Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

User Profiling and Sentiment Analysis for a brand using data from social medium Twitter

του

Πέτρου Αμοιρίδη του Αναστασίου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Επιβλέπων: Λεωνίδας Χατζηθωμάς

Επίκουρος Καθηγητής Πανεπιστημίου Μακεδονίας

ΘΕΣΣΑΛΟΝΙΚΗ

Αύγουστος 2022

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον κ. Λεωνίδα Χατζηθωμά, Επίκουρο Καθηγητή του Τμήματος Οργάνωσης & Διοίκησης Επιχειρήσεων του Πανεπιστημίου Μακεδονίας για την ανάθεση της παρούσας διπλωματικής εργασίας, τις συμβουλές, τις παρατηρήσεις και την ουσιαστική καθοδήγηση που μου παρείχε. Δεν θα μπορούσα ακόμη να λησμονήσω να ευχαριστήσω την σύζυγό μου Αννίτα για την ενθάρρυνση και την υποστήριξή της καθ' όλη τη διάρκεια εκπόνησης της εργασίας αυτής.

Περίληψη – Abstract

Στον σύγχρονο κόσμο των επιχειρήσεων και του εμπορίου η γνώση που αντλούν οι εταιρείες μέσω των κοινωνικών δικτύων (Twitter, Facebook, Instagram) για τα κοινά χαρακτηριστικά (profile) των πελατών τους, τους δίνει συγκριτικό πλεονέκτημα έναντι των υπολοίπων, καθώς συμβάλει δυναμικά, τόσο στην διαμόρφωση της εμπορικής τους στρατηγικής με στόχο την αύξηση, αλλά και την διεύρυνση, του καταναλωτικού τους κοινού, όσο και στην βελτίωση της εικόνας τους στο ευρύ κοινό μέσω στοχευμένων προωθητικών ενεργειών.

Σε αυτή την κατεύθυνση η τεχνική της Ανάλυσης Συναισθήματος (Sentiment Analysis) μέσω της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing (NLP)) των μηνυμάτων των χρηστών των κοινωνικών δικτύων για αυτές τις εταιρίες, οδηγεί σε δύο συχνά αλληλοσυνδεόμενες κατηγοριοποιήσεις. Η μία κατηγοριοποίηση (User Profiling) γίνεται βάσει των κοινών γνωρισμάτων (έκφραση συναισθημάτων, φρασεολογία, κ.α) των χρηστών και η άλλη τους κατατάσσει σε τρεις κατηγορίες βάσει της θετικότητας, αρνητικότητας ή ουδετερότητας που εκφράζουν μέσω των μηνυμάτων τους για την εκάστοτε εταιρεία (Ανάλυση συναισθήματος - Sentiment Analysis).

Στην παρούσα διπλωματική πραγματοποιείται αρχικά η ανάλυση συναισθήματος των μηνυμάτων (tweets) του κοινωνικού μέσου Twitter κατά την διάρκεια του SuperBowl, σχετικά με τον κλάδο της αυτοκινητοβιομηχανίας, αναδεικνύοντας παράλληλα τα προβλήματα και τους περιορισμούς που θέτουν τα εργαλεία που χρησιμοποιούνται για την ανάλυση αυτή. Τα αποτελέσματα έδειξαν ότι τα συναισθήματα που εκφράζουν οι χρήστες για τις εταιρίες που διαφημίζονται κατά την διάρκεια σημαντικών αθλητικών γεγονότων τείνουν να είναι περισσότερο θετικά, γεγονός που συνάδει με το γενικότερο αίσθημα ενθουσιασμού και χαράς που επικρατεί, ενώ υπήρξαν αυτοκινητοβιομηχανίες για τις οποίες τα μηνύματα εμφάνισαν σημαντικό ποσοστό αρνητικών σχολίων.

Στο δεύτερο μέρος επιχειρείται η εξόρυξη του προφίλ των χρηστών και η κατηγοριοποίησή τους βάση του μοντέλου προσωπικότητας OCEAN. Μία επιπλέον κατηγοριοποίηση γίνεται σύμφωνα με την έρευνα της Forrester Research, Inc. του 2010, η οποία ανέδειξε τις κατηγορίες των χρηστών των κοινωνικών μέσων. Για το σκοπό αυτό χρησιμοποιήθηκαν τα αποτελέσματα της ανάλυσης κειμένου που πραγματοποιήθηκε από την εφαρμογή LIWC2007 (Linguistic Inquiry and Word

Count) στα tweets της βάσης δεδομένων που αξιοποιήθηκαν για την ανάλυση συναισθήματος. Για την εκκαθάριση και προετοιμασία των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, ενώ για την εξόρυξη του προφίλ των χρηστών έγινε χρήση του προγράμματος στατιστικής ανάλυσης SPSS. Η έρευνά μας κατέληξε στο συμπέρασμα ότι οι δύο τύποι προσωπικότητας που αντιπροσωπεύουν το σύνολο σχεδόν των χρηστών του Twitter είναι αυτοί των εξωστρεφών και των νευρωτικών κατά το μοντέλο OCEAN, ενώ σύμφωνα με την κατηγοριοποίηση της Forrester Research οι χρήστες εμφανίζουν χαρακτηριστικά των κατηγοριών των συνομιλητών , των κριτικών και των συμμετεχόντων.

Λέξεις κλειδιά: Ανάλυση συναισθήματος, Λεξικό συναισθήματος, Εξόρυξη προφίλ (User Profiling), OCEAN, Επεξεργασία φυσικής γλώσσας, Πρόγραμμα ανάλυσης κειμένου LIWC, Twitter, Python, SPSS, Super Bowl, Αυτοκινητοβιομηχανίες

Περιεχόμενα

1.1. Εισαγωγή

Η εκτεταμένη χρήση των κοινωνικών μέσων από χρήστες του διαδικτύου οδηγεί όλο και μεγαλύτερο αριθμό εταιριών πάσης φύσεως και ενδιαφέροντος στην συλλογή και επεξεργασία δεδομένων που σκιαγραφούν το προφίλ των πελατών τους και τους προσφέρουν χρήσιμες πληροφορίες τόσο για τις αγοραστικές τους συνήθειες, όσο και για την γνώμη που έχουν γι' αυτές. Τους δίνουν έτσι την δυνατότητα να σχεδιάσουν την στρατηγική τους για την δημιουργία μιας σταθερότερης σχέσης μαζί τους, προτείνοντάς τους εξατομικευμένα προϊόντα και υπηρεσίες, εξυπηρετώντας με αυτό τον τρόπο τόσο τις υπάρχουσες όσο και μελλοντικές τους ανάγκες.

Στα πλαίσια της συλλογής και επεξεργασίας του τεράστιου πλέον όγκου δεδομένων των κοινωνικών μέσων, δημιουργήθηκαν νέοι τομείς έρευνας και αναπτύχθηκαν τεχνικές και μέθοδοι, οι οποίες βελτιώνονται συνεχώς προκειμένου να δίνουν όλο και πιο ακριβή αποτελέσματα και να οδηγούν σε όλο και πιο χρήσιμα και αξιοποιήσιμα συμπεράσματα.

Η ανάλυση συναισθήματος (ή εξόρυξη γνώμης) είναι ένας από αυτούς του τομείς έρευνας, που έχουν ως στόχο την ανάδειξη του συναισθήματος που εκφράζεται μέσω των δημοσιευμένων μηνυμάτων των χρηστών στα κοινωνικά μέσα, όπως το Twitter, και η κατηγοριοποίησή τους σε θετικό, αρνητικό ή ουδέτερο. Τα μηνύματα αυτά μπορούν να έχουν την μορφή σχολίων, κριτικών για προϊόντα και υπηρεσίες, συζητήσεων σε forum, δημοσίευση κατάστασης του χρήστη κ.α, και αποτελούν για τις εταιρίες, τις κυβερνήσεις και του διάφορους οργανισμούς πολύτιμο εργαλείο έρευνας. Τα τελευταία χρόνια η ανάλυση συναισθήματος αποτελεί αναπόσπαστο κομμάτι της έρευνας στο αντικείμενο της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), ενώ η μελέτη της πραγματοποιείται συχνά σε συνδυασμό με άλλους τομείς όπως η ανάλυση κειμένου, η σύνοψη κειμένου, η μηχανική μάθηση, η Εξόρυξη Προφίλ Χρήστη (User Profiling) κ.α. Υπάρχουν δύο βασικοί τύποι μεθόδων ανάλυσης συναισθήματος. Ο πρώτος βασίζεται στην μηχανική μάθηση και ο δεύτερος στην χρήση λεξικών όπως το λεξικό της εφαρμογής LIWC (Linguistic Inquiry and Word Count).

Η Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ) αφορά σε τεχνικές που επιτρέπουν στους υπολογιστές να επεξεργάζονται και να κατανοούν την ανθρώπινη φυσική γλώσσα, που εμφανίζεται είτε σε μορφή γραπτού κειμένου (π.χ. ένα tweet), είτε σε

λεκτική μορφή (π.χ. ομιλία). Υπάρχει πληθώρα εφαρμογών της ΕΦΓ, όπως είναι η μηχανική μετάφραση, η ανάπτυξη συστημάτων αναγνώρισης ομιλίας, η σύνοψη κειμένου, η κατηγοριοποίηση κειμένου, η ανάλυση κειμένου, στα πλαίσια της οποίας εντάσσεται η ανάλυση συναισθήματος, η μηχανική μάθηση κ.α.

Η μηχανική μάθηση είναι παρακλάδι της τεχνητής νοημοσύνης (AI – Artificial Intelligence) και αποτελεί πεδίο της επιστήμης των υπολογιστών που με βάση τις μαθηματικές μεθοδολογίες και τη Στατιστική επιτρέπει στους υπολογιστές να αποκτούν γνώση από τα δεδομένα που τους εισάγονται και να κάνουν προβλέψεις σχετικά με αυτά. Με αυτό τον τρόπο η μηχανική μάθηση βρίσκει εφαρμογή στον τομέα της ανάλυσης συναισθήματος των δημοσιευμένων tweets των χρηστών, κατηγοριοποιώντας τα ανάλογα με την θετικότητα, αρνητικότητα ή ουδετερότητα των συναισθημάτων που εκφράζουν.

Σε έναν σημαντικά μεγάλο αριθμό ερευνών για την πρόβλεψη του συνολικού συναισθήματος ενός κειμένου, η μηχανική μάθηση βασίζεται στην χρήση λεξικών, όπως το LIWC, δηλαδή προκαθορισμένων λιστών λέξεων, όπου κάθε λέξη μπορεί να συνδέεται με ένα συγκεκριμένο συναίσθημα και επισημαίνεται ως θετική, αρνητική ή ουδέτερη, βάσει μια προκαθορισμένες τιμής που αντικατοπτρίζει την ισχύ ή την ένταση του συναισθήματος.

Πεδίο εφαρμογής της ανάλυσης συναισθήματος αποτελεί και η εξόρυξη προφίλ χρήστη μέσω των δημοσιευμένων μηνυμάτων του στα κοινωνικά μέσα. Σύμφωνα με τον Pennebaker οι κατηγορίες των λέξεων που χρησιμοποιούν οι άνθρωποι στις καθημερινές τους συζητήσεις συνδέονται με συγκεκριμένα χαρακτηριστικά ενός τύπου προσωπικότητας όπως αυτή ορίζεται από το μοντέλο Big Five (ή αλλιώς OCEAN)[1] . Κατά συνέπεια, σχόλια σε μορφή κειμένου (tweets) στο κοινωνικό μέσο Twitter αποτελέσουν σημαντικά δεδομένα για τον προσδιορισμό του προφίλ του χρήστη που τα δημοσιεύει και συνδέεται σε σημαντικό βαθμό με τον τύπο προσωπικότητάς του. Οι πληροφορίες αυτές για το προφίλ των χρηστών είναι χρήσιμες τόσο σε επίπεδο έρευνας σε τομείς όπως η ψυχολογία, η ασφάλεια στο διαδίκτυο, η ανάδειξη κοινωνικών ανισοτήτων κ.α, όσο και σε επίπεδο εμπορικών και διαφημιστικών εφαρμογών.

Στα πλαίσια της έρευνας στο τομέα της ανάλυσης συναισθήματος, τα τελευταία χρόνια έχουν πραγματοποιηθεί μελέτες που εστιάζουν στην προσπάθεια εξόρυξης του προφίλ των χρηστών βάση του γραπτού τους λόγου στα κοινωνικά μέσα. Ωστόσο, τα μηνύματα αυτά αφορούν κυρίως πολιτικά και κοινωνικά γεγονότα. Στην

παρούσα διπλωματική επιχειρείται αρχικά η ανάλυση συναισθήματος μέσω των δημοσιευμένων tweets των χρηστών στο κοινωνικό μέσο Twitter και πιο συγκεκριμένα την γνώμη που εκφράζουν για συγκεκριμένες επωνυμίες αναφορικά με τον κλάδο της αυτοκινητοβιομηχανίας και πιο συγκεκριμένα για τις εταιρίες Audi, Chevrolet, Chrysler, KIA και Volkswagen. Τα δεδομένα έχουν συλλεχθεί κατά την διάρκεια του Super Bowl [1] του 2014, του σημαντικότερου και δημοφιλέστερου ετήσιου αγώνα του πρωταθλήματος του αμερικανικού ποδοσφαίρου.

Στο δεύτερο σκέλος της διπλωματικής επιχειρείται η εξόρυξη του προφίλ των χρηστών του κοινωνικού μέσου Twitter μέσω των μηνυμάτων τους (tweets) σε αυτό. Τα μηνύματα κι εδώ αφορούν το κλάδο της αυτοκινητοβιομηχανίας και πιο συγκεκριμένα τις εταιρίες Audi, Chevrolet, Chrysler, KIA και Volkswagen.

1.2. Σκοπός

Σκοπός της παρούσας εργασίας είναι η ανάδειξη της χρησιμότητας διεξαγωγής αναλύσεων σε μηνύματα των χρηστών των κοινωνικών μέσων από τις εταιρίες πάσης φύσεως και αντικειμένου κατά την διάρκεια διεξαγωγής σημαντικών γεγονότων, με στόχο την εξαγωγή χρήσιμων και αξιοποιήσιμων πληροφοριών αναφορικά με το προφίλ των χρηστών αυτών, τις ανάγκες τους, τις επιθυμίες τους και τις προτιμήσεις τους, αλλά και το συναίσθημα που αυτοί εκφράζουν για συγκεκριμένες επωνυμίες. Τα μεγάλα αθλητικά γεγονότα μάλιστα, που συνήθως συνοδεύονται από πληθώρα προωθητικών και διαφημιστικών ενεργειών από τις εταιρίες που συμμετέχουν σε αυτά ως χορηγοί, είναι ίσως η πιο πλούσια πηγή τέτοιων πληροφοριών, καθώς τότε υπάρχει πιο έντονη και ενθουσιώδης συμμετοχή του καταναλωτικού κοινού σε συζητήσεις για αυτές τις εταιρίες και τα πεπραγμένα τους. Για το σκοπό αυτό, στο πρώτο σκέλος της διπλωματικής πραγματοποιήσαμε μία ανάλυση συναισθήματος των tweets των χρηστών του Twitter που αφορούν προϊόντα και εταιρίες του κλάδου της αυτοκινητοβιομηχανίας, αποσκοπώντας στην εκμαίευση του γενικού αισθήματος που τα συνοδεύει, ενώ στο δεύτερο σκέλος, διερευνήσαμε και αναδείξαμε ως έναν βαθμό τα προφίλ των χρηστών (user profiling) αυτών με την βοήθεια των αποτελεσμάτων της ανάλυσης των παραπάνω tweets από την εφαρμογή LIWC2007.

[1] Το Super Bowl είναι αθλητικό γεγονός με την μεγαλύτερη τηλεθέαση κάθε χρόνο στις ΗΠΑ και δεύτερο παγκοσμίως σε ετήσια αθλητικά γεγονότα μετά τον τελικό του ΟΥΕΦΑ Τσάμπιονς Λιγκ.

1.3. Δομή διπλωματικής

Στο 2ο Κεφάλαιο δίνεται ο ορισμός των μέσων κοινωνικής δικτύωσης, αναπτύσσονται οι κατηγορίες και τα χαρακτηριστικά τους, ενώ στη συνέχεια παρουσιάζονται οι κατηγορίες των χρηστών των μέσων κοινωνικής δικτύωσης σύμφωνα με την διαδικτυακή έρευνα της Forrester Research, Inc., που πραγματοποιήθηκε το 2010. Τέλος, παρουσιάζεται το κοινωνικό δίκτυο Twitter, που αποτελεί και την πηγή των δεδομένων την ανάλυσή, αναπτύσσεται ο τρόπος λειτουργίας και τα χαρακτηριστικά του, καθώς επίσης αξιολογείται η χρησιμότητά του στην διεξαγωγή μελετών των κοινωνικών επιστημών.

Το 3ο Κεφάλαιο διαπραγματεύεται την Ανάλυση Συναισθήματος (Sentiment Analysis), τα επίπεδα και τις μεθόδου ανάλυσης. Πιο συγκεκριμένα γίνεται παρουσίαση της μεθόδου μηχανικής μάθησης και της μεθόδου ανάλυσης με χρήση λεξικού.

Στο 4ο Κεφάλαιο δίνονται οι ορισμοί της Εξόρυξης Προφίλ Χρηστών (User Profiling) και κατηγοριοποίησης βάσει της προσωπικότητάς τους σύμφωνα με το μοντέλο OCEAN.

Στο 5ο Κεφάλαιο παρουσιάζεται η μεθοδολογία που ακολουθήθηκε, γίνεται περιγραφή των δεδομένων προς ανάλυση, καθώς και των εργαλείων ανάλυσης που χρησιμοποιήθηκαν για την ερευνητική διαδικασία.

Στο 6ο Κεφάλαιο αναπτύσσεται η ερευνητική διαδικασία που ακολουθήθηκε για την ανάλυση συναισθήματος και την εξόρυξη του προφίλ των χρηστών του κοινωνικού μέσου Twitter.

Στο 7ο Κεφάλαιο παρατίθενται τα συμπεράσματα, οι περιορισμοί της έρευνας και οι προτάσεις για μελλοντικές βελτιώσεις.

Στο 8ο Κεφάλαιο παρουσιάζεται αναλυτικά η βιβλιογραφία που χρησιμοποιήθηκε για την συγγραφή της διπλωματικής εργασίας.

2. Μέσα Κοινωνικής Δικτύωσης

2.1. Εισαγωγή στα μέσα κοινωνικής δικτύωσης

Η εξέλιξη του Παγκόσμιου Ιστού (World Wide Web – WEB) από την αρχική του μορφή (WEB 1.0) όπου ο χρήστης του διαδικτύου ήταν ένας παθητικός δέκτης πληροφοριών, στην σύγχρονη μορφή του (WEB 2.0) σηματοδότησε την στροφή σε μια καθόλα ανθρωποκεντρική προσέγγιση της οργάνωσης της πληροφορίας και έδωσε την δυνατότητα σε κάθε χρήστη να αλληλοεπιδρά με άλλους χρήστες, να δημιουργεί πρωτότυπο περιεχόμενο και να το διαμοιράζεται, να εμπορεύεται προϊόντα και υπηρεσίες, να δημιουργεί και να συντηρεί online κοινότητες, να συνεργάζεται και να συμμετέχει σε συλλογικές δράσεις και να εκφράζει την γνώμη του για γεγονότα που συμβαίνουν σε όλο τον κόσμο.

Το WEB 2.0 βασισμένο σε χαρακτηριστικά όπως η ελεύθερη διακίνηση της πληροφορίας, η προώθηση του δημοκρατικού χαρακτήρα του διαδικτύου, η δυνατότητα για ανοιχτή και αμφίδρομη επικοινωνία των χρηστών μεταξύ τους αλλά και με επιχειρήσεις ή οργανισμούς, οδήγησε στην εμφάνιση και τη ραγδαία ανάπτυξη των μέσων κοινωνικής δικτύωσης, όπως το Facebook, το Twitter, το Instagram, το TikTok κ.α.

2.2. Ορισμός των μέσων κοινωνικής δικτύωσης

Από την αρχή της δημιουργίας τους μέχρι και σήμερα έχουν διατυπωθεί αρκετοί διαφορετικοί ορισμοί για τα κοινωνικά δίκτυα. Σύμφωνα με τους Boyd & Ellison [2], «Τα κοινωνικά δίκτυα ορίζονται ως βασισμένες στο διαδίκτυο (διαδικτυακές) υπηρεσίες που (1) επιτρέπουν στα άτομα να δημιουργήσουν ένα δημόσιο ή ημι-δημόσιο προφίλ μέσα σε ένα οριοθετημένο σύστημα, (2) να επικοινωνήσουν με μια λίστα από άλλους χρήστες με τους οποίους μοιράζονται μια μορφή σύνδεσης και (3) να δουν και να διανείμουν την δικιά τους λίστα των συνδέσεων και αυτών που φτιάχτηκαν από άλλους μέσα στο σύστημα». Οι όροι «μέσα κοινωνικής δικτύωσης» και «κοινωνικό δίκτυο» συχνά ταυτίζονται κάτω από τον όρο «κοινωνική δικτύωση». Ωστόσο, υπάρχει μια σημαντική διαφοροποίηση: ο όρος «κοινωνικά μέσα» (social media) αναφέρεται στα μέσα (εργαλεία) διαμοιρασμού της πληροφορίας, των δεδομένων και της επικοινωνίας στο κοινό, ενώ ο όρος «κοινωνική δικτύωση» αναφέρεται στη δημιουργία και την αξιοποίηση των κοινοτήτων για τη διασύνδεση ανθρώπων με κοινά ενδιαφέροντα [3]. Ένας σύγχρονος, αφαιρετικός, και περισσότερο περιγραφικός ορισμός δόθηκε από τους Caleb T. Carr και Rebecca A. Hayes (2015) [4] σύμφωνα με τον οποίον: «Τα κοινωνικά

μέσα είναι κανάλια που βασίζονται στο Διαδίκτυο και επιτρέπουν στους χρήστες να αλληλοεπιδρούν και να αυτοπαρουσιάζονται επιλεκτικά, είτε σε πραγματικό χρόνο είτε ασύγχρονα, τόσο σε ευρύ όσο και σε περιορισμένο κοινό, αντλώντας αξία από περιεχόμενο που δημιουργεί ο χρήστης και την εντύπωση που προκαλεί μέσω της αλληλεπίδρασης με άλλους». Τα κοινωνικά μέσα σήμερα αποτελούν κυρίαρχο κομμάτι της καθημερινότητας των σύγχρονων ανθρώπων. Το 2019, το διαδικτυακό λεξικό Merriam-Webster όρισε τα κοινωνικά μέσα ως «μορφές ηλεκτρονικής επικοινωνίας μέσω των οποίων οι χρήστες δημιουργούν διαδικτυακές κοινότητες για να μοιράζονται πληροφορίες, ιδέες, προσωπικά μηνύματα και άλλο περιεχόμενο [5] .

Τα βασικά χαρακτηριστικά των κοινωνικών μέσων σύμφωνα με τους Mayfield C., Perdue G. και Wooten K. (2008)[6] είναι:

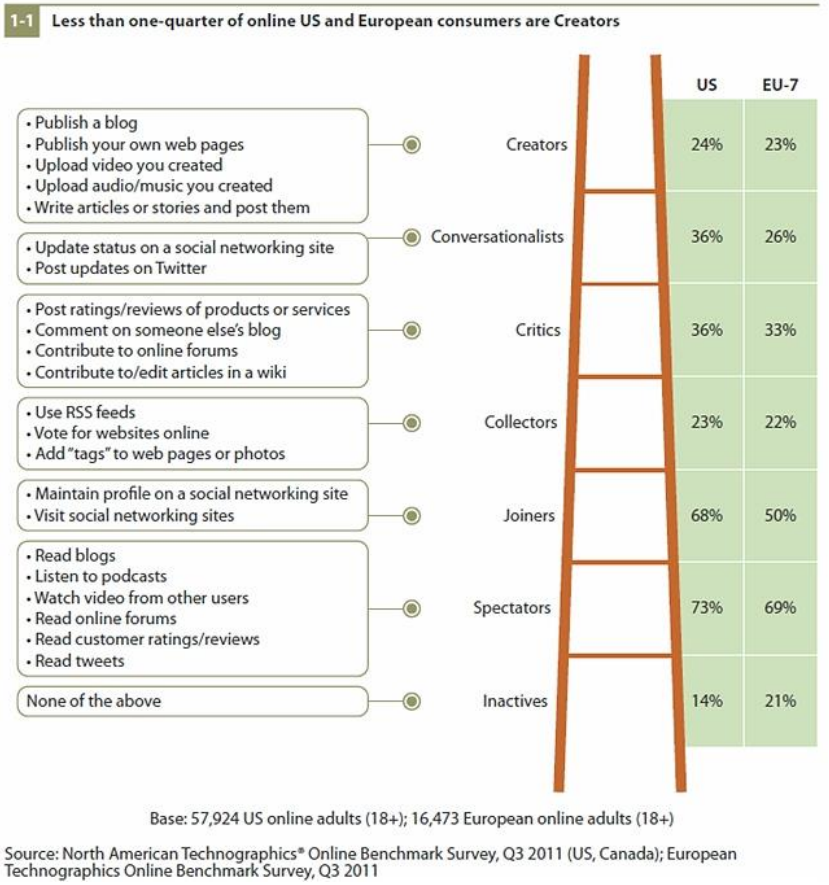
- ✓ ***Συμμετοχή (Participation)***: Τα κοινωνικά μέσα ενθαρρύνουν την συνεισφορά και τον σχολιασμό από τους ενδιαφερομένους χρήστες. Η συμμετοχή των χρηστών δημιουργεί ακαθόριστα όρια μεταξύ των μέσων ενημέρωσης και του κοινού, δημιουργώντας έτσι μια ελεύθερη επικοινωνία χωρίς οριοθετημένα πλαίσια. Τα όρια που σε περιπτώσεις, όπως τα οικογενειακά δίκτυα ή τα φιλικά, είναι εύκολο να προσδιοριστούν, είναι δυσδιάκριτα στα μεγαλύτερα κοινωνικά δίκτυα.
- ✓ ***Διαφάνεια (Openness)***: Οι περισσότερες υπηρεσίες των κοινωνικών μέσων είναι ανοιχτές σε ανατροφοδότηση και συμμετοχή, ενώ σπάνια υπάρχουν εμπόδια στην πρόσβαση και στην χρήση του περιεχομένου.
- ✓ ***Συνομιλία (Conversation)***: Σε αντίθεση με τα παραδοσιακά μέσα ενημέρωσης που αφορούν μόνο την μετάδοση ενός περιεχομένου σε ένα ακροατήριο, το οποίο είναι παθητικός δέκτης των πληροφοριών, τα κοινωνικά μέσα δίνουν την δυνατότητα αμφίδρομης επικοινωνίας, σχολιασμού και ανταλλαγής απόψεων.
- ✓ ***Κοινότητα (Community)***: Τα κοινωνικά μέσα επιτρέπουν την εύκολη και άμεση δημιουργία κοινοτήτων που μοιράζονται κοινά ενδιαφέροντα. Οι χρήστες επικοινωνούν ανταλλάσσοντας εμπειρίες και απόψεις πάνω στο αντικείμενο του ενδιαφέροντός τους, όπως την ενασχόλησή τους με τον αθλητισμό, την αστρονομία ή την φωτογραφία.
- ✓ ***Συνεκτικότητα (Connectedness)***: Τα περισσότερα είδη των κοινωνικών μέσων αναπτύσσουν την συνεκτικότητα τους, κάνοντας χρήση συνδέσεων με άλλες ιστοσελίδες, πόρους και ανθρώπους.

2.3. Κατηγορίες χρηστών των κοινωνικών μέσων

Η διαδικτυακή έρευνα της Forrester Research, Inc., που πραγματοποιήθηκε το 2010 στην Αμερική σε δείγμα ενήλικων, οδήγησε στην δημιουργία 7 διαφορετικών ομάδων χρηστών, κάθε μια από τις οποίες εμφανίζει μια λίστα από δραστηριότητες στις οποίες συμμετέχουν οι χρήστες της αντίστοιχης ομάδας κάθε μήνα (ή στην περίπτωση των Conversatinalist, εβδομαδιαίως) κατά την ενασχόλησή τους με τα κοινωνικά μέσα. Οι ομάδες αυτές κατατάσσονται στα 7 επίπεδα της «Σκάλας των κοινωνικών τεχνολογικών συμπεριφορών» (Εικόνα 1) [7]. Οι ομάδες αυτές είναι οι ακόλουθες.

1. ***Δημιουργοί (Creators)***: Σε αυτή την ομάδα ανήκουν οι χρήστες που είναι ενεργοί στα μέσα κοινωνικής δικτύωσης, δημοσιεύοντας τουλάχιστον μια φορά το μήνα, ένα ιστολόγιο ή ένα άρθρο, αναρτώντας video, εικόνες ή μουσική σε πλατφόρμες όπως το YouTube.
2. ***Συνομιλητές (Conversationalists):*** Η ομάδα αυτή αποτελείται από χρήστες που συμμετέχουν σε εκατέρωθεν διαλόγους, που χαρακτηρίζουν τις ενημερώσεις κατάστασης στο Facebook και το Twitter.
3. ***Κριτικοί (Critics)***: Αυτή η ομάδα περιλαμβάνει χρήστες που αντιδρούν σε περιεχόμενο που έχουν δημιουργήσει άλλοι χρήστες, δημοσιεύουν σχόλια, αξιολογήσεις ή κριτικές για προϊόντα ή υπηρεσίες.
4. ***Συλλέκτες (Collectors)***: Είναι οι χρήστες που αποθηκεύουν διευθύνσεις URL και σελιδοδείκτες σε ιστότοπους με υπηρεσίες κοινωνικής σελιδοσήμανσης ειδήσεων (social bookmarking) (όπως το Digg και το Delicious), ψηφίζουν άρθρα και δημοσιεύσεις που έχουν εντοπίσει στο διαδίκτυο, ή χρησιμοποιούν ροές RSS (Really Simple Syndication feeds). Αυτή η πράξη συλλογής και συγκέντρωσης πληροφοριών διαδραματίζει σημαντικό ρόλο στην οργάνωση του τεράστιου περιεχομένου που παράγεται από δημιουργούς και κριτικούς
5. ***Συμμετέχοντες (Joiners)***: Τα άτομα που συμμετέχουν ή διατηρούν λογαριασμούς σε ιστότοπους κοινωνικής δικτύωσης, όπως το Twitter και το Facebook.
6. ***Θεατές (Spectators)***: Οι θεατές οι είναι εκείνοι που «καταναλώνουν» ότι παράγουν οι υπόλοιποι χρήστες, και αποτελούν την μεγαλύτερη ομάδα αυτής της κατάταξης. Θεατής είναι το άτομο που διαβάζει ιστολόγια, ακούει διαδικτυακές ραδιοφωνικές εκπομπές (podcasts), παρακολουθεί video άλλων χρηστών, διαβάζει συζητήσεις σε forums, διαβάσει κριτικές και αξιολογήσεις καταναλωτών και διαβάζει tweets.

7. ***Αδρανείς (Inactives)***: Στην ομάδα αυτή ανήκουν τα άτομα που δεν χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης και περιορίζονται στο να κάνουν απλή χρήση του διαδικτύου.



*Εικόνα 1. Κατηγορίες χρηστών των κοινωνικών μέσων (**Πηγή:** Forrester Research, Inc.)*

2.4. Το κοινωνικό μέσο Twitter

Το Twitter είναι μια αμερικανική υπηρεσία μικρο-ιστολογίου (micro-blogging) και κοινωνικής δικτύωσης (social network) μέσω της οποία οι χρήστες δημοσιεύουν και αλληλεπιδρούν με μηνύματα γνωστά ως «tweets» («τιτιβίσματα»). Οι εγγεγραμμένοι χρήστες μπορούν να δημοσιεύουν tweets όπως και retweet, αλλά οι μη εγγεγραμμένοι χρήστες μπορούν να διαβάσουν μόνο αυτά που είναι διαθέσιμα στο κοινό. Οι χρήστες αλληλεπιδρούν με το Twitter μέσω ενός προγράμματος περιήγησης ή εφαρμογής σε κινητές συσκευές και σε προγραμματιστικό επίπεδο μέσω των API (application programming interface) τους. Τα tweets είναι σύντομα μηνύματα 140 χαρακτήρων, αλλά τον Νοέμβριο του 2017 το όριό τους διπλασιάστηκε σε 280 χαρακτήρες σε όλες τις γλώσσες εκτός των Ιαπωνικών, Κινέζικων και Κορεάτικων[8]. Τα tweets ήχου και βίντεο παραμένουν περιορισμένα σε 140 δευτερόλεπτα για τους περισσότερους λογαριασμούς.

Το Twitter δημιουργήθηκε από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams τον Μάρτιο του 2006 και ξεκίνησε την λειτουργία του τον Ιούλιο του ίδιου έτους. Η εταιρεία Twitter εδρεύει στο Σαν Φρανσίσκο της Καλιφόρνια και διαθέτει περισσότερα από 25 γραφεία σε όλο τον κόσμο[9]. Μέχρι το 2012, περισσότεροι από 100 εκατομμύρια χρήστες δημοσίευσαν 340 εκατομμύρια tweets την ημέρα [10], και η υπηρεσία χειρίστηκε κατά μέσο όρο 1,6 δισεκατομμύρια ερωτήματα αναζήτησης την ημέρα [10] [11], [11]. Το 2013, ήταν ένας από τους δέκα ιστότοπους με τις περισσότερες επισκέψεις και έχει περιγραφεί ως «το SMS του Διαδικτύου» [12]. Από το πρώτο τρίμηνο του 2019, το Twitter είχε περισσότερους από 330 εκατομμύρια μηνιαίους ενεργούς χρήστες [13]. Στην πράξη, η συντριπτική πλειοψηφία των tweets γράφεται από μια μειονότητα χρηστών[14], [15]. Στις 25 Απριλίου του 2022, το διοικητικό συμβούλιο του Twitter συμφώνησε σε εξαγορά ύψους 44 δισεκατομμυρίων δολαρίων από τον Elon Musk, τον διευθύνοντα σύμβουλο της SpaceX και της Tesla [16], [17].

2.4.1. Λειτουργία και χαρακτηριστικά

Ο χρήστης μπορεί να δημιουργήσει έναν δωρεάν λογαριασμό στο Twitter εισάγοντας το όνομά του και ένα ψευδώνυμο που φέρει αρχικά το σύμβολο at sign (@) και στη συνέχεια μπορεί να ακολουθήσει άλλους χρήστες (follow) για να διαβάσει τις δημοσιεύσεις τους, δηλαδή τα tweets τους, στην αρχική του σελίδα.

Το περιβάλλον του Twitter περιλαμβάνει τις εξής σελίδες:

1. **Αρχική σελίδα (Home)**

Η αρχική σελίδα παρέχει μια σύνοψη της δραστηριότητας του λογαριασμού του χρήστη και επισημαίνει τα κορυφαία tweets, αναφορές και ακολούθους ανά μήνα.

2. **Εξερεύνηση (Explore)**

Η εξερεύνηση εμφανίζει μια σειρά από εξατομικευμένα θέματα βάσει της δραστηριότητας του χρήστη, όπως δημοφιλή hashtag, θέματα και ενημερώσεις, τα οποία επίσης διαφέρουν ανάλογα με την τοποθεσία και τις ρυθμίσεις που έχει επιλέξει ο χρήστης.

3. **Ειδοποιήσεις (Notifications)**

Το χρονοδιάγραμμα ειδοποιήσεων δίνει πληροφορίες για την αλληλεπίδραση ενός χρήστη με τους άλλους χρήστες, όπως ποια tweets του έχουν αρέσει, τα πιο πρόσφατα retweets (των tweets του), τα tweets που απευθύνονται σε αυτόν, καθώς και τους νέους ακολούθους του προφίλ του και της λίστας του.

4. **Προσωπικά μηνύματα (Messages)**

Τα προσωπικά μηνύματα δίνουν την δυνατότητα στους χρήστες να έχουν ιδιωτικές συνομιλίες με άλλα άτομα σχετικά με δημοσιευμένα tweets και άλλο περιεχόμενο.

5. **Σελιδοδείκτες (Bookmarks)**

Οι σελιδοδείκτες επιτρέπουν στους χρήστες να αποθηκεύουν tweets σε ένα χρονοδιάγραμμα για εύκολη και γρήγορη πρόσβαση ανά πάσα στιγμή.

6. **Λίστες (Lists)**

Οι λίστες επιτρέπουν στους χρήστες να προσαρμόζουν, να οργανώνουν και να δίνουν προτεραιότητα στα Tweets που εμφανίζονται στο χρονολόγιό τους, να επιλέγουν την συμμετοχή τους σε λίστες που δημιουργήθηκαν από άλλους στο Twitter ή να δημιουργούν από τον δικό τους λογαριασμό λίστες άλλων λογαριασμών κατά ομάδα, θέμα ή ενδιαφέρον.

7. **Προφίλ του χρήστη (Profile)**

Το προφίλ περιέχει πληροφορίες του χρήστη, όπως το όνομα, η βιογραφία, η τοποθεσία και η εικόνα του, οι οποίες είναι πάντα δημόσιες. Για ορισμένα πεδία πληροφοριών του προφίλ υπάρχουν ρυθμίσεις ώστε να επιλέγεται η ορατότητά τους σε άλλους χρήστες.

Η λειτουργία του Twitter βασίζεται ουσιαστικά στην ανταλλαγή σύντομων συμβολοσειρών με περιεχόμενο σε μορφή κειμένου και σχετίζονται με την ενημέρωση κατάστασης ενός χρήση (tweets). Τα tweets, εκτός από το περιεχόμενο κειμένου, συνοδεύονται από δύο επιπλέον κομμάτια μεταδεδομένων (metadata), τις οντότητες και τα μέρη. Οι οντότητες (entities) είναι ουσιαστικά οι αναφορές των χρηστών, τα hashtags, οι διευθύνσεις URL και τα μέσα που μπορεί να σχετίζονται με ένα tweet. Τα μέρη (places) είναι τοποθεσίες του πραγματικού κόσμου που μπορεί να επισυναφθούν σε ένα tweet και μπορεί να αφορούν είτε την πραγματική τοποθεσία στην οποία συντάχθηκε ένα tweet, είτε κάποιο μέρος που αναφέρεται το tweet[18]. Ένα Tweet μπορεί επίσης να περιέχει φωτογραφίες, GIF, βίντεο, συνδέσμους και κείμενο.

Ένας χρήστης μπορεί να συμμετάσχει σε μια συνομιλία απαντώντας σε κάποιο tweet, πατώντας στο εικονίδιο 💬. Η συνομιλία μεταξύ δύο ατόμων είναι ορατή μόνο σε ακολούθους των δύο αυτών ατόμων, ενώ πατώντας πάνω στο tweet εμφανίζεται η πλήρη λίστα των ονομάτων των χρηστών που συμμετέχουν στη συνομιλία των συμμετεχόντων στη συνομιλία κάνοντας κλικ ή πατώντας την ερώτηση πάνω από το tweet. Τα μηνύματα που δημοσιεύει ένας χρήστης είναι δημόσια από προεπιλογή, ενώ υπάρχει η δυνατότητα

επιλογής του κοινού στο οποίο θα είναι ορατά, μέσα από τις ρυθμίσεις του λογαριασμού του.

Ο διαμοιρασμός ενός tweet από έναν χρήστη ονομάζεται retweet και μπορεί να συνοδεύεται από τα δικά του σχόλια επί του συγκεκριμένου tweet, ενώ μπορεί να γίνεται αναφορά και στον συγγραφέα του αρχικού tweet με την χρήση του συμβόλου at sign (@).

Μια επιπλέον δυνατότητα του Twitter είναι η χρήση του hashtag - γραμμένο με σύμβολο # (κάγκελο) - το οποίο χρησιμοποιείται για τη δημιουργία ευρετηρίου λέξεων-κλειδιών ή θεμάτων, επιτρέποντας στους χρήστες να ακολουθούν εύκολα θέματα που τους ενδιαφέρουν. Οι χρήστες χρησιμοποιούν το σύμβολο hashtag (#) πριν από μια σχετική λέξη-κλειδί ή φράση στο tweet τους για να κατηγοριοποιήσουν αυτά τα tweets και να τους βοηθήσουν να εμφανίζονται πιο εύκολα στην αναζήτηση στο Twitter. Κάνοντας κλικ ή πατώντας μια λέξη με hashtag, εμφανίζονται άλλα tweets που περιλαμβάνουν αυτό το hashtag. Οι λέξεις με hashtag που γίνονται πολύ δημοφιλείς αναφέρονται συχνά σε δημοφιλή θέματα.

2.4.2. Η αξία των Tweets

Στη σύγχρονη κοινωνία της πληροφορίας και της άμεσης και ποικιλότροπης διάδρασης των ανθρώπων μεταξύ τους, αλλά και με εταιρίες πάσης φύσεως, τα δημοσιευμένα σχόλια, μηνύματα και κριτικές στα κοινωνικά μέσα έχουν αποκτήσει μεγάλη δυναμική και αξία και αποτελούν τα τελευταία 20 χρόνια αντικείμενο έρευνας τόσο της ακαδημαϊκής κοινότητας, όσο και των εταιριών ή και σε αρκετές περιπτώσεις σε συνεργασία μεταξύ τους. Η έρευνα αυτή καλύπτει ένα ευρύ φάσμα κοινωνικοπολιτικών φαινομένων και ανθρώπινων συμπεριφορών, με απώτερο στόχο την εξαγωγή συμπερασμάτων και την χρήση αυτών για την βελτίωση υπηρεσιών και προϊόντων, την χάραξη νέων εμπορικών και πολιτικών στρατηγικών, αλλά και την εξέλιξη της κοινωνίας στο σύνολό της.

Εκείνοι που χρησιμοποιούν τα κοινωνικά μέσα μπορούν να δημοσιεύσουν τις σκέψεις, τα συναισθήματα ή/και τις απόψεις τους για σχεδόν κάθε πτυχή της ζωής τους [19]. Το περιεχόμενο των κοινωνικών μέσων, επομένως, παρουσιάζει στους ακαδημαϊκούς ερευνητές νέες σημαντικές ευκαιρίες μελέτης μιας σειράς θεμάτων σε ένα φυσικό περιβάλλον. Υπάρχουν ορισμένα ηθικά ζητήματα που σχετίζονται με τη διεξαγωγή αυτής της έρευνας, ωστόσο, υπάρχουν τεράστια οφέλη που μπορούν

να προκύψουν από αυτήν την έρευνα, στην κατανόηση του τι και πώς επικοινωνούν οι άνθρωποι σε συγκεκριμένες καταστάσεις.

Τα κοινωνικά μέσα αλλάζουν τον τρόπο επικοινωνίας των ανθρώπων, τόσο στην καθημερινή τους ζωή, όσο και σε ακραίες συνθήκες, όπως για παράδειγμα, καταστροφές που μπορεί να απειλήσουν άτομα, ομάδες ανθρώπων και τη συνολική δημόσια υγεία σε τοπικές και περιφερειακές περιοχές [20]. Οι Merchant, Elmer και Lurie (2011) αναφέρουν ότι η εμπλοκή και η χρήση πλατφορμών κοινωνικών μέσων όπως το Twitter μπορεί να θέσει την κοινότητα διαχείρισης έκτακτης ανάγκης σε καλύτερη θέση για να είναι σε θέση να ανταποκριθεί σε αναδυόμενες καταστροφές. Η έρευνα στο Twitter καλύπτει ένα ευρύ φάσμα, όπως η ανάλυση tweets που σχετίζονται με εξεγέρσεις [21], φυσικές καταστροφές [23],[22] και κρίσιμα γεγονότα [25],[23]. Καθώς η χρήση των κοινωνικών μέσων έχει αλλάξει τον τρόπο επικοινωνίας των ανθρώπων [24], π.χ., κατά τη διάρκεια καταστάσεων έκτακτης ανάγκης, οι πληροφορίες είναι πλέον διαθέσιμες από το κοινό και μπορούν να χρησιμοποιηθούν για την ενημέρωση της επίγνωσης της κατάστασης των καταστάσεων έκτακτης ανάγκης και για να βοηθήσουν τους υπεύθυνους συντονισμού κρίσεων να ανταποκριθούν κατάλληλα. Οι μελέτες σε σχέση με τις φυσικές καταστροφές έχουν διαπιστώσει ότι το Twitter προσφέρει ένα αποφασιστικό κανάλι επικοινωνίας μεταξύ της κυβέρνησης, των ανταποκριτών έκτακτης ανάγκης και του κοινού κατά τη διάρκεια κρίσεων [25] [23]. Αν και αυτές οι νέες πηγές πληροφοριών δεν θα αντικαταστήσουν τις υπάρχουσες, μπορούν να παράσχουν μια νέα πηγή δεδομένων που δυνητικά θα μπορούσε να έχει πολλές εφαρμογές στο πλαίσιο της διαχείρισης καταστάσεων έκτακτης ανάγκης και του συντονισμού κρίσεων.

Τα tweets μπορούν να συνιστούν επίσης ένα ισχυρό εργαλείο για τις εταιρίες/επιχειρήσεις. Η αξιολόγηση των tweets των πελατών μια επιχείρησης μπορεί να συνεισφέρει στο τομέα του μάρκετινγκ και της διαφήμισης, βοηθώντας στην ενίσχυση μιας καμπάνιας ή/και την δημιουργία νέων προωθητικών ενεργειών. Σε αυτή την κατηγορία, εντάσσεται και ο εντοπισμός κακόβουλων υποστηρικτών σε καμπάνιες που αναπτύσσονται στα κοινωνικά μέσα. Δεδομένου ότι τα tweets μπορούν να αναρτηθούν και να προσπελαστούν από ένα ευρύ φάσμα υπηρεσιών στο διαδίκτυο σε πραγματικό χρόνο, η διάδοση των πληροφοριών σε ένα μεγάλο κοινό έχει γίνει το επίκεντρο των εμπόρων, των κυβερνήσεων, και ακόμη και κακόβουλους αποστολείς ανεπιθύμητων μηνυμάτων. Σε αντίθεση με τις καμπάνιες σε παραδοσιακές πλατφόρμες

μέσων μαζικής ενημέρωσης, στα μέσα κοινωνικής δικτύωσης οι εκστρατείες συχνά επηρεάζουν τους ανθρώπους με κρυφό ή σιωπηρό τρόπο χωρίς να αποκαλύπτουν την πραγματική τους πρόθεση. Έτσι, οι αναγνώστες συχνά δεν γνωρίζουν ότι τα μηνύματα που βλέπουν είναι στρατηγικές που στοχεύουν στο να πείσουν τους χρήστες να αγοράσουν κάποια προϊόντα/υπηρεσίες-στόχους ή να αποδεχτούν κάποιες στοχευμένες ιδέες ή ιδεολογίες.

Μια μεγάλη κατηγορία μελετών αφορούν την εξόρυξη του προφίλ των χρηστών του Twitter και την ανίχνευση του τύπου προσωπικότητάς τους βάσει των δημοσιευμένων tweets τους, για διάφορα γεγονότα, προϊόντα ή/και υπηρεσίες. Η επιλογή των λέξεων, η χρήση των σημείων στίξης (συντακτικών και σχολιαστικών), αλλά και η επιλογή των emoticons για την σύνταξη των tweets από έναν χρήστη είναι δυνατόν να αποκαλύψει τόσο την ψυχολογική του κατάσταση εκείνη τη χρονική στιγμή, όσο και να οδηγήσει στην εξαγωγή πληροφοριών για τον χαρακτήρα και την προσωπικότητάς του. Το σύνολο αυτών των πληροφοριών αποτελεί χρήσιμο εργαλείο για τις κοινωνικές, πολιτικές και ιατρικές επιστήμες, το ερευνητικό φάσμα των οποίων καλύπτει θέματα όπως οι καταναλωτικές συνήθειες, η πολιτικές πεποιθήσεις, η ψυχολογία κ.α.

2.4.3. Η δημοτικότητα του Twitter

Ερευνητές της κοινότητας των Νέων Κοινωνικών Μέσων Νέα Κοινωνική Επιστήμη (New Social Media New Social Science - NSMNSS), έχουν καταλήξει στους λόγους για τους οποίους το Twitter έχει προσελκύσει περισσότερη ακαδημαϊκή έρευνα σε σύγκριση με άλλες πλατφόρμες κοινωνικών μέσων. Το δίκτυο NSMNSS είναι μια διαδικτυακή κοινότητα που ιδρύθηκε το 2011 προκειμένου να παρέχει ένα χώρο για αναστοχαστικές συζητήσεις σχετικά με τον τρόπο με τον οποίο η εργασία με νέες μορφές δεδομένων, συμπεριλαμβανομένων των δεδομένων των κοινωνικών μέσων, ήταν πιθανό να αμφισβητήσει τις συμβατικές προσεγγίσεις στην έρευνα των κοινωνικών επιστημών [26].

Υπάρχουν τουλάχιστον πέντε πιθανοί λόγοι για τη δημοτικότητα του Twitter στην ακαδημαϊκή έρευνα [27]:

1. Το Twitter API (application programming interface)[2], είναι πιο ανοιχτό και προσβάσιμο σε σύγκριση με άλλες πλατφόρμες κοινωνικών μέσων. Αυτό καθιστά το Twitter πιο ευνοϊκό για τους προγραμματιστές που δημιουργούν εργαλεία για πρόσβαση σε δεδομένα. Κατά συνέπεια, αυτό αυξάνει τη διαθεσιμότητα λογισμικού και διαδικτυακών εργαλείων στους ερευνητές. Τα δεδομένα του Facebook, συγκριτικά, είναι πολύ δύσκολο να αποκτηθούν και είναι διαθέσιμα μόνο σε συγκεντρωτικό επίπεδο για σκοπούς μάρκετινγκ.

2. Το Twitter διευκολύνει την εύρεση και την παρακολούθηση συνομιλιών, καθώς διαθέτει μια λειτουργία αναζήτησης που επιτρέπει στους χρήστες να αναζητούν tweets, ενώ και τα tweets εμφανίζονται επίσης στα αποτελέσματα αναζήτησης στο Google, γεγονός που διευκολύνει τον εντοπισμό τους. Το Facebook μπορεί να θεωρηθεί περισσότερο μια ιδιωτική πλατφόρμα και δεν εμφανίζονται όλες οι δημόσιες αναρτήσεις στα αποτελέσματα μιας μηχανής αναζήτησης. Το Facebook παρέχει επίσης στους χρήστες περισσότερα στοιχεία ελέγχου απορρήτου.

3. Το Twitter έχει μια ισχυρή κουλτούρα χρήσης του hashtag (#) που διευκολύνει τη συλλογή, την ταξινόμηση και την επέκταση των αναζητήσεων κατά τη συλλογή δεδομένων. Επομένως, τα δεδομένα του Twitter είναι πιο εύκολο να ανακτηθούν καθώς τα σημαντικά περιστατικά, οι ειδήσεις και τα γεγονότα στο Twitter τείνουν να επικεντρώνονται σε ένα hashtag. Το Facebook έχει δυνατότητα hashtag, ωστόσο, η χρήση των hashtags δεν φαίνεται να είναι τόσο διαδεδομένη όσο στο Twitter.

4. Το Twitter μπορεί να είναι μια δημοφιλής πλατφόρμα λόγω της προσοχής που μπορεί να λάβει από τα υπόλοιπα κοινά μέσα ενημέρωσης και μπορεί να προσελκύσει περισσότερη έρευνα λόγω της πολιτιστικής του θέσης. Το Twitter χρησιμοποιείται επίσης ευρέως από δημοσιογράφους, τόσο για τον εντοπισμό γεγονότων που αποτελούν ειδήσεις όσο και για τη διανομή έκτακτων ειδήσεων. Σε σύγκριση με το Facebook, το Twitter λαμβάνει πολύ περισσότερη προσοχή από τα μέσα ενημέρωσης επειδή διασημότητες, πολιτικοί και αθλητές δημοσιεύουν tweets

[2] Το API (Application Programming Interface) ή αλλιώς Διεπαφή Προγραμματισμού Εφαρμογών αναφέρεται στις εντολές (διεπαφή) των προγραμματιστικών διαδικασιών που παρέχει ένα λειτουργικό σύστημα, μια βιβλιοθήκη ή μία εφαρμογή έτσι ώστε να επιτρέπει από άλλα προγράμματα να κάνουν αιτήσεις προς αυτά για ανταλλαγή ή επεξεργασία δεδομένων. Ένα από τα κύρια πλεονεκτήματα της API είναι ότι επιτρέπει την άντληση πληροφοριών από το ένα σύστημα στο άλλο.
(https://www.ip.gr/el/dictionary/378-API___Application_Programming_Interface)

για τα τρέχοντα γεγονότα, ορισμένα από τα οποία μπορεί να είναι αμφιλεγόμενα και ως εκ τούτου αποτελούν ειδήσεις.

5. Πολλοί ερευνητές χρησιμοποιούν οι ίδιοι το Twitter και, λόγω των ευνοϊκών προσωπικών τους εμπειριών, μπορεί να αισθάνονται πιο άνετα όταν ερευνούν μια πιο οικεία πλατφόρμα.

3. Ανάλυση συναισθήματος (Sentiment Analysis)

3.1. Ορισμός

Ανάλυση Συναισθήματος (Sentiment Analysis) ή Εξόρυξη Γνώμης (Opinion Mining) είναι η υπολογιστική μελέτη των απόψεων (opinions), των συναισθημάτων (sentiments), των εκτιμήσεων (appraisals) και των στάσεων (attitudes) των ανθρώπων προς τις οντότητες και τα χαρακτηριστικά τους, όπως προϊόντα, υπηρεσίες, οργανισμούς, εκδηλώσεις, γεγονότα, άτομα και θέματα [28]. Ο όρος *ανάλυση συναισθήματος* εμφανίστηκε ίσως για πρώτη φορά στην εργασία των Nasukawa και Yi (2003)[29], ενώ ο όρος *εξόρυξη γνώμης* από τους Dave K, Lawrence S., Pennock D. (2003)[30]. Ωστόσο, η έρευνα στον τομέα αυτό ξεκίνησε νωρίτερα από τους Wiebe (2000)[31], Das και Chen (2001)[32], Tong (2001)[33], Morinaga et al.(2002)[34], Pang B., Lee L. και Vaithyanathan S.(2002)[35] και Turney (2002)[36]. Ακόμη νωρίτερα σχετικές εργασίες περιλάμβαναν ανάλυση μεταφορικών εκφράσεων, εξαγωγή επιθέτων με συναισθηματικό βάρος, και οι προηγούμενες σχετικές εργασίες περιλαμβάνουν την ερμηνεία και εξαγωγή συναισθημάτων επιθέτων, συναισθηματική υπολογιστική (affective computing), ανάλυση υποκειμενικότητας, απόψεων και επιπτώσεων [40] [37] [38] [39] [39].

Στόχος της ανάλυσης συναισθήματος είναι η κατάταξη των προς μελέτη κειμένων σε μία από τις καθοριζόμενες από το εκάστοτε πρόβλημα κατηγορίες συναισθήματος (π.χ θετικό, αρνητικό ή ουδέτερο συναίσθημα). Η ταχεία ανάπτυξη του πεδίου συμπίπτει με την ανάπτυξη των κοινωνικών μέσων, τα οποία παρέχουν την δυνατότητα συγκέντρωσης ενός τεράστιου όγκου ψηφιακών δεδομένων, σε μορφές όπως δημοσιεύσεις σε κοινωνικά δίκτυα, σχόλια, κριτικές, συζητήσεις σε forum κ.α. Αν και η ανάλυση των κοινωνικών δικτύων δεν είναι ένας νέος ερευνητικός τομέας, καθώς ξεκίνησε στις δεκαετίες του 1940 και του 1950, όταν οι ερευνητές της επιστήμης του management άρχισαν να μελετούν τους κοινωνικούς παράγοντες (ανθρώπους σε οργανισμούς) και τις σχέσεις μεταξύ τους, η έλευση των μέσων κοινωνικής δικτύωσης έχει τροφοδοτήσει την εκρηκτική ανάπτυξή της τα τελευταία είκοσι χρόνια. Η ανάλυση συναισθήματος, αντίθετα, είναι ένας νέος τομέας έρευνας που ουσιαστικά αναπτύχθηκε μέσα από τα κοινωνικά μέσα. Η ανάλυση συναισθήματος έχει εξελιχθεί τις τελευταίες δεκαετίας σε έναν από τους πιο δραστήριους τομείς της έρευνας στο αντικείμενο της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP), ενώ η μελέτη της συνδυάζεται συχνά με έρευνες σε άλλους τομείς όπως η Εξόρυξη Δεδομένων, η Μελέτη

Ψυχολογικού προφίλ (Profiling), η εξόρυξη Προφίλ Χρηστών (User Profiling) η Εξόρυξη Κειμένου κ.α. Τα τελευταία χρόνια, οι ερευνητές έχουν μελετήσει την πολυτροπική ανάλυση συναισθημάτων, η οποία χρησιμοποιεί πληροφορίες εικόνας ή/και βίντεο, κειμένου και ήχου για την ταξινόμηση των ανθρώπινων συναισθημάτων.

Η ανάλυση συναισθημάτων επικεντρώνεται κυρίως σε απόψεις που εκφράζουν ή υποδηλώνουν θετικά ή αρνητικά συναισθήματα. Οι προτάσεις που εκφράζουν απόψεις ή συναισθήματα είναι συνήθως υποκειμενικές προτάσεις, σε αντίθεση με τις αντικειμενικές προτάσεις, οι οποίες δηλώνουν γεγονότα, επειδή οι απόψεις και τα συναισθήματα είναι εγγενώς υποκειμενικά. Ωστόσο, οι αντικειμενικές προτάσεις μπορεί να υποδηλώνουν και τα θετικά ή αρνητικά συναισθήματα των συγγραφέων τους, επειδή μπορεί να περιγράφουν επιθυμητά ή ανεπιθύμητα γεγονότα. Η ανάλυση συναισθημάτων μελετά επίσης τέτοιες αντικειμενικές προτάσεις.

Η φύση των κοινωνικών μέσων επιτρέπει και ενθαρρύνει την αλληλεπίδραση μεταξύ των χρηστών, διαμορφώνοντας μια νέα συμμετοχική κουλτούρα και αναπτύσσοντας την ελεύθερη έκφραση και επικοινωνία των απόψεών τους για οποιοδήποτε θέμα, πέρα από γεωγραφικά και χωρικά όρια. Αυτή η συμμετοχική επανάσταση στον ιστό και τις επικοινωνίες έχει μεταμορφώσει τόσο την καθημερινότητά μας όσο και την κοινωνία στο σύνολό της.

Από το 2002, η έρευνα στην ανάλυση συναισθημάτων είναι πολύ ενεργή, καθώς εκτός από τη διαθεσιμότητα μεγάλου αριθμού δεδομένων στα μέσα κοινωνικής δικτύωσης, οι απόψεις και τα συναισθήματα ως βασικό χαρακτηριστικό των ανθρώπινων δραστηριοτήτων, συναντώνται σε ένα πολύ ευρύ φάσμα εφαρμογών. Κάθε απόφαση του σύγχρονου ανθρώπου βασίζεται συχνά στις απόψεις του κοινωνικού του περίγυρου. Το ίδιο ισχύει και για τους σύγχρονους οργανισμού και επιχειρήσεις. Οι εφαρμογές της ανάλυσης συναισθήματος αποτέλεσαν ισχυρό κίνητρο για την ανάπτυξη της έρευνας, αποκαλύπτοντας στην πορεία προκλήσεις και συναρπαστικά ερευνητικά προβλήματα.

Τα θέματα, οι εκδηλώσεις, τα γεγονότα και οι άνθρωποι που συζητούνται στα μέσα κοινωνικής δικτύωσης καθίστανται σημαντικά, καθώς αποτελούν σημαντικές πηγές πληροφοριών για εξαγωγής συναισθημάτων και απόψεων. Επιπλέον, η μελέτη μπορεί να εστιαστεί και στους ίδιους τους συμμετέχοντες. Υπάρχει η δυνατότητα δημιουργίας ενός προφίλ συναισθήματος κάθε συμμετέχοντα στα κοινωνικά μέσα, με βάση τις απόψεις του για τα επίκαιρα θέματα αλλά και τις αναρτήσεις των άλλων χρηστών, καθώς είναι αυτά που αντικατοπτρίζουν τη φύση

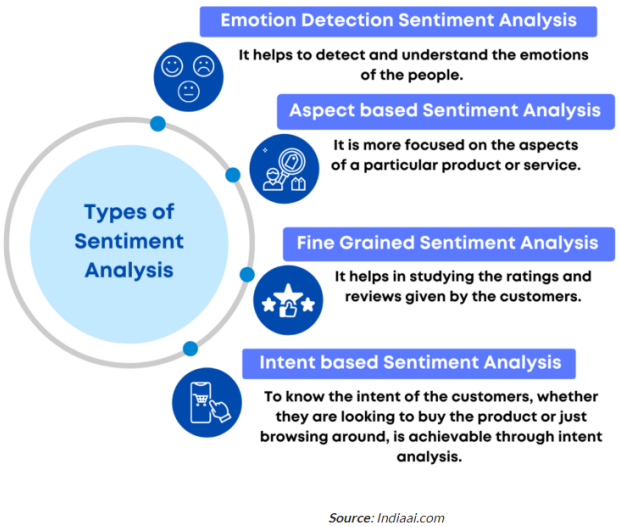
και τις προτιμήσεις του ατόμου. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν σε πολλές εφαρμογές όπως αναφέρονται αναλυτικά παρακάτω.

Αν και η ανάλυση συναισθημάτων προήλθε από την επιστήμη των υπολογιστών, τα τελευταία χρόνια, έχει εξαπλωθεί στις κοινωνικές και οικονομικές επιστήμες, καθώς και στις επιστήμες της οργάνωσης και διοίκησης επιχειρήσεων λόγω της σημασίας της για τις επιχειρήσεις και την κοινωνία στο σύνολό της, καθώς είναι αυτές που ασχολούνται τόσο με τον καταναλωτισμό, όσο και με την έκφραση της δημόσιας γνώμης. Αποτελεί γεγονός το ότι η ανάλυση συναισθημάτων με τη βοήθεια των μέσων κοινωνικής δικτύωσης μπορεί να αλλάξει ριζικά την κατεύθυνση της έρευνας και της πρακτικής σε αυτούς τους τομείς.

3.2. Τύποι ανάλυσης συναισθήματος

Η κατηγοριοποίηση της πολικότητας είναι ένα σημαντικό μέρος της ανάλυσης συναισθημάτων. Το συνολικό συναίσθημα που εκφράζεται από μια παράγραφο, φράση ή λέξη αναφέρεται ως πολικότητα. Αυτή η πολικότητα μπορεί να μετρηθεί χρησιμοποιώντας μια «βαθμολογία συναισθήματος», η οποία είναι μια αριθμητική βαθμολογία και μπορεί να υπολογιστεί για το πλήρες κείμενο ή για μία μόνο φράση.

Ανάλογα με τον τρόπο με τον οποίο επιδιώκετε η ερμηνεία των σχολίων και των ερωτήσεων των πελατών, μπορούν να οριστούν και να προσαρμοστούν οι κατηγορίες , ώστε να ταιριάζουν στις ανάγκες της ανάλυσης συναισθήματος (Εικόνα 2).



Εικόνα 2. Τύποι ανάλυσης συναισθήματος (Πηγή: Indiaai.com)

Στη συνέχεια, παρουσιάζονται μερικές από τις πιο κοινές μεθοδολογίες για την ανάλυση συναισθήματος:

- **Λεπτομερής ανάλυση συναισθήματος (Fine Grained Sentiment Analysis):** Διασπά την πολικότητα σε μικρότερες ομάδες, συνήθως εξαιρετικά θετικές έως πολύ αρνητικές, για να παρέχει ένα πιο συγκεκριμένο επίπεδο πολικότητας. Αυτό μπορεί να συγκριθεί με ένα σύστημα αξιολόγησης 5 αστέρων όσον αφορά τη γνώμη.
- **Ανάλυση συναισθήματος βάσει διαστάσεων (Aspect-based Sentiment Analysis ABSA):** Αυτή η ανάλυση είναι ιδιαίτερα χρήσιμη, όταν σχετίζεται με μια συγκεκριμένη ιδιότητα ή χαρακτηριστικό που περιγράφεται στο κείμενο. Η ABSA είναι η διαδικασία ανακάλυψης αυτών των τάσεων ή χαρακτηριστικών και του συναισθήματός τους. Αυτά τα χαρακτηριστικά αναφέρονται ως «θέματα» (themes) στο Θεματικό (Thematic).
- **Ανίχνευση συναισθημάτων (Emotion detection):** Αντί να ανιχνεύει θετικά και αρνητικά συναισθήματα, η ανίχνευση συναισθημάτων ανιχνεύει συγκεκριμένα συναισθήματα. Η ευτυχία, η απογοήτευση, το σοκ, ο θυμός και η θλίψη είναι μερικά παραδείγματα.
- **Βάσει πρόθεσης (Intent based Sentiment Analysis):** Η ανάλυση βάσει πρόθεσης κάνει διάκριση μεταξύ γεγονότων και απόψεων σε ένα κείμενο. Ένα διαδικτυακό σχόλιο που δείχνει δυσαρέσκεια για την αλλαγή μιας μπαταρίας, για παράδειγμα, μπορεί να παρακινήσει την εξυπηρέτηση πελατών να επικοινωνήσει με τον πελάτη ώστε να διορθώσει το πρόβλημα. [40]

3.3. Εφαρμογές της ανάλυσης συναισθήματος

Τα τελευταία χρόνια, οι εφαρμογές ανάλυσης συναισθημάτων έχουν εξαπλωθεί σχεδόν σε κάθε πιθανό τομέα, από τα καταναλωτικά προϊόντα, την υγειονομική περίθαλψη, τον τουρισμό, τη φιλοξενία και τις χρηματοπιστωτικές υπηρεσίες έως τις κοινωνικές εκδηλώσεις και τις εκλογές πολιτικών προσώπων.

Οι απόψεις των καταναλωτών για τα διάφορα προϊόντα είναι σημαντικές για τις επιχειρήσεις. Το ίδιο σημαντικές είναι και οι γνώμες του κοινού για τις υπηρεσίες που παρέχουν οι διάφοροι οργανισμοί. Οι τοπικές και ομοσπονδιακές κυβερνήσεις, θέλουν επίσης να καθορίσουν τις δημόσιες απόψεις σχετικά με τις υπάρχουσες ή προτεινόμενες πολιτικές τους. Οι γνώμες αυτές επιτρέπουν στους σχετικούς κυβερνητικούς φορείς να λάβουν αποφάσεις που ανταποκρίνονται άμεσα σε ένα ταχέως μεταβαλλόμενο κοινωνικά, οικονομικά και πολιτικά κλίμα. Στη διεθνή πολιτική σκηνή, κάθε κυβέρνηση παρακολουθεί τα μέσα κοινωνικής δικτύωσης

άλλων χωρών προκειμένου να διαπιστώσει τι συμβαίνει σε αυτές τις χώρες και ποιες είναι οι απόψεις και τα συναισθήματα των ανθρώπων για τα τρέχοντα τοπικά και διεθνή ζητήματα και γεγονότα. Αυτές οι πληροφορίες είναι πολύ χρήσιμες για τη διπλωματία, τις διεθνείς σχέσεις και τη λήψη οικονομικών αποφάσεων. Εκτός από τις επιχειρήσεις, τους οργανισμούς και τις κυβερνητικές υπηρεσίες, οι μεμονωμένοι καταναλωτές θέλουν να γνωρίζουν τις απόψεις άλλων σχετικά με τα προϊόντα, τις υπηρεσίες και τους πολιτικούς υποψηφίους πριν αγοράσουν τα προϊόντα, χρησιμοποιήσουν τις υπηρεσίες και λάβουν εκλογικές αποφάσεις. Η συλλογή και η ανάλυση των απόψεων του κοινού και των καταναλωτών είναι εδώ και πολύ καιρό μια τεράστια επιχείρηση για το μάρκετινγκ, τις δημόσιες σχέσεις και τις εταιρείες πολιτικών εκστρατειών.

Σήμερα, τα άτομα, οι οργανισμοί και οι κυβερνητικές υπηρεσίες χρησιμοποιούν όλο και περισσότερο το περιεχόμενο των μέσων κοινωνικής δικτύωσης για τη λήψη αποφάσεων. Εάν ένα άτομο θέλει να αγοράσει ένα καταναλωτικό προϊόν, δεν περιορίζεται στην γνώμη των συγγενών και φίλων, αλλά αναζητά κριτικές χρηστών και συζητήσεις σε δημόσια φόρουμ στο διαδίκτυο σχετικά με το προϊόν. Για έναν οργανισμό, ενδέχεται να μην είναι πλέον απαραίτητο να διεξάγει έρευνες, δημοσκοπήσεις ή ομάδες εστίασης για να συγκεντρώσει δημόσιες ή καταναλωτικές απόψεις σχετικά με τα προϊόντα και τις υπηρεσίες του οργανισμού, επειδή μια πληθώρα τέτοιων πληροφοριών είναι άμεσα διαθέσιμη μέσω του διαδικτύου.

Οι εφαρμογές είναι επίσης ευρέως διαδεδομένες στις κυβερνητικές υπηρεσίες, οι οποίες παρακολουθούν τα μέσα κοινωνικής δικτύωσης για να ανακαλύψουν τα δημόσια συναισθήματα και τις ανησυχίες των πολιτών.

Πέρα από την εφαρμογή της ανάλυσης συναισθήματος στους διαφορετικούς τομείς της καθημερινότητας, έχουν δημοσιευθεί και πολλές ερευνητικές εργασίες που άπτονται μιας ποικιλίας άλλων εφαρμογών. Για παράδειγμα, αρκετοί ερευνητές έχουν χρησιμοποιήσει πληροφορίες συναισθημάτων για να προβλέψουν την επιτυχία των ταινιών και τα έσοδα από το box-office. Οι Mishne and Glance (2006) έδειξαν ότι το θετικό συναίσθημα είναι ο καλύτερος προγνωστικός παράγοντας της επιτυχίας μιας ταινίας από ότι οι φήμες που κυκλοφορούν[41]. Οι Sadikov, Parameswaran και Venetis (2009) έκαναν την ίδια πρόβλεψη χρησιμοποιώντας συναισθήματα και άλλα χαρακτηριστικά [42]. Ο Liu et al. (2007) ανέφερε ένα μοντέλο ανάλυσης συναισθήματος για την πρόβλεψη των εσόδων box-office που αποτελείται από δύο βήματα [43]. Στο πρώτο βήμα χτίζεται ένα μοντέλο θέματος

βασισμένο σε πιθανή λανθάνουσα σημασιολογική ανάλυση (PLSA) (Hofmann, 1999)[44] χρησιμοποιώντας μόνο λέξεις συναισθήματος (ή λέξεις γνώμης) σε ένα σύνολο κριτικών ταινιών. Οι λέξεις συναισθήματος υποδεικνύουν επιθυμητές ή ανεπιθύμητες καταστάσεις. Για παράδειγμα, οι λέξεις «καλό», «μεγάλο» και «όμορφο» είναι θετικές λέξεις συναισθήματος, και το «κακό», «απαίσιο» και «άσχημο» είναι λέξεις αρνητικού συναισθήματος. Στο δεύτερο βήμα δημιουργείται ένα μοντέλο αυτοπαλινδρόμησης που χρησιμοποιεί τόσο τα έσοδα όσο και τις καταγραφές των συναισθημάτων των τελευταίων ημερών για να προβλέψει τα μελλοντικά έσοδα. Το ίδιο πρόβλημα πρόβλεψης εσόδων αντιμετωπίστηκε επίσης από τους Asur και Huberman (2010) αναλύοντας τις συναισθηματικές λέξεις που περιέχονταν σε έναν μεγάλο αριθμό tweets [45] . Μια άλλη προσέγγιση, βασισμένη στην γραμμική παλινδρόμηση, χρησιμοποιώντας κριτικές και μετα-δεδομένα ταινιών δόθηκε από τους Joshi et al. (2010)[46].

Ένας άλλος τομέας εφαρμογής της ανάλυση συναισθήματος είναι εκλογές πολιτικών κομμάτων. Για παράδειγμα, οι O'Connor et al. (2010) καταμετρώντας λέξεις με θετικό και αρνητικό συναίσθημα δημιούργησαν μια βαθμολογία συναισθημάτων, η οποία αποδείχθηκε ότι συσχετίζεται καλά με την προεδρική αποδοχή, τις πολιτικές εκλογικές δημοσκοπήσεις και τις έρευνες εμπιστοσύνης των καταναλωτών [47]. Οι Bermingham και Smeaton (2011) χρησιμοποίησαν θετικά και αρνητικά tweets ως ανεξάρτητες μεταβλητές και αποτελέσματα δημοσκοπήσεων ως τιμές για την εξαρτώμενη μεταβλητή, ώστε να καταρτίσουν ένα μοντέλο γραμμική παλινδρόμησης για την πρόβλεψη των αποτελεσμάτων των Ιρλανδικών εκλογών το 2011[48]. Οι Chung και Mustafaraj (2011)[49] και οι Gayo-Avello et al. (2011)[50] συζήτησαν αρκετούς περιορισμούς των τρεχουσών εργασιών σχετικά με τη χρήση δεδομένων twitter για την πρόβλεψη πολιτικών εκλογών, ένας από τους οποίους είναι η κακή ακρίβεια ανάλυσης συναισθημάτων. Στην εργασία τους οι Διακόπουλος και Shamma (2010)[51] και Sang και Bos (2012)[52] χρησιμοποίησαν χειροκίνητα κατηγοριοποιημένα συναισθήματα των tweets για την πρόβλεψη εκλογικού αποτελέσματος. Οι Tumasjan et al. (2010) έδειξαν ότι ακόμη και οι απλές αναφορές για ένα πολιτικό κόμμα στο Twitter μπορεί να είναι ένας καλός προγνωστικός παράγοντας των εκλογικών αποτελεσμάτων[53]. Σε άλλα σχετικά έργα, οι Yano και Smith (2010) ανέφεραν μια μέθοδο για την πρόβλεψη του όγκου των σχολίων των πολιτικών ιστολογίων [54] , οι Chen et al. (2010) μελέτησαν τις πολιτικές απόψεις [55]και οι Khoo

et al. (2012) ανέλυσαν το συναίσθημα σε άρθρα πολιτικών ειδήσεων σχετικά με τις οικονομικές πολιτικές και τα πολιτικά πρόσωπα [56].

Μια άλλη δημοφιλής περιοχή εφαρμογής είναι οι προβλέψεις του χρηματιστηρίου. Οι Das και Chen (2007) προσδιόρισαν τις γνώμες από δημοσιευμένα μηνύματα κατατάσσοντας κάθε θέση σε μία από τις τρεις κατηγορίες συναισθημάτων: bullish (αισιόδοξη), bearish (απαισιόδοξη) ή ουδέτερη (ούτε bullish ούτε bearish). Τα συναισθήματα που προέκυψαν σε όλες τις μετοχές στη συνέχεια συγκεντρώθηκαν και χρησιμοποιήθηκαν για την πρόβλεψη του Δείκτη Υψηλής Τεχνολογίας Morgan Stanley [32]. Οι Zhang et al. (2010c) εντόπισαν θετικές και αρνητικές δημόσιες διαθέσεις στο Twitter και τις χρησιμοποίησε για να προβλέψει την κίνηση των δεικτών χρηματιστηρίου, όπως ο βιομηχανικός μέσος όρος Dow Jones (DJIA), S&p 500, και NASDAQ. Έδειξαν ότι όταν τα συναισθήματα στο Twitter είναι αρνητικά - δηλαδή, όταν οι άνθρωποι εκφράζουν μικρή ελπίδα, φόβο και ανησυχία – ο δείκτης Dow Jones κατεβαίνει την επόμενη μέρα, ενώ όταν εκφράζουν ελπίδα, λιγότερο φόβο και ανησυχία, ο δείκτης ανεβαίνει [57]. Ομοίως, οι Bollen et al. (2011) χρησιμοποίησαν δεδομένα του Twitter για να προβλέψει την κίνηση του δείκτη DJIA [58]. Οι Zhang και Skiena (2010) χρησιμοποίησαν δεδομένα συναισθημάτων σε ιστολόγια και ιστότοπους ειδήσεων για να σχεδιάσουν στρατηγικές συναλλαγών[59].

Εκτός από την έρευνα στους παραπάνω τομείς, έχουν επίσης δημοσιευθεί πολυάριθμες εργασίες σχετικά με τη χρήση ανάλυσης συναισθημάτων σε άλλους τύπους εφαρμογών. Για παράδειγμα, στην εργασία τους οι McGlohon et al. (2010), χρησιμοποίησαν τις κριτικές των πελατών για την δημιουργία κατάταξης προϊόντων και εμπόρων [60]. Οι Hong και Skiena (2010), μελέτησαν τις σχέσεις μεταξύ της στοιχηματικής γραμμής της Εθνικής Ένωσης Ποδοσφαίρου και των απόψεων του κοινού σε ιστολόγια και στο Twitter [61]. Οι Miller et al. (2011), διερεύνησαν τη ροή συναισθημάτων στα κοινωνικά δίκτυα [62]. Οι Sakunkoo P. και Sakunkoo N. (2009), μελέτησαν τις κοινωνικές επιρροές των διαδικτυακών κριτικών βιβλίων [63] και οι Groh και Hauffa (2011), χρησιμοποίησαν την ανάλυση συναισθημάτων για τον χαρακτηρισμό των κοινωνικών σχέσεων[64].

3.4. Έρευνα στην Ανάλυση Συναισθήματος

Το μεγάλο φάσμα εφαρμογών της ανάλυσης συναισθήματος παρέχει ισχυρά κίνητρα για περαιτέρω έρευνα, αλλά και για την προσπάθεια επίλυσης δύσκολων τεχνικών προβλημάτων που προκύπτουν τόσο κατά την εφαρμογή της, όσο και κατά την ερμηνεία

των αποτελεσμάτων που προκύπτουν από αυτή. Από το 2000 και έπειτα, ο τομέας της ανάλυσης συναισθήματος είναι ένας από τους πιο ενεργούς τομείς έρευνας στο NLP (Natural Language Processing), την εξόρυξη δεδομένων και την εξόρυξη ιστού και μελετάται επίσης ευρέως στις επιστήμες διοίκησης και διαχείρισης επιχειρήσεων [65]–[70]. Αν και η ανάλυση συναισθημάτων έχει μελετηθεί σε διαφορετικούς κλάδους, η εστίασή τους δεν είναι η ίδια. Για παράδειγμα, στις επιστήμες της διοίκησης και διαχείρισης επιχειρήσεων, ο κύριος στόχος είναι ο αντίκτυπος των απόψεων των καταναλωτών στις επιχειρήσεις και οι τρόποι αξιοποίησης αυτών των απόψεων για την ενίσχυση των επιχειρηματικών πρακτικών. Αντίθετα, για το NLP και την εξόρυξη δεδομένων, ο στόχος είναι ο σχεδιασμός αποτελεσματικών αλγορίθμων και μοντέλων για την εξαγωγή απόψεων από ένα κείμενο φυσικής γλώσσας και την περαιτέρω κατάλληλη σύνοψή του με σκοπό την εξαγωγή συμπερασμάτων. Όσον αφορά την κατανόηση της φυσικής γλώσσας, η ανάλυση συναισθημάτων μπορεί να θεωρηθεί υπο-πεδίο της σημασιολογικής ανάλυσης, επειδή στόχος της είναι να αναγνωρίζει τόσο τα θέματα για τα οποία μιλούν οι άνθρωποι, όσο και τα συναισθήματά τους γι' αυτά.

3.5. Επίπεδα Ανάλυσης

Η έρευνα της ανάλυσης συναισθήματος πραγματοποιείται κυρίως σε τρία επίπεδα λεπτομέρειας, σε επίπεδο εγγράφου (document level), σε επίπεδο πρότασης (sentence level, και σε επίπεδο πτυχής (ή οπτικής) (aspect level).

1. **Επίπεδο εγγράφου:** Στόχος της ανάλυσης σε επίπεδο εγγράφου είναι να ταξινομεί ολόκληρα έγραφα ανάλογα με το θετικό ή αρνητικό συναίσθημα που εκφράζεται μέσα σε αυτά [38] [36]. Η ταξινόμηση αυτή είναι γνωστή ως ταξινόμηση συναισθημάτων σε επίπεδο εγγράφου (document-level sentiment classification). Για παράδειγμα, δεδομένης μιας κριτικής ενός προϊόντος, το σύστημα καθορίζει εάν αυτή εκφράζει συνολικά θετική ή αρνητική γνώμη για το συγκεκριμένο προϊόν. Αυτό το επίπεδο ανάλυσης υποθέτει ότι κάθε έγγραφο εκφράζει απόψεις για μία μόνο οντότητα (π.χ. ένα μόνο προϊόν ή υπηρεσία). Κατά συνέπεια, δεν εφαρμόζεται σε έγγραφα που αξιολογούν ή συγκρίνουν πολλαπλές οντότητες, για τις οποίες απαιτείται πιο λεπτομερή ανάλυση.
2. **Επίπεδο πρότασης:** Το επόμενο επίπεδο είναι η ταξινόμηση μιας πρότασης, ανάλογα με το αν εκφράζει θετική, αρνητική ή ουδέτερη γνώμη (ή καμία γνώμη). Αυτό το επίπεδο ανάλυσης συνδέεται στενά με την ταξινόμηση βάσει της υποκειμενικότητας (subjectivity classification) [39], που διακρίνει τις προτάσεις που εκφράζουν

πραγματικές πληροφορίες (αντικειμενικές προτάσεις) από προτάσεις που εκφράζουν υποκειμενικές απόψεις (υποκειμενικές προτάσεις). Ωστόσο, η υποκειμενικότητα δεν είναι ισοδύναμη με το συναίσθημα ή τη γνώμη, καθώς πολλές αντικειμενικές προτάσεις μπορεί να υποδηλώνουν συναισθήματα ή απόψεις. Αντίθετα, πολλές υποκειμενικές προτάσεις μπορεί να μην εκφράζουν καμία γνώμη ή συναίσθημα.

3. **Επίπεδο πτυχής (ή οπτικής):** Παρά την χρησιμότητά τους σε συγκεκριμένες εφαρμογές, οι αναλύσεις σε επίπεδο εγγράφου και σε επίπεδο πρότασης δεν δίνουν σαφή απάντηση για το τι αρέσει και τι ακριβώς αντιπαθούν οι άνθρωποι. Για παράδειγμα, αν γνωρίζουμε μόνο ότι η πρόταση «Μου αρέσει το iPhone 5» είναι θετική, είναι περιορισμένης χρήσης εκτός αν γνωρίζουμε ότι η θετική γνώμη αφορά το iPhone 5. Επίσης, σε περίπτωση που μια πρόταση χαρακτηριστεί θετική ως προς το συναίσθημα, σημαίνει ότι όλα τα τμήματα που την συνθέτουν εκφράζουν εξίσου θετική γνώμη. Ωστόσο, αυτό δεν είναι ακριβές, καθώς μια πρόταση μπορεί να έχει πολλαπλές εκφάνσεις. Για παράδειγμα, «Η Apple τα πάει πολύ καλά σε αυτή την φτωχή οικονομία». Δεν έχει νόημα να χαρακτηρίσουμε αυτή την πρόταση θετική ή αρνητική επειδή είναι θετική για την Apple, αλλά αρνητική για την οικονομία. Για την επίτευξη αυτού του επιπέδου ακρίβειας αποτελεσμάτων, πραγματοποιείται ανάλυση σε επίπεδο πτυχής (ή οπτικής) (aspect-based sentiment analysis). Αντί για την εξέταση γλωσσικών μονάδων (έγγραφα, παραγράφους, προτάσεις, υποπροτάσεις ή φράσεις), η ανάλυση σε επίπεδο πτυχής εξετάζει άμεσα μια γνώμη και τον στόχο της (που ονομάζεται στόχος γνώμης – opinion target). Η συνειδητοποίηση της σημασίας των στόχων γνώμης μας επιτρέπει να κατανοήσουμε πολύ καλύτερα το πρόβλημα της ανάλυσης συναισθήματος. Για την καλύτερη κατανόηση του επιπέδου πτυχής, έστω ότι έχουμε το εξής παράδειγμα πρότασης: «Αν και η εξυπηρέτηση είναι κακή, εξακολουθώ να αγαπώ αυτό το εστιατόριο». Αυτή η πρόταση έχει σαφώς θετικό τόνο, αλλά δεν μπορούμε να πούμε ότι είναι απολύτως θετική. Μπορούμε μόνο να πούμε ότι η πρόταση είναι θετική για το εστιατόριο (τονίστηκε), αλλά εξακολουθεί να είναι αρνητική για την εξυπηρέτησή του (δεν τονίζεται). Εάν κάποιος που διαβάζει αυτή τη γνώμη ενδιαφέρεται πολύ για την εξυπηρέτηση, πιθανότατα δεν θα προτιμήσει το συγκεκριμένο εστιατόριο. Σε τέτοιες εφαρμογές, οι στόχοι γνώμης (π.χ. το εστιατόριο και η υπηρεσία στην προηγούμενη πρόταση) περιγράφονται συχνά από τις οντότητες (π.χ. εστιατόριο) ή/και τις διάφορες πτυχές τους (π.χ. εξυπηρέτηση του εστιατορίου). Έτσι, ο στόχος αυτού του επιπέδου ανάλυσης είναι να ανακαλύψει συναισθήματα σχετικά με τις οντότητες ή/και τις πτυχές τους. Με βάση αυτό το επίπεδο ανάλυσης,

μπορεί να συνταχθεί σύνοψη των απόψεων σχετικά με τις οντότητες και τις πτυχές τους. Να σημειωθεί ότι σε ορισμένες εφαρμογές, ο χρήστης μπορεί να ενδιαφέρεται μόνο για απόψεις σχετικά με τις οντότητες. Σε αυτή την περίπτωση, το σύστημα μπορεί απλά να αγνοήσει τις πτυχές του. Η ανάλυση σε επίπεδο πτυχής είναι αυτό που απαιτείται σχεδόν σε όλα τα συστήματα ανάλυσης συναισθήματος στην πραγματική ζωή.

Εκτός από τα διαφορετικά επίπεδα ανάλυσης, υπάρχουν και δύο διαφορετικοί τύποι απόψεων: οι κανονικές γνώμες (regular opinions) και οι συγκριτικές γνώμες (comparative opinions) [71].

1. Μια κανονική γνώμη εκφράζει ένα συναίσθημα για μια συγκεκριμένη οντότητα ή μια πτυχή της οντότητας. Για παράδειγμα, η πρόταση «Η Coca Cola έχει πολύ καλή γεύση» εκφράζει ένα θετικό συναίσθημα ή άποψη για την γεύση της Coca Cola. Αυτός είναι ο πιο κοινός τύπος γνώμης.
2. Μια συγκριτική γνώμη συγκρίνει πολλαπλές οντότητες με βάση ορισμένες από τις κοινές πτυχές τους. Για παράδειγμα, η πρόταση «Η Coca Cola έχει καλύτερη γεύση από την Pepsi» περιέχει σύγκριση της Coca Cola και της Pepsi με βάση την γεύση τους (μια πτυχή) και εκφράζει μια προτίμηση για την Coca Cola

3.6. Μέθοδοι Ανάλυσης Συναισθήματος

Υπάρχουν δύο βασικοί τύποι μεθόδων ανάλυσης συναισθήματος. Ο πρώτος βασίζεται στην μηχανική μάθηση (machine-learning based) και ο δεύτερος στην χρήση λεξικών (lexicon-based).

A. Μέθοδοι Μηχανικής Μάθησης

Οι μέθοδοι μηχανικής μάθησης συχνά βασίζονται σε τεχνικές εποπτευόμενης ταξινόμησης (supervised classification), όπου η ανίχνευση συναισθημάτων οδηγεί σε δυαδικό αποτέλεσμα (θετικό ή αρνητικό συναίσθημα). Η προσέγγιση αυτή απαιτεί δεδομένα με σήμανση (labeled data) για την ανάπτυξη και εκπαίδευση των ταξινομητών [35]. Οι ταξινομητές ταξινομούν τα κείμενα ανάλογα με το συνολικό θετικό ή αρνητικό συναίσθημα που εκφράζουν. Οι πιο συχνά χρησιμοποιούμενες μέθοδοι μηχανικής μάθησης στην ανάλυση συναισθήματος είναι η μηχανή διανυσμάτων υποστήριξης (Support Vector Machine - SVM) [72][73] και η μέθοδος Naïve Bayes [74]. Οι Wang και Manning [75] διαπίστωσαν ότι η μέθοδος Naïve Bayes ήταν πιο αποτελεσματική για αποσπάσματα ή σύντομες κριτικές, ενώ η SVM ήταν πιο αποτελεσματική για μεγαλύτερα έγγραφα ή εκτενείς κριτικές. Το βασικό πλεονέκτημα των μεθόδων που βασίζονται στη εκπαίδευση-μάθηση, είναι η ικανότητά τους να προσαρμόζουν και να δημιουργούν εκπαιδευμένα μοντέλα για συγκεκριμένους σκοπούς, πλαίσια ή/και εφαρμογές. Η έννοια των λέξεων και το συναίσθημα που αντικατοπτρίζουν εξαρτάται σε κάποιο βαθμό από τον τομέα και το πλαίσιο στο οποίο συναντώνται [76] [77]. Βασικό πλεονέκτημα των μεθόδων μηχανικής μάθησης είναι ότι ο ρόλος κάθε λέξης στη διαδικασία κατηγοριοποίησης συναισθημάτων προσαρμόζεται στο σώμα και την εφαρμογή. Το μειονέκτημα των μεθόδων μηχανικής μάθησης, είναι η διαθεσιμότητα δεδομένων με σήμανση (ετικέτα) (labeled data) για την ανάπτυξη του ταξινομητή και, ως εκ τούτου, η χαμηλή δυνατότητα εφαρμογής της μεθόδου σε νέα δεδομένα. Αυτό συμβαίνει επειδή η επισήμανση δεδομένων μπορεί να είναι δαπανηρή ή ακόμη και απαγορευτική για ορισμένες εργασίες. Ωστόσο, με τον πολλαπλασιασμό των ιστότοπων που φιλοξενούν αξιολογήσεις προϊόντων, συνοδευόμενα από σχόλια χρηστών και κριτικές, έχει αυξηθεί σημαντικά ο αριθμός των πηγών από τις οποίες μπορούν να αντληθούν δεδομένα με σήμανση. Στις περιπτώσεις που δεν είναι διαθέσιμος ένας σημαντικά μεγάλος όγκος τέτοιων δεδομένων, ένα υπάρχον λεξικό συναισθήματος είναι απαραίτητο για την κατηγοριοποίηση των συναισθημάτων.

Β. Μέθοδοι με τη χρήση λεξικών

Οι μέθοδοι που βασίζονται σε λεξικά (stentiment lexicon) κάνουν χρήση μιας προκαθορισμένης λίστας λέξεων, όπου κάθε λέξη μπορεί να συσχετίζεται με ένα συγκεκριμένο συναίσθημα, ή/και να επισημαίνεται ως θετική, αρνητική ή ουδέτερη, βάσει μιας προκαθορισμένης τιμής που αντικατοπτρίζει τη ισχύ ή την ένταση του συναισθήματος. Η ανάπτυξη του λεξικού μπορεί να γίνει είτε με χειροκίνητο τρόπο[78], [79], είτε με την χρήση αυτόματων συσχετισμών λέξεων [37] [80], είτε ημιαυτόματα αντλώντας τιμές συναισθημάτων από πηγές όπως το [76]. Για την πρόβλεψη του συνολικού συναισθήματος ενός κειμένου, απαιτείται η χρήση κατάλληλου αλγορίθμου που συγκεντρώνει τις τιμές των συναισθημάτων των μεμονωμένων λέξεων του κειμένου. Οι βασισμένοι σε λεξικά μέθοδοι ποικίλλουν ανάλογα με το πλαίσιο στο οποίο δημιουργήθηκαν. Για παράδειγμα, το LIWC [81] προτάθηκε αρχικά για την ανάλυση συναισθηματικών μοτίβων σε επίσημα γραπτά αγγλικά κείμενα, ενώ *τα λεξικά* PANAS-t [82] και POMS-ex [83] προτάθηκαν ως ψυχομετρικές κλίμακες προσαρμοσμένες στο περιεχόμενο του Διαδικτύου. Παρόλο που οι λεκτικές μέθοδοι δεν βασίζονται σε δεδομένα με σήμανση, είναι δύσκολο να δημιουργηθεί ένα μοναδικό λεξικό που θα μπορούσε να χρησιμοποιηθεί σε διαφορετικά περιβάλλοντα και εφαρμογές. Για παράδειγμα, η αργκό είναι κοινή στα μέσα κοινωνικής δικτύωσης, αλλά σπάνια υποστηρίζεται από λεκτικές [84]

Στα μειονεκτήματα της εφαρμογής των μεθόδων ανάλυσης συναισθήματος με την χρήση λεξικών συγκαταλέγονται τα ακόλουθα:

1. Μια θετική ή αρνητική λέξη συναισθημάτων μπορεί να έχει αντίθετους προσανατολισμούς ή πολικότητες σε διαφορετικούς τομείς εφαρμογών ή περιεχόμενο προτάσεων. Με τον προσανατολισμό ή την πολικότητα, εννοούμε αν ένα συναίσθημα ή μια γνώμη είναι θετική, αρνητική ή ουδέτερη. Για παράδειγμα, η λέξη «φοβερός» υποδεικνύει αρνητικό συναίσθημα, όπως για παράδειγμα στην πρόταση «Έχω ένα φοβερό πονοκέφαλο», αλλά μπορεί επίσης να υποδηλώνει θετικό συναίσθημα, όπως για παράδειγμα, «Η ταινία που είδαμε εχθές ήταν φοβερή». Η εξάρτηση της πολικότητας από τον τομέα εφαρμογής ή το περιεχόμενο της πρότασης θέτει περιορισμούς στην αξιοπιστία της μεθόδου.

2. Μια πρόταση που περιέχει συναισθηματικές λέξεις θα μπορούσε να μην εκφράζει κάποιο συναίσθημα. Το φαινόμενο αυτό παρατηρείται σε διάφορους τύπους προτάσεων, όπως οι ερωτήσεις (ανακριτικές) και οι προτάσεις υπό όρους. Για παράδειγμα, η έκφραση «Μπορείτε να μου πείτε ποιο κινητό της Samsung είναι καλό;» και η έκφραση «Αν το

κινητό που θα μου προτείνεται έχει καλή κάμερα, θα το αγοράσω». Και οι δύο προτάσεις περιέχουν τη συναισθηματική λέξη καλό/ή, αλλά καμία δεν εκφράζει θετική ή αρνητική γνώμη για κάποιο συγκεκριμένο κινητό. Ωστόσο, αυτό δεν σημαίνει ότι όλες οι ερωτήσεις ή υπό όρους προτάσεις δεν εκφράζουν γνώμη ή συναίσθημα. Για παράδειγμα, η πρόταση «Ξέρει κανείς ποιος μπορεί να επισκευάσει αυτό το άθλιο κινητό;" και η πρόταση «Αν ψάχνετε για ένα καλό αυτοκίνητο, πάρτε ένα Tayota».

3. Οι σαρκαστικές προτάσεις με ή χωρίς συναισθηματικά λόγια είναι δύσκολο να ερμηνευτούν σωστά, όπως για παράδειγμα, η έκφραση «Τι υπέροχο πλυντήριο! Σταμάτησε να λειτουργεί σε μόλις δύο μέρες.» Ο σαρκασμός δεν είναι τόσο συνηθισμένος στις κριτικές των καταναλωτών σχετικά με τα προϊόντα και τις υπηρεσίες, αλλά είναι κοινός στις πολιτικές συζητήσεις.

4. Πολλές προτάσεις χωρίς συναισθηματικές λέξεις μπορεί να υποδηλώνουν θετικά ή αρνητικά συναισθήματα ή απόψεις των συγγραφέων τους. Για παράδειγμα, η πρόταση «Αυτό το πλυντήριο χρησιμοποιεί πολύ νερό» υποδηλώνει αρνητική γνώμη για το πλυντήριο επειδή χρησιμοποιεί πολλούς πόρους (νερό). Πολλές τέτοιες προτάσεις είναι στην πραγματικότητα αντικειμενικές προτάσεις που εκφράζουν ορισμένες πραγματικές πληροφορίες. Για παράδειγμα, η πρόταση «Αρκούν δύο μέρες ύπνου σε αυτό το στρώμα, για να σχηματιστεί μία μόνιμη κοιλάδα» εκφράζει αρνητική άποψη για την ποιότητα του στρώματος. Αυτή η πρόταση μπορεί να θεωρηθεί αντικειμενική επειδή δηλώνει ένα γεγονός, αν και η λέξη «κοιλάδα» χρησιμοποιείται ως μεταφορά εδώ, αντικαθιστώντας την λέξη «βαθούλωμα». Όπως μπορούμε να δούμε, αυτές οι δύο προτάσεις δεν περιέχουν λέξεις συναισθήματος, αλλά και οι δύο εκφράζουν κάτι ανεπιθύμητο και άρα δείχνουν αρνητικές απόψεις.

Όλα αυτά τα ζητήματα παρουσιάζουν σημαντικές προκλήσεις και στην πραγματικότητα, είναι μόνο μερικά από τα μειονεκτήματα που καλούμαστε να αντιμετωπίσουμε.

Η ανάγκη για μεγαλύτερη ακρίβεια και συνολικά καλύτερα αποτελέσματα, οδήγησε στην ανάπτυξη πολλαπλών μεθόδων ανάλυσης και μέτρησης συναισθημάτων. Κάθε μία από αυτές τις μεθόδους εμφανίζει περιορισμούς, πλεονεκτήματα και μειονεκτήματα σε σχέση με τις υπόλοιπες. Μια προσπάθεια σύγκρισης των πιο διαδεδομένων μεθόδων ανάλυσης συναισθήματος σε μηνύματα των κοινωνικών μέσων πραγματοποιήθηκε το 2013 από τους Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto και Meeyoung Cha [85]. Οι προς σύγκριση μέθοδοι ανάλυσης συναισθήματος ήταν οι: SentiWordNet, SASA, PANAS-t, Emoticons, SentiStrength, LIWC, SenticNet και

Happiness Index. Η μελέτη επικεντρώθηκε στην ανίχνευση της πολικότητας του περιεχομένου 1.7 δισεκατομμυρίων tweets (δηλαδή, θετικό ή αρνητικό συναίσθημα), χωρίς να εξετάζει άλλους τύπους συναισθημάτων (π.χ. ψυχολογικές διαδικασίες όπως ο θυμός ή η ηρεμία). Τα αποτελέσματα έδειξαν ότι καμία μεμονωμένη μέθοδος δεν είναι πάντα καλύτερη από κάποια άλλη σε διαφορετικά σενάρια και πηγές κειμένου. Ωστόσο σύμφωνα με την μελέτη, οι μέθοδοι μηχανικής μάθησης (Naïve Bayes, Maximum Entropy, and SVM – Support-vector machine) είναι πιο κατάλληλες για την ανάλυση συναισθήματος σε tweets, απ' ότι οι λεκτικές μέθοδοι όπως η LIWC [86].

Στην συνέχεια παρατίθενται τα πιο γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά, καθώς και η μεθοδολογία που ακολουθείται σε καθένα από αυτά.

a. SentiWordNet

Το SentiWordNet, είναι ένα λεξικό συναισθήματος που προέκυψε από τον εμπλουτισμό του αγγλικού λεξικού WordNet με πληροφορίες συναισθήματος [87], από τους Esuli και Sebastiani [76] και αποτελείται από περισσότερες από 38000 πολικές και αρκετές άλλες αυστηρά αντικειμενικές λέξεις. Το λεξικό WordNet ομαδοποιεί επίθετα, ουσιαστικά, ρήματα και άλλα μέρη του λόγου σε συνώνυμα σύνολα που ονομάζονται synsets. Κάθε synset έχει τρεις βαθμολογίες, μία θετική, μία αρνητική και μία ουδέτερη. Το SentiWordNet συσχετίζει αυτές τις τρεις βαθμολογίες με το synset από το λεξικό WordNet για να υποδείξει το συναίσθημα ενός κειμένου ως θετικό (positive), αρνητικό (negative) ή αντικειμενικό-ουδέτερο (objective-neutral) (ουδέτερο). Οι βαθμολογίες, οι οποίες παίρνουν τιμές στο διάστημα [0, 1], και δίνουν μαζί άθροισμα ίσο με την μονάδα, υπολογίζονται με την χρήση ημι-εποπτευόμενης μεθόδου μηχανικής μάθησης (SVM και Rocchio). Έτσι, στο SentiWordNet, το συναίσθημα συνδέεται με την έννοια μιας λέξης και όχι με την ίδια τη λέξη. Αυτή η αναπαράσταση επιτρέπει σε μια λέξη να έχει πολλαπλά συναισθήματα που αντιστοιχούν σε κάθε έννοια. Επειδή υπάρχουν τρεις βαθμολογίες, κάθε έννοια από μόνη της μπορεί να είναι τόσο θετική όσο και αρνητική, ή ούτε θετική ούτε αρνητική.

b. WordNet-Affect

Το WordNet-Affect όπως το SentiWordNet, είναι μια πηγή αποτελούμενη από 2874 synsets στα οποία, χρησιμοποιώντας μια ημι-εποπτευόμενη μέθοδο μηχανικής μάθησης,

έχουν προστεθεί ετικέτες συναισθήματος που ονομάζονται a-labels, στο λεξικό WordNet [88]. Το WordNet-Affect δημιουργήθηκε ως εξής:

1ον. Δημιουργείται ένα σύνολο από βασικά synsets, στα οποία έχουν προστεθεί χειροκίνητα ετικέτες συναισθήματος της μορφής a-labels.

2ον. Αυτές οι ετικέτες προβάλλονται σε άλλα synsets χρησιμοποιώντας σχεσιακές συνάψεις του λεξικού WordNet.

3ον. Οι ετικέτες a-labels αξιολογούνται και διορθώνονται με μη αυτόματο τρόπο, όπου αυτό κρίνεται απαραίτητο.

c. SenticNet

Το SenticNet [89] είναι μια μέθοδος εξόρυξης απόψεων και ανάλυσης συναισθημάτων που ερευνά τεχνικές τεχνητής νοημοσύνης και τεχνικές σημασιολογικού Ιστού και είναι ένας συνδυασμός του λεξικού WordNet-Affect και του ConceptNet. Στόχος του SenticNet είναι να συμπεράνει την πολικότητα των εννοιών σε ένα φυσικό γλωσσικό κείμενο σε σημασιολογικό επίπεδο και όχι σε συντακτικό επίπεδο. Η ανάλυση συναισθημάτων έτσι γίνεται σε επίπεδο έννοιας, αξιοποιώντας υπονοούμενα και υποδηλωτικές πληροφορίες που σχετίζονται με λέξεις και εκφράσεις πολλαπλών λέξεων αντί να βασίζονται αποκλειστικά σε συχνότητες επανεμφάνισης λέξεων. Η μέθοδος χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) για να δημιουργήσει πολικότητα για σχεδόν 100.000 έννοιες.

Συγκεκριμένα, για τις έννοιες που μοιάζουν περισσότερο με την έννοια του κειμένου που εισάγεται προς ανάλυση, έχουν δημιουργηθεί τιμές ταξινόμησης συναισθημάτων που αναπαράγονται βάσει τεσσάρων ειδικά διαμορφωμένων συναισθηματικών διαστάσεων, όπως για παράδειγμα «Η τερπνότητα» (Pleasantness), «Η προσήλωση» (Attention), «Η ευαισθησία»(Sensitivity) και «Η κλίση (ταλέντο)»(Aptitude). Για αυτές τις τέσσερις διαστάσεις η πολικότητα έχει κυμαινόμενη τιμή μεταξύ -1 και +1, όπου το -1 δηλώνει εξαιρετικά αρνητική πολικότητα και το +1 εξαιρετικά θετική. Αυτή η γνωσιακή βάση είναι διαθέσιμη για δωρεάν λήψη σε ένα αυτόνομο αρχείο XML και η τελευταία έκδοση (που εκδίδεται κάθε δύο χρόνια) μπορεί επίσης να προσεγγιστεί ως ενσωματωμένο API.

Από την δημιουργία του έως σήμερα το SenticNet εξελίχθηκε περνώντας από πολλά στάδια και εκδόσεις, με την τελευταία να παρουσιάζεται το 2020 (SenticNet 6). Οι παλιές εκδόσεις του SenticNet επικεντρώθηκαν στη συλλογή πολικότητας της ανάλυσης σκέψης χρησιμοποιώντας την κοινή λογική, αλλά λόγω της αδυναμίας τους να εξάγουν

συμπεράσματα, δεν κατάφερναν να φτάνουν στον επιθυμητό στόχο. Στο παρελθόν, το SenticNet έχει χρησιμοποιηθεί σε πολλές εργασίες εκτός από την ανίχνευση πολικότητας, όπως π.χ. συστήματα σύστασης [90], χρηματιστηριακή πρόβλεψη [91], πολιτικές προβλέψεις [92], ανίχνευση ειρωνείας [93], μέτρηση αποτελεσματικότητας φαρμάκων [94], ανίχνευση κατάθλιψης [95], διαλογή ψυχικής υγείας [96], ανίχνευση συμπεριφοράς εμβολιασμού [97], ψυχολογικές μελέτες [98] και πολλά άλλα.

Η έκδοση SenticNet 6 έφερε βελτίωση στην ακρίβεια όλων αυτών των εργασιών, χρησιμοποιώντας μια προσέγγιση στην αναπαράσταση της γνώσης που είναι τόσο από πάνω προς τα κάτω όσο και από κάτω προς τα πάνω. Από πάνω προς τα κάτω με την έννοια ότι αξιοποιεί συμβολικά μοντέλα (δηλαδή, λογικά και σημασιολογικά δίκτυα) για να κωδικοποιήσει το νόημα και από κάτω προς τα πάνω, επειδή χρησιμοποιεί υποσυμβολικές μεθόδους (π.χ. biLSTM και BERT) για να μάθει σιωπηρά συντακτικά μοτίβα από τα δεδομένα. Αυτή η σύζευξη συμβολικής και υποσυμβολικής τεχνητής νοημοσύνης είναι το κλειδί για την σύζευξη των δρόμων της Επεξεργασίας Φυσικής Γλώσσα (NLP) και της κατανόησης της φυσικής γλώσσας.

d. Senti-Strength

Η μέθοδος SentiStrength [99] σχεδιάστηκε για να ανιχνεύσει την πολικότητα ενός συνόλου δεδομένων ως θετική ή αρνητική καθώς και τις αντίστοιχες τιμές συνοχής τους (από 1 έως 5) και για τις δύο πολώσεις. Η διαδικασία του stemming (ορισμός παρακάτω) που χρησιμοποιείται εδώ είναι πολύ απλή και με την βοήθεια του λεξικού αναζητά λέξεις με παρόμοια ρίζα (π.χ η λέξη ποδόσφαιρο θα ταιριάζει με τις λέξεις που ξεκινούν με «ποδόσφαιρ-» όπως η λέξη ποδοσφαιριστής). Οι βαθμολογίες μεταξύ 1 και 5 που αναφέρθηκαν παραπάνω αρχικά ανατέθηκαν χειροκίνητα (ανθρώπινη συμβολή) αναπτύσσοντας ένα σώμα 2.600 σχολίων από το MySpace και αργότερα ενημερώθηκε μέσω πρόσθετων δοκιμών. Η ανθρώπινη συμβολή είναι σημαντικοί καθώς πολλοί συναισθηματικοί όροι εμφανίζονται σπάνια σε ένα κείμενο και η αναγνώρισή τους είναι δύσκολη. Ο πυρήνας του SentiStrength αποτελείται από 2310 λέξεις και όρους συναισθήματος του λεξικού της εφαρμογής ανάλυσης κειμένου LIWC, ενώ στην πορεία προστέθηκαν μια λίστα αρνητικών και θετικών λέξεων, ένας κατάλογος βοηθητικών λέξεων για την ενίσχυση (π.χ. "πολύ") ή την αποδυνάμωση (π.χ. "κάπως") των συναισθημάτων, μια λίστα από emoticons, ενώ έγινε και χρήση επαναλαμβανόμενων σημείων στίξης (π.χ. "Cool!!!!") για επιπλέον ενίσχυση των συναισθημάτων. Το SentiStrength χωρίζει το κείμενο σε λέξεις, στη συνέχεια διαχωρίζει τα emoticons και τα

σημεία στίξης και μετά τη διαίρεση των λέξεων ελέγχεται με την αντιστοίχιση λεξικού για οποιονδήποτε όρο συναισθήματος. Η βαθμολογία της κάθε πρότασης εξαρτάται από τον βαθμό αντιστοίχισης με τους όρους συναισθήματος.

e. Emo-Lexicon

Το Emo-Lexicon [100] είναι ένα λεξικό 14000 όρων που δημιουργήθηκε χρησιμοποιώντας πύλες πληθοπορισμού (crowdsourcing) όπως το Amazon Mechanical Turk. Αν και δημιουργείται χειροκίνητα, το λεξικό αυτό είναι μεγαλύτερο από άλλα λεξικά συναισθημάτων, γεγονός που υποδεικνύει ότι το crowdsourcing αποτελεί έναν ισχυρό μηχανισμό δημιουργίας μεγάλης κλίμακας λεξικού συναισθημάτων για τη δημιουργία μεγάλης κλίμακας λεξικό συναισθημάτων. Ωστόσο, επειδή η διαδικασία δημιουργίας του είναι ανοιχτή στο ευρύ κοινό, ο ποιοτικός έλεγχος αποτελεί μια πρόκληση. Για να μετριαστεί αυτό, το λεξικό δημιουργείται βάσει κάποιων επιπλέον οδηγιών, ως εξής:

1. Μια λίστα λέξεων δημιουργείται από έναν θησαυρό λέξεων.
2. Όταν ένας σχολιαστής σχολιάζει μια λέξη με συναίσθημα, πρέπει πρώτα να εξακριβώσει την αίσθηση της λέξης. Η λέξη-στόχος εμφανίζεται μαζί με τέσσερις λέξεις. Ο σχολιαστής πρέπει να επιλέξει μία από αυτές που είναι πιο κοντά στη λέξη-στόχο.
3. Μόνο εάν ο σχολιαστής ήταν σε θέση να καθορίσει σωστά την αίσθηση της λέξης, ο σχολιασμός του για την ετικέτα συναισθημάτων λαμβάνεται υπόψιν.

f. SO-CAL

Το σύστημα SO-CAL (Sentiment Orientation CALculator) αναπτύχθηκε από τον Brooke το 2009 [101] και βασίζεται σε ένα χειροκίνητα κατασκευασμένο αποθετήριο ακατέργαστων λέξεων. Σε αντίθεση με το SentiWordNet, δεν υπάρχουν πληροφορίες συναισθήματος που σχετίζονται με μια λέξη. Το SO-CAL χρησιμοποιεί ως βάση έναν λεξικό συναισθημάτων που αποτελείται από περίπου 5000 λέξεις. Κάθε λέξη στο SO-CAL έχει μια ετικέτα συναισθήματος που παίρνει ακέραιες τιμές από -5 έως 5, εκτός από το 0 που αφορά αντικειμενικές λέξεις, οι οποίες αποκλείονται από την ανάλυση. Στα δυνατά σημεία του SO-CAL συγκαταλέγεται η ακρίβειά του, καθώς βασίζεται στη χρήση λεπτομερών χαρακτηριστικών που χειρίζονται το συναίσθημα σε διάφορες περιπτώσεις με τρόπους που συμμορφώνονται με γλωσσικά φαινόμενα. Επιπλέον, μερικά ειδικά

χαρακτηριστικά λειτουργούν εκτός του πεδίου εφαρμογής του λεξικού, προκειμένου να επηρεάσουν το συναίσθημα σε επίπεδο εγγράφου.

g. Happiness Index & ANEW (Affective Norms for English Words)

Το Happiness Index (Δείκτης Ευτυχίας) [102] είναι μια κλίμακα συναισθημάτων που βασίζεται στο λεξικό ANEW (Affective Norms for English Word) [103]. Το λεξικό ANEW είναι μια συλλογή από 2447 λέξεις έχουν βαθμολογηθεί από προπτυχιακούς φοιτητές ως προς τρεις συναισθηματικές διαστάσεις, την δυναμικότητα/σθένος (valence), την διέγερση (arousal) και την κυριαρχίας τους (dominance). Στον Δείκτης Ευτυχίας ένα δεδομένο κείμενο βαθμολογείται μεταξύ 1 έως 9, υποδεικνύοντας την ποσότητα ευτυχίας που υπάρχει στο κείμενο. Οι συγγραφείς υπολόγισαν τη συχνότητα εμφάνισης κάθε λέξης από το λεξικό ANEW στο κείμενο και στη συνέχεια υπολόγισαν έναν σταθμισμένο μέσο όρο του σθένους των λέξεων μελέτης ANEW. Η επικύρωση της βαθμολογίας του Δείκτη Ευτυχίας βασίζεται σε παραδείγματα. Συγκεκριμένα, οι συγγραφείς το εφάρμοσαν σε ένα σύνολο δεδομένων στίχων τραγουδιών, τίτλων τραγουδιών και προτάσεων ιστολογίου. Διαπίστωσαν ότι η βαθμολογία ευτυχίας για τους στίχους των τραγουδιών είχε μειωθεί από το 1961 έως το 2007, ενώ η βαθμολογία για αναρτήσεις ιστολογίου την ίδια περίοδο είχε αυξηθεί.

h. LIWC (Linguistic Inquiry and Word Count)

Το LIWC είναι μια εφαρμογή ανάλυσης κειμένου που αναπτύχθηκε το 1992, στα πλαίσια μιας διερευνητικής μελέτης για την θεραπευτική χρήση της γλώσσας, από τους Martha E. Francis και James W.Pennebaker [104]. Στα χρόνια που ακολούθησαν η εφαρμογή δέχθηκε τέσσερις ενημερώσεις (LIWC2001, LIWC2007, LIWC2015, LIWC-22), που περιλάμβαναν βελτιώσεις που αφορούν τόσο στο περιβάλλον χρήσης της εφαρμογής, όσο στον εμπλουτισμό του λεξικού με βάση το οποίο γίνεται η ανάλυση του κειμένου. Σε όλες τις εκδόσεις παρέχεται η δυνατότητα ανάλυσης μεμονωμένων ή πολλαπλών αρχείων κειμένου, γρήγορα και αποτελεσματικά, επιτρέποντας στο χρήστη να διερευνήσει τους πολλαπλούς τρόπους με τους οποίους μπορούν να χρησιμοποιηθούν οι λέξεις. Η εφαρμογή LIWC βασίζεται σε ένα ενσωματωμένο προεπιλεγμένο λεξικό που καθορίζει ποιες λέξεις πρέπει να υπολογίζονται στα αρχεία κειμένου προορισμού. Οι λέξεις κειμένου που διαβάζονται και αναλύονται από το LIWC αναφέρονται ως **target words (*λέξεις-στόχοι)***. Οι λέξεις στο αρχείο λεξικού του LIWC αναφέρονται ως **dictionary words (*λέξεις λεξικού)***. Οι ομάδες λέξεων λεξικού που άπτονται ενός

συγκεκριμένου τομέα (π.χ. λέξεις αρνητικού συναισθήματος) αναφέρονται ως υποτμήματα ή κατηγορίες λέξεων. Κάθε μια από αυτές τις κατηγορίες περιλαμβάνει μια σειρά από μεταβλητές που αντιστοιχούν σε συγκεκριμένες λέξεις λεξικού.

Το LIWC έχει σχεδιαστεί να δέχεται γραπτό ή μεταγραμμένο κείμενο το οποίο έχει αποθηκευτεί σε ψηφιακό αρχείο σε μία από τις εξής μορφές: α) ακατέργαστο κείμενο, β) pdf, γ) rtf , δ) αρχεία Microsoft Word, ε) αρχεία Microsoft Excel ή στ) αρχεία csv. Το LIWC έχει πρόσβαση σε ένα μόνο αρχείο ή ομάδα αρχείων και αναλύει το καθένα διαδοχικά, γράφοντας την έξοδο σε ένα μοναδικό αρχείο. Το LIWC διαβάζει από κάθε καθορισμένο αρχείο κειμένου, μία λέξη-στόχο κάθε φορά. Καθώς επεξεργάζεται κάθε λέξη-στόχος, γίνεται αναζήτηση στο αρχείο λεξικού, αναζητώντας μια αντιστοιχία λέξης λεξικού με την τρέχουσα λέξη-στόχο. Εάν η λέξη-στόχος ταιριάζει με τη λέξη λεξικού, η/οι αντίστοιχη/ες κλίμακα/ες (ή βαθμολογία/ες) κατηγορίας λέξεων για αυτήν τη λέξη αυξάνεται. Κατά την επεξεργασία του αρχείου κειμένου προορισμού, αυξάνονται επίσης οι μετρήσεις για διάφορα δομικά στοιχεία σύνθεσης του κειμένου (π.χ. πλήθος λέξεων και σημεία στίξης προτάσεων). Το τελικό αρχείο που εξάγεται για κάθε κείμενο, αποτελείται από μια γραμμή των μεταβλητών εξόδου, το όνομα του αρχείου και την καταμέτρηση των λέξεων που αντιστοιχούν στην κάθε μεταβλητή. Το πλήθος των μεταβλητών έχει αυξηθεί από τις 84 μεταβλητές στην $2^{η}$ έκδοση (LIWC2001) σε 90 στην $4^{η}$ έκδοση, ενώ στην τελευταία έκδοση (LIWC-22) ακολουθείτε μια αρκετά διαφορετική προσέγγιση για τον καθορισμό των μεταβλητών εξόδου, ενώ έχουν προστεθεί επιπλέον 4 μεταβλητές σύνοψης, όπως περιγράφεται παρακάτω.

Στη συνέχεια παρατίθενται τα βασικά χαρακτηριστικά, οι αλλαγές και οι βελτιώσεις που πραγματοποιήθηκαν σε κάθε έκδοση.

i. LIWC2001

Η δεύτερη έκδοση της εφαρμογής είχες την δυνατότητα να διαβάσει αρχεία κειμένου με κατάληξη .txt, και να εξάγει τα αποτελέσματα αποθηκεύοντας τα σε κείμενο οριοθετημένο με στηλοθέτες. Η έκδοση αυτή καταγράφει έως και 84 μεταβλητές στο αρχείο εξόδου το οποίο περιλαμβάνει το όνομα αρχείου, 17 τυπικές γλωσσικές διαστάσεις (π.χ. καταμέτρηση λέξεων, ποσοστό αντωνυμιών, άρθρα), 25 κατηγορίες λέξεων που αφορούν ψυχολογικές καταστάσεις (π.χ. επιρροή, γνωστική λειτουργία), 10 διαστάσεις που σχετίζονται με τη «σχετικότητα» (χρόνος, χώρος, κίνηση) και 19 κατηγορίες προσωπικών δραστηριοτήτων (π.χ. εργασία, σπίτι, δραστηριότητες αναψυχής).

Το λεξικό της έκδοσης LIWC2001 αποτελούνταν από 2.300 λέξεις και στελέχη λέξεων. Κάθε στέλεχος λέξης ή λέξης ορίζει μία or περισσότερες κατηγορίες λέξεων ή υποτμήματα. Για παράδειγμα, η λέξη "έκλαψε" είναι μέρος τεσσάρων κατηγοριών λέξεων: θλίψη, συνολική επίδραση, αρνητικό συναίσθημα, και ένα ρήμα του παρελθόντος. Ως εκ τούτου, εάν βρεθεί στο κείμενο προορισμού, κάθε μία από αυτές τις τέσσερις κατηγορίες η αντίστοιχες βαθμολογίες κλίμακας θα αυξηθούν. Όπως σε αυτό το παράδειγμα, πολλές από τις κατηγορίες του LIWC2001 έχουν ιεραρχική ταξινόμηση. Το σύνολο των λέξεων που δηλώνουν θυμό, εξ ορισμού, θα κατηγοριοποιηθούν ως αρνητικά συναισθήματα και συνολικά ως λέξεις συναισθημάτων. Επιπλέον, ακόμη και οι ρίζες των λέξεων μπορούν να ανιχνευτούν. Για παράδειγμα, το στέλεχος "hungr*" που επιτρέπει οποιαδήποτε λέξη-στόχο που ταιριάζει με τα πρώτα πέντε γράμματα να υπολογίζεται ως λέξη που αναφέρεται στην κατηγορία διατροφής (π.χ *hungry, hungrier, hungriest*). Ο αστερίσκος, υποδηλώνει την αποδοχή όλων των γραμμάτων, παύλες ή αριθμών μετά την εμφάνισή του.

Κάθε μία από τις 74 προεπιλεγμένες κατηγορίες LIWC2001 αποτελείται από μια λίστα λέξεων λεξικού που καθορίζουν αυτήν την κλίμακα. Ο πίνακας παρέχει μια ολοκληρωμένη λίστα με τις προεπιλεγμένες κατηγορίες λεξικών LIWC2001, κλίμακες, λέξεις κλίμακας δείγματος και σχετικές μετρήσεις λέξεων κλίμακας (Εικόνες 3 και 4).

Dimension	Abbrev	Examples	# Words	Judge 1	Judge 2
I. STANDARD LINGUISTIC DIMENSIONS					
Word Count	WC				
Words per sentence	WPS				
Sentences ending with ?	Qmarks				
Unique words (type/token ratio)	Unique				
% words captured, dictionary words	Dic				
% words longer than 6 letters	Sixltr				
Total pronouns	Pronoun	I, our, they, you're	70		
1st person singular	I	I, my, me	9		
1st person plural	We	we, our, us	11		
Total first person	Self	I, we, me	20	.78	.47
Total second person	You	you, you'll	14		
Total third person	Other	she, their, them	22		
Negations	Negate	no, never, not	31		
Assents	Assent	yes, OK, mmhmm	18		
Articles	Article	a, an, the	3		
Prepositions	Preps	on, to, from	43		
Numbers	Number	one, thirty, million	29		
II. PSYCHOLOGICAL PROCESSES					
Affective or Emotional Processes	Affect	happy, ugly, bitter	615		
Positive Emotions	Posemo	happy, pretty, good	261	.63	.33
Positive feelings	Posfeel	happy, joy, love	43		
Optimism and energy	Optim	certainty, pride, win	69	.37	.22
Negative Emotions	Negemo	hate, worthless, enemy	345	.75	.38
Anxiety or fear	Anx	nervous, afraid, tense	62	.57	.40
Anger	Anger	hate, kill, pissed	121	.57	.41
Sadness or depression	Sad	grief, cry, sad	72	.66	.29
Cognitive Processes	Cogmech	cause, know, ought	312		
Causation	Cause	because, effect, hence	49	.39	.31
Insight	Insight	think, know, consider	116	.73	.23
Discrepancy	Discrep	should, would, could	32	.53	.20
Inhibition	Inhib	block, constrain	64		
Tentative	Tentat	maybe, perhaps, guess	79	.49	.21
Certainty	Certain	always, never	30		
Sensory and Perceptual Processes	Senses	see, touch, listen	111		
Seeing	See	view, saw, look	31		
Hearing	Hear	heard, listen, sound	36		
Feeling	Feel	touch, hold, felt	30		
Social Processes	Social	talk, us, friend	314		

Εικόνα 3. Πληροφορίες μεταβλητών εξόδου του λεξικού LIWC2001

Communication	Comm	talk, share, converse	124		
Other references to people	Othref	1st pl, 2nd, 3rd per prns	54		
Friends	Friends	pal, buddy, coworker	28	.74	.69
Family	Family	mom, brother, cousin	43	.81	.80
Humans	Humans	boy, woman, group	43		
III. RELATIVITY					
Time	Time	hour, day, oclock	113		
Past tense verb	Past	walked, were, had	144	.75	.75
Present tense verb	Present	walk, is, be	256		
Future tense verb	Future	will, might, shall	14		
Space	Space	around, over, up	71		
Up	Up	up, above, over	12		
Down	Down	down, below, under	7		
Inclusive	Incl	with, and, include	16		
Exclusive	Excl	but, except, without	19		
Motion	Motion	walk, move, go	73		
IV. PERSONAL CONCERNS					
Occupation	Occup	work, class, boss	213		
School	School	class, student, college	100	.27	.25
Job or work	Job	employ, boss, career	62		
Achievement	Achieve	try, goal, win	60		
Leisure activity	Leisure	house, TV, music	102		
Home	Home	house, kitchen, lawn	26		
Sports	Sports	football, game, play	28		
Television and movies	TV	TV, sitcom, cinema	19		
Music	Music	tunes, song, cd	31		
Money and financial issues	Money	cash, taxes, income	75		
Metaphysical issues	Metaph	God, heaven, coffin	85		
Religion	Relig	God, church, rabbi	56		
Death and dying	Death	dead, burial, coffin	29		
Physical states and functions	Physcal	ache, breast, sleep	285		
Body states, symptoms	Body	ache, heart, cough	200	.45	.61
Sex and sexuality	Sexual	lust, penis, fuck	49		
Eating, drinking, dieting	Eating	eat, swallow, taste	52		
Sleeping, dreaming	Sleep	asleep, bed, dreams	21		
Grooming	Groom	wash, bath, clean	15		
APPENDIX: EXPERIMENTAL DIMENSIONS					
Swear words	Swear	damn, fuck, piss	29		
Nonfluencies	Nonfl	uh, rr*	6		
Fillers	Fillers	youknow, Imean	6		

Εικόνα 4. Πληροφορίες μεταβλητών εξόδου του λεξικού LIWC2001

ii. LIWC2007

Η LIWC2007 ήταν η τρίτη έκδοση της εφαρμογής, από την χρήση της οποίας προέκυψαν και τα αποτελέσματα στα οποία στηρίχθηκε και η παρούσα διπλωματική, Στην έκδοση αυτή ενημερώθηκε η αρχική εφαρμογή με ένα εκτεταμένο λεξικό και ένα πιο σύγχρονο και λειτουργικό σχεδιασμό λογισμικού (Pennebaker, Francis & Booth, 2001)[105]. Παρέχεται πλέον η δυνατότητα εισαγωγής κειμένου από ψηφιακά αρχεία σε μία από τις πολλαπλές μορφές, συμπεριλαμβανομένου του ακατέργαστου κειμένου, ASCII, unicode ή αρχεία Microsoft Word. Το LIWC2007 αποκτά πρόσβαση σε ένα μόνο αρχείο ή ομάδα αρχείων και τα αναλύει διαδοχικά, γράφοντας την έξοδο σε ένα μόνο αρχείο. Με κάθε αρχείο κειμένου, περίπου 80 μεταβλητές εξόδου γράφονται ως μία γραμμή δεδομένων σε καθορισμένο αρχείο εξόδου. Αυτή η εγγραφή δεδομένων περιλαμβάνει το όνομα αρχείου, 4 γενικές περιγραφικές κατηγορίες (συνολικός αριθμός λέξεων, λέξεις ανά πρόταση, ποσοστό λέξεων που καταγράφονται από το λεξικό και ποσοστό λέξεων μεγαλύτερο από έξι γράμματα), 22 τυπικές γλωσσικές διαστάσεις (π.χ. ποσοστό λέξεων στο κείμενο που είναι αντωνυμίες, άρθρα, βοηθητικά ρήματα κ.λπ.), 32 κατηγορίες λέξεων που αφορούν ψυχολογικές καταστάσεις (π.χ., επιρροή, γνωστική λειτουργία, βιολογικές διεργασίες), 7 κατηγορίες προσωπικών δραστηριοτήτων (π.χ. εργασία, σπίτι, δραστηριότητες αναψυχής), 3 παρα-γλωσσολογικές διαστάσεις (π.χ. συναίνεση) και 12 κατηγορίες σημείων στίξης (περίοδοι, κόμματα κ.τ.λ.). Το προεπιλεγμένο λεξικό του LIWC2007 αποτελείται από σχεδόν 4.500 λέξεις και στελέχη λέξεων. Ένας πλήρης κατάλογος των τυποποιημένων κατηγοριών του LIWC2007 εμφανίζεται στον παρακάτω πίνακα (Εικόνες 5 και 6).

Η έκδοση LIWC2007, περιλάμβανε ουσιαστική ενημέρωση των λεξικών και τροποποίηση στη δομή του λεξικού. Αντλώντας πάνω από αρκετές εκατοντάδες χιλιάδες αρχεία κειμένου που αποτελούνται από αρκετές εκατοντάδες εκατομμύρια λέξεις τόσο από δείγματα γραπτής όσο και από ομιλούμενης γλώσσας, έγινε προσπάθεια προσδιορισμού κοινών λέξεων και κατηγοριών λέξεων που δεν έχουν καταγραφεί στις προηγούμενες εκδόσεις LIWC. Εξετάζοντας τις λέξεις που χρησιμοποιούνται συχνότερα το 2000, μια ομάδα τεσσάρων δικαστών συμφώνησε ατομικά και συλλογικά ποιες νέες λέξεις και νέες κατηγορίες λέξεων ήταν κατάλληλες για συμπερίληψη. Με βάση πρόσφατες μελέτες που υποδηλώνουν ότι οι λέξεις λειτουργίας είναι ιδιαίτερα σχετικές με τις ψυχολογικές διαδικασίες, προστέθηκαν κατηγορίες συνδυασμών, επιρρημάτων, ποσοτικοποιητών, βοηθητικών ρημάτων, κοινώς χρησιμοποιούμενων ρημάτων, απρόσωπων αντωνυμιών, λέξεων ολικής λειτουργίας και λέξεων ολικής σχετικότητας.

Επιπλέον, οι αντωνυμίες τρίτου προσώπου χωρίστηκαν σε 3ο ενικό και 3ου πληθυντικό. Τέλος, μια μεγάλη ομάδα σημείων στίξης έχουν προστεθεί ως ξεχωριστές κατηγορίες.

Category	Abbrev	Examples	Words in category	Validity (judges)	Alpha: Binary/raw
Linguistic Processes					
Word count	wc				
words/sentence	wps				
Dictionary words	dic				
Words>6 letters	sixltr				
Total function words	funct		464		.97/.40
Total pronouns	pronoun	I, them, itself	116		.91/.38
Personal pronouns	ppron	I, them, her	70		.88/.20
1st pers singular	i	I, me, mine	12	.52	.62/.44
1st pers plural	we	We, us, our	12		.66/.47
2nd person	you	You, your, thou	20		.73/.34
3rd pers singular	shehe	She, her, him	17		.75/.52
3rd pers plural	they	They, their, they'd	10		.50/.36
Impersonal pronouns	ipron	It, it's, those	46		.78/.46
Articles	article	A, an, the	3		.14/.14
[Common verbs][a]	verb	Walk, went, see	383		.97/.42
Auxiliary verbs	auxverb	Am, will, have	144		.91/.23
Past tense [a]	past	Went, ran, had	145	.79	.94/.75
Present tense [a]	present	Is, does, hear	169		.91/.74
Future tense [a]	future	Will, gonna	48		.75/.02
Adverbs	adverb	Very, really, quickly	69		.84/.48
Prepositions	prep	To, with, above	60		.88/.35
Conjunctions	conj	And, but, whereas	28		.70/.21
Negations	negate	No, not, never	57		.80/.28
Quantifiers	quant	Few, many, much	89		.88/.12
Numbers	number	Second, thousand	34		.87/.61
Swear words	swear	Damn, piss, fuck	53		.65/.48
Psychological Processes					
Social processes[b]	social	Mate, talk, they, child	455		.97/.59
Family	family	Daughter, husband, aunt	64	.87	.81/.65
Friends	friend	Buddy, friend, neighbor	37	.70	.53/.12
Humans	human	Adult, baby, boy	61		.86/.26
Affective processes	affect	Happy, cried, abandon	915		.97/.36
Positive emotion	posemo	Love, nice, sweet	406	.41	.97/.40
Negative emotion	negemo	Hurt, ugly, nasty	499	.31	.97/.61
Anxiety	anx	Worried, fearful, nervous	91	.38	.89/.33
Anger	anger	Hate, kill, annoyed	184	.22	.92/.55
Sadness	sad	Crying, grief, sad	101	.07	.91/.45
Cognitive processes	cogmech	cause, know, ought	730		.97/.37
Insight	insight	think, know, consider	195		.94/.51
Causation	cause	because, effect, hence	108	.44	.88/.26
Discrepancy	discrep	should, would, could	76	.21	.80/.28
Tentative	tentat	maybe, perhaps, guess	155		.87/.13
Certainty	certain	always, never	83		.85/.29
Inhibition	inhib	block, constrain, stop	111		.91/.20
Inclusive	incl	And, with, include	18		.66/.32

Εικόνα 5. Πληροφορίες μεταβλητών εξόδου του λεξικού LIWC2007

Category	Abbrev	Examples	Words in category	Validity (judges)	Alpha: Binary/raw
Exclusive	excl	But, without, exclude	17		.67/.47
Perceptual processes[c]	percept	Observing, heard, feeling	273		.96/.43
See	see	View, saw, seen	72		.90/.43
Hear	hear	Listen, hearing	51		.89/.37
Feel	feel	Feels, touch	75		.88/.26
Biological processes	bio	Eat, blood, pain	567	.53	.95/.53
Body	body	Cheek, hands, spit	180		.93/.45
Health	health	Clinic, flu, pill	236		.85/.38
Sexual	sexual	Horny, love, incest	96		.69/.34
Ingestion	ingest	Dish, eat, pizza	111		.86/.68
Relativity	relativ	Area, bend, exit, stop	638		.98/.51
Motion	motion	Arrive, car, go	168		.96/.41
Space	space	Down, in, thin	220		.96/.44
Time	time	End, until, season	239		.94/.58
Personal Concerns					
Work	work	Job, majors, xerox	327		.91/.69
Achievement	achieve	Earn, hero, win	186		.93/.37
Leisure	leisure	Cook, chat, movie	229		.88/.50
Home	home	Apartment, kitchen, family	93		.81/.57
Money	money	Audit, cash, owe	173		.90/.53
Religion	relig	Altar, church, mosque	159		.91/.53
Death	death	Bury, coffin, kill	62		.86/.40
Spoken categories					
Assent	assent	Agree, OK, yes	30		.59/.41
Nonfluencies	nonflu	Er, hm, umm	8		.28/.23
Fillers	filler	Blah, Imean, youknow	9		.63/.18

Εικόνα 6. Πληροφορίες μεταβλητών εξόδου του λεξικού LIWC2007

Επιπλέον, ορισμένες από τις αρχικές κατηγορίες έχουν καταργηθεί - κυρίως επειδή αυτές οι κατηγορίες είχαν σταθερά χαμηλά βασικά ποσοστά και σπάνια χρησιμοποιήθηκαν. Αυτές είναι οι: Optimism, Positive Feelings, Communication Verbs, Other References, Metaphysical, Sleeping, Grooming, School, Sports, Television, Up, and Down. Η κατηγορία των μοναδικών λέξεων (Unique Words) (γνωστή και ως Type.Token ratio έχει επίσης καταργηθεί.

1. Σύγκριση LIWC2007 με LIWC2001

Πολλά από τα παλαιότερα λεξικά έχουν αλλάξει ελαφρώς, μερικά έχουν ενημερωθεί ουσιαστικά (π.χ. αποκλειστικές λέξεις, γνωστικοί μηχανισμοί) και άλλα έχουν αφαιρεθεί ή προστεθεί. Στον Πίνακα 4 παρατίθενται τα μέσα, οι τυπικές αποκλίσεις και οι συσχετίσεις μεταξύ των δύο εκδόσεων λεξικού. Οι αναλύσεις αυτές βασίζονται σε

σύγκριση με πάνω από 2800 τυχαία επιλεγμένα κείμενα από καθένα από τα είδη που απαριθμούνται στον Πίνακα 1.

	LIWC2007		LIWC2001		
	mean	sd	mean	sd	correlation
Word count	1687.84	7697.27	1687.84	7697.27	1.00
Words per sentence	22.38	44.38	22.38	44.38	1.00
Dictionary words	86.31	10.13	75.32	10.64	0.97
Words>6 letters	13.26	4.56	13.26	4.56	1.00
Pronouns	12.14	4.09	14.16	4.52	0.97
1st person singular	7.82	3.68	7.78	3.67	1.00
1st person plural	0.78	0.90	0.78	0.90	1.00
2nd person	1.08	1.57	1.09	1.60	1.00
Articles	5.36	1.94	5.33	1.94	1.00
Past tense verbs	4.62	3.09	4.74	3.14	1.00
Present tense verbs	8.77	3.80	10.46	4.07	0.96
Future tense verbs	1.14	1.07	1.28	1.22	0.88
Prepositions	12.24	2.85	12.23	2.82	0.99
Negations	1.91	1.11	1.85	1.11	0.97
Numbers	2.52	2.15	2.51	2.15	1.00
Swear words	0.31	0.64	0.30	0.63	0.99
Social words	8.63	3.97	7.92	3.82	0.98
Family	0.53	0.85	0.51	0.84	0.99
Friends	0.33	0.46	0.32	0.46	0.99
Humans	0.73	0.66	0.67	0.61	0.95
Affect	5.12	2.25	4.04	1.91	0.93
Positive emotions	3.02	1.62	2.26	1.33	0.89
Negative emotions	2.04	1.43	1.76	1.31	0.97
Anxiety	0.39	0.46	0.28	0.39	0.91
Anger	0.69	0.86	0.59	0.79	0.97
Sadness	0.41	0.50	0.37	0.47	0.97
Cognitive mechanisms	16.34	4.02	6.41	2.50	0.75
Insight	2.20	1.26	1.86	1.05	0.86
Causal	1.44	0.80	0.90	0.61	0.83
Discrepancy	1.63	0.98	2.14	1.13	0.87
Tentative	2.60	1.30	2.45	1.27	0.84
Certainty	1.31	0.80	1.08	0.71	0.81
Inhibition	0.43	0.39	0.30	0.30	0.73
Inclusive	4.96	1.90	5.80	1.62	0.72
Exclusive	2.89	1.49	3.56	1.35	0.61
Seeing	0.79	0.72	0.68	0.53	0.61
Hearing	0.56	0.56	0.96	0.77	0.60
Feeling	0.69	0.63	0.44	0.53	0.68
Body	0.77	0.86	0.69	0.81	0.79
Sexual	0.36	0.66	0.33	0.59	0.91
Motion	2.33	1.34	1.54	1.07	0.86
Space	5.86	2.02	3.41	1.41	0.76
Time	5.75	2.40	4.60	2.10	0.93
Occupation	1.87	1.63	2.12	1.55	0.89
Achievement	1.27	0.87	0.78	0.59	0.80
Leisure	1.20	1.05	1.25	1.11	0.67
Home	0.77	0.90	0.73	0.80	0.89
Money	0.49	0.60	0.35	0.46	0.91
Religion	0.23	0.47	0.20	0.43	0.79
Death	0.14	0.32	0.12	0.30	0.96
Assent	0.73	1.28	0.45	0.87	0.92
Nonfluencies	0.30	0.49	0.10	0.38	0.82
Fillers	0.22	0.80	0.21	0.79	0.99

Πίνακας 1. Σύγκριση μεταξύ των λεξικών LIWC2007 και LIWC2001: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

iii. LIWC2015

Η δεύτερη (LIWC2001) και τρίτη έκδοση (LIWC2007) επικαιροποίησαν την αρχική εφαρμογή με ένα διευρυμένο λεξικό και ένα πιο σύγχρονο σχεδιασμό λογισμικού [105] Pennebaker, Booth, Boyd, & Francis,, 2007). Στην έκδοση LIWC2015 (Pennebaker, Booth, Boyd, & Francis, 2015)[106], άλλαξε σημαντικά τόσο το λεξικό όσο και τις επιλογές λογισμικού. Πρέπει να σημειωθεί ότι το λογισμικό και το λεξικό του LIWC2015 είναι μια νέα εφαρμογή και όχι μια βασική ενημέρωση σε προηγούμενες εκδόσεις, ενώ παρέχεται πλέον και έκδοση βασισμένη στο web.

Το LIWC2015 μπορούσε πλέον να δέχεται γραπτό ή μεταγραμμένο λεκτικό κείμενο το οποίο έχει αποθηκευτεί ως ψηφιακό αρχείο από τις πολλαπλές μορφές, συμπεριλαμβανομένου απλού κειμένου, PDF, RTF ή τυπικών αρχείων του Microsoft Word (π.χ. .doc και .docx). Σε αντίθεση με τις προηγούμενες εκδόσεις, το λογισμικό μπορεί τώρα να επεξεργάζεται κείμενο σε γραμμική βάση εντός και μεταξύ στηλών μέσα σε πολλές μορφές υπολογιστικών φύλλων, συμπεριλαμβανομένων εκείνων που έχουν αποθηκευτεί ως αρχεία .xls, .xlsx και .csv.

Για κάθε αρχείο κειμένου, γράφονται ως μία γραμμή δεδομένων σε ένα αρχείο εξόδου περίπου 90 μεταβλητές εξόδου (10 επιπλέον από την έκδοση LIW2007). Όπως και στις προηγούμενες εκδόσεις, αυτή η εγγραφή δεδομένων περιλαμβάνει το όνομα αρχείου και τον αριθμό λέξεων, 4 συνοπτικές μεταβλητές γλώσσας (αναλυτική σκέψη, επιρροή, αυθεντικότητα και συναισθηματικός τόνος), 3 γενικές περιγραφικές κατηγορίες (λέξεις ανά πρόταση, ποσοστό λέξεων-στόχων που καταγράφονται από το λεξικό και ποσοστό λέξεων στο κείμενο που είναι μεγαλύτερες από έξι γράμματα), 21 τυπικές γλωσσικές διαστάσεις (π.χ. ποσοστό λέξεων στο κείμενο που είναι αντωνυμίες, άρθρα, βοηθητικά ρήματα κ.λπ.), 41 κατηγορίες λέξεων που αφορούν ψυχολογικές καταστάσεις (π.χ. επιρροή, γνωστική λειτουργία, βιολογικές διεργασίες, κινήσεις), 6 κατηγορίες προσωπικών δραστηριοτήτων (π.χ. εργασία, σπίτι, δραστηριότητες αναψυχής), 5 παρα-γλωσσολογικές διαστάσεις (πχ. συναίνεση, βρισιές) και 12 κατηγορίες σημείων στίξης (περίοδοι, κόμματα κ.λπ.). Το προεπιλεγμένο λεξικό LIWC2015 αποτελείται από σχεδόν 6.400 λέξεις, στελέχη λέξεων και επιλεγμένα emoticons. Κάθε μία από τις προεπιλεγμένες κατηγορίες LIWC2015 αποτελείται από μια λίστα λέξεων λεξικού που καθορίζουν αυτήν την κλίμακα.

Μια σημαντική αλλαγή από προηγούμενες εκδόσεις του LIWC είναι η συμπερίληψη τεσσάρων νέων συνοπτικών μεταβλητών: αναλυτική σκέψη (Pennebaker et al., 2014), επιρροή (Kacewicz et al., 2012), αυθεντικότητα (Newman et al., 2003) και

συναισθηματικός τόνος (Cohn et al., 2004). Κάθε συνοπτική μεταβλητή προήλθε από ευρήματα που δημοσιεύθηκαν προηγουμένως από το εργαστήριό μας και μετατράπηκε σε εκατοστημόριο με βάση τυποποιημένες βαθμολογίες από μεγάλα δείγματα σύγκρισης.

Ορισμένες από τις αρχικές κατηγορίες έχουν καταργηθεί, κυρίως λόγω των σταθερά χαμηλών βασικών ποσοστών τους, της χαμηλής εσωτερικής αξιοπιστίας ή της σπάνιας χρήσης τους από τους ερευνητές:

1. **Ρήματα σε παρελθοντικό χρόνο (Past tense verbs)**
2. **Ρήματα σε ενεστώτα (Present tense verbs)**
3. **Ρήματα σε μελλοντικό χρόνο Future tense verbs**
4. **Λέξεις τη κατηγορίας Human**
5. **Λέξεις της κατηγορίας Inhibition**
6. **Λέξεις που δηλώνουν μη αποκλειστικότητα (Inclusives)**
7. **Λέξεις που δηλώνουν αποκλειστικότητα (Exclusives)**

Στον παρακάτω πίνακα (Πίνακας 2) παρουσιάζεται ο κατάλογος κατηγοριών που είναι είτε α) νέες στο LIWC2015, είτε β) ουσιαστικά διαφορετικές από τις αντίστοιχες κατηγορίες σε προηγούμενες εκδόσεις. Ενώ άλλες κατηγορίες LIWC2015 μπορεί επίσης να διαφέρουν ελαφρώς από εκείνες σε προηγούμενες εκδόσεις, κατηγορίες από προηγούμενες εκδόσεις του LIWC που παρουσιάζονται στον παρακάτω κατάλογο έχουν υποβληθεί σε ουσιαστική αναθεώρηση.

Common verbs	Κοινά ρήματα
Common adjectives	Κοινά επίθετα
Common comparison words	Κοινές λέξεις σύγκρισης
Interrogatives	Ανακριτικές
Female references	Γυναικείες αναφορές
Male references	Αρσενικές αναφορές
Cognitive processes	Γνωστικές διαδικασίες
Differentiation words	Διαφοροποιήσεις λέξεις
Drives Affiliation words	Οδηγεί λέξεις σχέσης
Achievement words	Λέξεις επιτεύγματος
Power words	Λέξεις δύναμης
Risk words	Λέξεις ανταμοιβής
Reward words	Λέξεις Ανταμείβοντας λέξεις
Past focus words	Παρελθόν λέξεις εστίασης

Present focus words	Λέξεις εστίασης
Future focus words	Λέξεις μελλοντικής εστίασης
Informal language	Άτυπη γλώσσα
Netspeak words	Λέξεις διαδικτυακής γλώσσας
Quantifiers	Ποσοτικοποιητές

Πίνακας 2. Κατάλογος νέων και αναθεωρημένων κατηγοριών λεξικού LIWC2015

Ωστόσο η εφαρμογή LIWC2015 συνοδεύεται από τα αρχικά εσωτερικά λεξικά και για τα δύο LIWC2001 και LIWC2007 για όσους θέλουν να βασίζονται σε παλαιότερες εκδόσεις του λεξικού καθώς και να συγκρίνουν τις αναλύσεις LIWC2015 με εκείνες που παρέχονται από παλαιότερες εκδόσεις του λογισμικού.

a. Σύγκριση LIWC2015 με LIWC2007

Τα περισσότερα από τα παλιά λεξικά έχουν αλλάξει ελαφρώς, μερικά έχουν υποστεί ουσιαστική επανεπεξεργασία (π.χ. κοινωνικές λέξεις, λέξεις γνωστικής διαδικασίας) και πολλά άλλα έχουν αφαιρεθεί ή προστεθεί. Στους Πίνακες *3, 4 και 5* παρατίθενται οι μέσοι όροι, οι τυπικές αποκλίσεις και οι συσχετίσεις μεταξύ των δύο εκδόσεων λεξικού. Όσο χαμηλότερη είναι η συσχέτιση, τόσο περισσότερες αλλαγές εμφανίζονται μεταξύ των δύο εκδόσεων.

LIWC Dimension	Output Label	LIWC2015 mean	LIWC2007 mean	LIWC 2015/2007 Correlation[1]
Word count	WC	11,921.82	11,852.99	1.00
Summary Variables				
Analytical thinking	Analytic	56.34	-	-
Clout	Clout	57.95	-	-
Authentic	Authentic	49.17	-	-
Emotional tone	Tone	54.22	-	-
Language Metrics				
Words per sentence*	WPS	17.40	25.07	0.74
Words>6 letters	Sixltr	15.60	15.89	0.98
Dictionary words	Dic	85.18	83.95	0.94
Function Words	function	51.87	54.29	0.95
Total pronouns	pronoun	15.22	14.99	0.99
Personal pronouns	ppron	9.95	9.83	0.99
1st pers singular	i	4.99	4.97	1.00
1st pers plural	we	0.72	0.72	1.00
2nd person	you	1.70	1.61	0.98
3rd pers singular	shehe	1.88	1.87	1.00
3rd pers plural	they	0.66	0.66	0.99

Πίνακας 3. Σύγκριση μεταξύ των λεξικών LIWC2007 και LIWC2015: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

Impersonal pronouns	ipron	5.26	5.17	0.99
Articles	article	6.51	6.53	0.99
Prepositions	prep	12.93	12.59	0.96
Auxiliary verbs	auxverb	8.53	8.82	0.96
Common adverbs	adverb	5.27	4.83	0.97
Conjunctions	conj	5.90	5.87	0.99
Negations	negate	1.66	1.72	0.96
Other Grammar				
Regular verbs	verb	16.44	15.26	0.72
Adjectives	adj	4.49	-	-
Comparatives	compare	2.23	-	-
Interrogatives	interrog	1.61	-	-
Numbers	number	2.12	1.98	0.98
Quantifiers	quant	2.02	2.48	0.88
Affect Words	affect	5.57	5.63	0.96
Positive emotion	posemo	3.67	3.75	0.96
Negative emotion	negemo	1.84	1.83	0.96
Anxiety	anx	0.31	0.33	0.94
Anger	anger	0.54	0.6	0.97
Sadness	sad	0.41	0.39	0.92
Social Words	social	9.74	9.36	0.96
Family	family	0.44	0.38	0.94
Friends	friend	0.36	0.23	0.78
Female referents	female	0.98	-	-
Male referents	male	1.65	-	-
Cognitive Processes[2]	cogproc	10.61	14.99	0.84
Insight	insight	2.16	2.13	0.98
Cause	cause	1.40	1.41	0.97
Discrepancies	discrep	1.44	1.45	0.99
Tentativeness	tentat	2.52	2.42	0.98
Certainty	certain	1.35	1.27	0.92
Differentiation[3]	differ	2.99	2.48	0.85
Perceptual Processes	percept	2.70	2.36	0.92
Seeing	see	1.08	0.87	0.88
Hearing	hear	0.83	0.73	0.94
Feeling	feel	0.64	0.62	0.92
Biological Processes	bio	2.03	1.88	0.94
Body	body	0.69	0.68	0.96
Health/illness	health	0.59	0.53	0.87
Sexuality	sexual	0.13	0.28	0.76
Ingesting	ingest	0.57	0.46	0.94
Drives and Needs	drives	6.93	-	-
Affiliation	affiliation	2.05	-	-
Achievement	achieve	1.30	1.56	0.93
Power	power	2.35	-	-
Reward focus	reward	1.46	-	-
Risk focus	risk	0.47	-	-

Πίνακας 4. Σύγκριση μεταξύ των λεξικών LIWC2007 και LIWC2015: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

Time Orientations[4]				
Past focus	focuspast	4.64	4.14	0.97
Present focus	focuspresent	9.96	8.1	0.92
Future focus	focusfuture	1.42	1.00	0.63
Relativity	relativ	14.26	13.87	0.98
Motion	motion	2.15	2.06	0.93
Space	space	6.89	6.17	0.96
Time	time	5.46	5.79	0.94
Personal Concerns				
Work	work	2.56	2.27	0.97
Leisure	leisure	1.35	1.37	0.95
Home	home	0.55	0.56	0.99
Money	money	0.68	0.70	0.97
Religion	relig	0.28	0.32	0.96
Death	death	0.16	0.16	0.96
Informal Speech	informal	2.52	-	-
Swear words	swear	0.21	0.17	0.89
Netspeak	netspeak	0.97	-	-
Assent	assent	0.95	1.11	0.68
Nonfluencies	nonfl	0.54	0.30	0.84
Fillers	filler	0.11	0.40	0.29
All Punctuation*	Allpunc	21.35	21.65	0.98
Periods	Period	7.49	7.56	0.98
Commas	Comma	4.75	4.75	1.00
Colons	Colon	0.64	0.73	0.98
Semicolons	SemiC	0.3	0.29	0.97
Question marks	QMark	0.58	0.58	1.00
Exclamation marks	Exclam	1.00	1.00	1.00
Dashes	Dash	1.19	1.21	0.98
Quotation marks	Quote	1.67	1.64	0.93
Apostrophes	Apostro	2.46	2.52	0.94
Parentheses (pairs)	Parenth	0.53	0.63	0.90
Other punctuation	OtherP	0.73	0.72	0.95

Πίνακας 5. Σύγκριση μεταξύ των λεξικών LIWC2007 και LIWC2015: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

iv.LIWC-22

Όπως και στις προηγούμενες εκδόσεις LIWC-22 (Pennebaker et al., 2022), έχει αλλάξει σημαντικά τόσο το λεξικό όσο και τις επιλογές λογισμικού για να αντικατοπτρίζει νέες κατευθύνσεις στην ανάλυση κειμένου. Στον πυρήνα του, το LIWC-22 αποτελείται από λογισμικό και ένα «λεξικό» – δηλαδή έναν χάρτη που συνδέει σημαντικές ψυχοκοινωνικές κατασκευές και θεωρίες με λέξεις, φράσεις και άλλες γλωσσικές κατασκευές. Όπως και στις προηγούμενες εκδόσεις, οι λέξεις που περιέχονται σε κείμενα που διαβάζονται και αναλύονται από το LIWC-22 αναφέρονται ως λέξεις-στόχοι. Οι λέξεις στο αρχείο λεξικού LIWC-22 θα αναφέρονται ως λέξεις λεξικού. Οι ομάδες λέξεων λεξικού που αγγίζουν έναν συγκεκριμένο τομέα (π.χ. λέξεις αρνητικού συναισθήματος) αναφέρονται ως "υπο-λεξικά" ή "κατηγορίες λέξεων" ή απλά

«κατηγορίες». Το LIWC-22 δέχεται γραπτό ή μεταγραμμένο προφορικό κείμενο το οποίο έχει αποθηκευτεί ως ψηφιακό αρχείο αναγνώσιμο από μηχανήματα σε μία από τις πολλαπλές μορφές, συμπεριλαμβανομένου του απλού κειμένου (.txt), pdf, RTF ή τυποποιημένων αρχείων microsoft Word (.docx). Το λογισμικό μπορεί επίσης να επεξεργαστεί κείμενα μέσα σε πολλές μορφές υπολογιστικών φύλλων, συμπεριλαμβανομένων εκείνων που έχουν αποθηκευτεί ως μορφή CSV (Comma Separated Values) ή Microsoft Excel (.xlsx). Το προεπιλεγμένο λεξικό LIWC-22 μπορεί να εκτελεστεί με δύο τρόπους: 1) μέσω του τυπικού γραφικού περιβάλλοντος εργασίας χρήστη (GUI), όπως συμβαίνει με όλες τις προηγούμενες εκδόσεις του LIWC και 2) μέσω διεπαφής γραμμής εντολών (Command Line Interface - CLI) από τη γραμμή εντολών σας ή από άλλες πλατφόρμες που μπορούν να συνδεθούν με τη γραμμή εντολών του συστήματος του χρήστη (για παράδειγμα, R ή Python).

Κατά τη λειτουργία, η μονάδα επεξεργασίας LIWC-22 αποκτά πρόσβαση σε κάθε κείμενο του συνόλου δεδομένων σας, συγκρίνει τη γλώσσα μέσα σε κάθε κείμενο με το λεξικό LIWC-22. Όπως όλες οι προηγούμενες εκδόσεις του LIWC, η ενότητα επεξεργασίας κειμένου LIWC-22 λειτουργεί μετρώντας όλες τις λέξεις σε ένα κείμενο προορισμού και, στη συνέχεια, υπολογίζοντας το ποσοστό των συνολικών λέξεων που αντιπροσωπεύονται σε καθένα από τα υποανάπτυκτες LIWC. Μετά την επεξεργασία κάθε δείγματος κειμένου, η λειτουργική μονάδα επεξεργασίας του LIWC εγγράφει την έξοδο σε έναν πίνακα δεδομένων που μπορεί να εξαχθεί σε διάφορες μορφές, συμπεριλαμβανομένων υπολογιστικών φύλλων (π.χ. CSV για το MS Excel), αλλά μορφή JSON.

Όλες οι προηγούμενες εκδόσεις του LIWC έχουν καθοριστεί σε μεγάλο βαθμό από το κεντρικό τους χαρακτηριστικό: τη βασική μονάδα επεξεργασίας του LIWC. Ένα σημαντικό χαρακτηριστικό του LIWC-22 είναι ότι το πρόγραμμα περιλαμβάνει μια ομάδα συνοδευτικών ενοτήτων επεξεργασίας που παρέχουν πρόσθετες αναλυτικές μεθόδους για τους ερευνητές. Αυτές οι πρόσθετες δυνατότητες παρέχουν νέους τρόπους ανάλυσης και κατανόησης των δειγμάτων κειμένου.

Το λεξικό του LIWC-22 αποτελείται πλέον από πάνω από 12.000 λέξεις, στελέχη λέξεων, φράσεις και επιλεγμένα emoticons. Κάθε καταχώρηση λεξικού αποτελεί μέρος μιας ή περισσότερων κατηγοριών, ή υποτμημάτων, σχεδιασμένων για την αξιολόγηση διαφόρων ψυχοκοινωνικών κατασκεύων. Για παράδειγμα, η λέξη *cried* είναι μέρος των κατηγοριών 10 λέξεων: affect, tone_pos, emotion, emo_neg, emo_sad, verbs, focuspast, communication, linguistic, and cognition. Ως εκ τούτου, εάν η λέξη *cried* βρίσκεται στο

κείμενο προορισμού, κάθε μία από αυτές τις 10 βαθμολογικές κλίμακες της κάθε υποκατηγορίας λέξεων θα αυξηθεί. Οι περισσότερες, αλλά όχι όλες, από τις κατηγορίες του LIWC-22 είναι διατεταγμένες ιεραρχικά. Όλες οι λέξεις θλίψης (sadness), εξ ορισμού, ανήκουν στην ευρύτερη κατηγορία “emo_neg”, “emotion”, “tone_neg”, καθώς και στη συνολική κατηγορία “affect”.

Όταν το LIWC σχεδιάστηκε για πρώτη φορά, η ιδέα ήταν να προσδιοριστεί μια ομάδα λέξεων που αξιοποιούσαν βασικές συναισθηματικές και γνωστικές διαστάσεις που συχνά μελετήθηκαν στην κοινωνική ψυχολογία, την υγεία και την ψυχολογία προσωπικότητας. Καθώς η κατανόησή μας για την ψυχολογία της λεκτικής συμπεριφοράς έχει ωριμάσει, το εύρος και το βάθος των κατηγοριών λέξεων στο λεξικό LIWC έχει επεκταθεί σημαντικά. Στην νέα αυτή έκδοση έχει ξαναχτίσει πλήρως ο μηχανισμός επεξεργασίας κειμένου καθώς και η ευελιξία χρήσης των λεξικών LIWC. Τα λεξικά μπορούν τώρα να φιλοξενήσουν αριθμούς, σημεία στίξης, σύντομες φράσεις, ακόμη και κανονικές εκφράσεις. Αυτές οι προσθήκες επιτρέπουν στον χρήστη να διαβάζει τη γλώσσα "netspeak" που είναι κοινή στις αναρτήσεις στα κοινωνικά μέσα Twitter και Facebook, καθώς και τρόπους επικοινωνίας τύπου SMS (π.χ. Snapchat, ανταλλαγή άμεσων μηνυμάτων). Για παράδειγμα, το "b4" κωδικοποιείται ως πρόθεση και το « :) » κωδικοποιείται ως θετική λέξη.

Σε αυτή την τελευταία έκδοση του LIWC, έχουν προστεθεί αρκετές νέες κατηγορίες, άλλες αναθεωρήθηκαν σημαντικά και ένας μικρός αριθμός καταργήθηκε. Με την εμφάνιση ισχυρότερων αναλυτικών μεθόδων και διαφορετικών γλωσσικών δειγμάτων, δημιουργήθηκαν πιο συνεπή γλωσσικά λεξικά με βελτιωμένες ψυχομετρικές ιδιότητες. Μια σημαντική ομάδα νέων μεταβλητών προστέθηκε στο LIWC. Ορισμένα βασίζονται σε πρόσφατα ερευνητικά ευρήματα και ψυχολογικούς τομείς που έχουν αγνοηθεί στο παρελθόν. Άλλες μεταβλητές έχουν συμπεριληφθεί λόγω θεωρητικών μετατοπίσεων στις κοινωνικές επιστήμες ή, γενικότερα, στον πολιτισμό. Εκτός από τις τυπικές διαστάσεις LIWC με βάση το ποσοστό των συνολικών λέξεων, υπολογίζονται πλέον και τέσσερις συνοπτικές μεταβλητές: η αναλυτική σκέψη (analytical think) (Pennebaker et al., 2014), επιρροή (clut) (Kacewicz et al., 2014), αυθεντικότητα (authenticity) (M. L. Newman et al., 2003) και ο συναισθηματικός τόνος (emotional tone) (Cohn et al., 2004).

Στους πίνακες 6, 7 και 8, παρατίθενται οι μέσοι όροι και οι τυπικές αποκλίσεις όλων των μεταβλητών LIWC-22 μαζί με τους αντίστοιχες του LIWC2015. Περιλαμβάνονται επίσης απλές συσχετίσεις μεταξύ των νέων και των παλαιών μεταβλητών LIWC. Όσο χαμηλότερη είναι η συσχέτιση, τόσο περισσότερες αλλαγές στις δύο εκδόσεις.

Οι μεταβλητές που έχουν αφαιρεθεί λόγω των σταθερά χαμηλών βασικών ποσοστών τους, της χαμηλής εσωτερικής αξιοπιστίας ή της σπάνιας χρήσης τους από τους ερευνητές περιλαμβάνουν:

[1].Λέξεις σύγκρισης (μεγαλύτερες, καλύτερες, μετά)

[2].Ανακριτές (ποιος, τι, πού)

[3].Σχετικότητα (άθροισμα χρόνου, χώρου, λέξεων κίνησης)

[4].Ορισμένα σημεία στίξης χαμηλού βασικού επιτοκίου (άνω και κάτω τελείες, ερωτηματικά, παύλες, εισαγωγικά, παρενθέσεις)

LIWC-22 (LIWC2015) Variable	LIWC-22		LIWC2015		LIWC-22/ LIWC2015 Correlation
	Mean	SD	Mean	SD	
WC	2070.5	2466.4	2070.1	2466.1	1.00
Analytic	49.6	29.8	59.5	28.0	0.99
Clout	49.6	28.4	62.5	22.1	0.95
Authentic	50.0	28.0	42.7	26.9	0.94
Tone	48.2	26.4	55.9	28.3	0.84
Words/sentence	17.2	34.7	17.2	34.7	1.00
Big words	17.2	6.7	17.2	6.7	1.00
Dictionary words	88.0	5.9	86.6	6.2	0.96
Linguistic	69.6	8.2			
function	54.6	7.1	51.8	6.8	0.90
pronoun	14.8	5.0	14.9	5.0	1.00
ppron	9.9	4.1	9.7	4.0	1.00
i	4.3	3.4	4.2	3.4	1.00
we	0.9	1.1	0.9	1.1	1.00
you	1.7	1.9	1.7	1.9	1.00
shehe	2.0	2.5	2.0	2.5	1.00
they	0.9	1.0	0.9	1.0	1.00
ipron	4.9	2.0	5.2	2.0	0.98
det	14.3	2.9			
article	6.7	2.6	6.7	2.6	1.00
number	2.1	2.6	2.2	2.6	1.00
prep	13.2	2.7	13.1	2.7	1.00
auxverb	8.7	2.9	8.6	2.8	0.99
adverb	5.3	2.4	4.9	2.3	0.97
conj	6.2	1.9	6.0	1.8	0.97
negate	1.6	1.1	1.7	1.1	0.99
verb	17.0	4.4	16.4	4.4	0.99
adj	6.0	1.7	4.5	1.5	0.74
quantity	4.0	1.7	2.0	0.9	0.56
Drives	4.4	2.3	7.5	2.5	0.79
affiliation	1.9	1.4	2.2	1.5	0.90
achieve	1.2	1.0	1.5	1.1	0.90
power	1.3	1.3	2.5	1.4	0.71

Πίνακας 6. Σύγκριση μεταξύ των λεξικών LIWC2015 και LIWC2022: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

LIWC-22 (LIWC2015) Variable	LIWC-22		LIWC2015		LIWC-22/ LIWC2015 Correlation
	Mean	SD	Mean	SD	
Cognition	22.2	4.9			
allnone	1.3	0.9			0.65[a]
cogproc	10.4	3.4	10.7	3.5	0.95
insight	2.4	1.2	2.2	1.1	0.91
cause	1.4	0.8	1.5	0.8	0.91
discrep	1.7	1.0	1.6	0.9	0.82
tentat	2.1	1.3	2.4	1.4	0.93
certitude (certainty)	0.6	0.6	1.5	0.8	0.33[a]
differ	3.1	1.4	2.9	1.4	0.91
memory	0.1	0.2			
Affect	5.4	2.6	5.6	2.4	0.87
tone_pos (posemo)	3.5	2.2	3.7	2.1	0.86
tone_neg (negemo)	1.5	1.1	1.8	1.2	0.84
emotion	1.9	1.5			0.75[b]
emo_pos	1.1	1.2			0.74[b]
emo_neg	0.7	0.6			0.68[b]
emo_anx (anx)	0.1	0.2	0.3	0.3	0.69[b]
emo_anger (anger)	0.1	0.2	0.6	0.7	0.51[b]
emo_sad (sad)	0.1	0.2	0.4	0.4	0.48[b]
swear	0.3	0.6	0.3	0.6	0.98
Social	12.1	4.2	10.2	3.9	0.89[c]
socbehav	3.9	2.0			
prosocial	0.7	0.8			
polite	0.4	1.1			
conflict	0.2	0.3			
moral	0.3	0.3			
comm	1.6	1.1			
socrefs	8.1	3.5			0.91[c]
family	0.4	0.6	0.5	0.7	0.93
friend	0.2	0.3	0.4	0.4	0.67
female	1.3	2.2	1.3	2.2	1.00
male	1.5	1.7	1.5	1.7	0.99
Culture	0.8	1.3			
politic	0.4	1.0			
ethnicity	0.1	0.4			
tech	0.3	0.6			
lifestyle	4.3	2.7			
leisure	0.6	0.8	1.2	1.1	0.84
home	0.4	0.5	0.5	0.6	0.89
work	2.5	2.4	3.1	2.7	0.95
money	0.7	1.3	0.8	1.3	0.97
relig	0.2	0.5	0.3	0.6	0.97
Physical (bio)	2.4	2.2	2.4	1.9	0.92
health	0.7	1.4	0.8	1.2	0.94[d]
illness	0.2	0.4			0.69[d]
wellness	0.0	0.2			
mental	0.0	0.1			
substances	0.1	0.2			
sexual	0.1	0.3	0.2	0.3	0.76
food (ingest)	0.7	1.6	0.7	1.3	0.95
death	0.1	0.3	0.2	0.3	0.91
need	0.5	0.5			
want	0.4	0.4			
acquire	0.8	0.6			
lack	0.1	0.3			
fulfill	0.1	0.2			
fatigue	0.1	0.2			
reward	0.2	0.3	1.5	1.0	0.29
risk	0.2	0.3	0.5	0.5	0.63
curiosity	0.4	0.5			
allure	7.0	3.1			

Πίνακας 7. Σύγκριση μεταξύ των λεξικών LIWC2015 και LIWC2022: Μέσοι όροι, Τυπική Απόκλιση και Συσχετίσεις

Perception	9.3	2.9	2.5	1.4	0.53
attention	0.5	0.5			
motion	1.7	0.9	2.1	1.0	0.79
space	6.0	2.0	6.8	2.1	0.78
visual (see)	1.0	0.8	1.0	0.8	0.88
auditory (hear)	0.3	0.4	0.7	0.7	0.65
feeling (feel)	0.5	0.5	0.6	0.5	0.86
time	4.4	1.9	5.2	2.0	0.90
focuspast	4.7	2.9	4.4	2.7	0.98
focuspresent	4.7	2.5	10.1	4.4	0.89
focusfuture	1.5	1.1	1.4	0.9	0.86
Conversation	1.3	2.3			
netspeak	0.7	1.5	0.6	1.3	0.95
assent	0.4	1.0	0.5	1.1	0.98
nonflu	0.2	0.6	0.4	0.7	0.92
filler	0.1	0.3	0.1	0.3	0.88
AllPunc	21.6	16.8	21.7	16.9	1.00
Period	8.6	7.3	8.6	7.3	1.00
Comma	4.1	3.0	4.1	3.0	1.00
QMark	1.2	9.7	1.2	9.7	1.00
Exclam	1.1	3.3	1.1	3.4	1.00
Apostro	2.1	2.2	2.1	2.2	1.00
OtherP	4.5	6.4	4.5	6.4	1.00

Πίνακας 8. Σύγκριση μεταξύ των λεξικών LIWC2007 και LIWC2015: Μέσοι όροι, Τυπική Απόκλιση, και Συσχετίσεις

a.Μεταφράσεις λεξικών του LIWC

Με τα χρόνια, τα λεξικά LIWC έχουν μεταφραστεί σε διάφορες γλώσσες σε συνεργασία με ερευνητές σε όλο τον κόσμο, όπως:

- Πορτογαλικά Βραζιλίας (Carvalho et al., 2019; Filho et al., 2013)
- Κινέζικα (Huang et al., 2012)
- Ολλανδικά (Boot et al., 2017, van Wissen & Boot, 2017)
- Γαλλικά (Piolat et al., 2011)
- Γερμανικά (Meier et al., 2018; Wolf et al., 2008)
- Ιταλικά (Agosti & Rellini, 2007)
- Ιαπωνικά (Igarashi et al., 2021)
- Νορβηγικά (Goksøyr, 2019)
- Ρουμανικά (Dudău & Sava, 2020)
- Ρωσικά (Kailer & Chung, 2011)
- Σερβικά (Bjekić et al.,2014)
- Ισπανικά (Ramírez-Esparza et al.,2007)
- Τουρκικά (Müderrisoğlu, 2012)
- Ουκρανικά (Zasiekin et al.,2018)

Μέχρι σήμερα, αυτές οι μεταφράσεις βασίζονταν κυρίως στα λεξικά LIWC2001, LIWC2007 ή LIWC2015. Οι διάφορες μεταφράσεις λεξικών LIWC, καθώς και άλλα δημοσιευμένα λεξικά, είναι διαθέσιμα στους ακαδημαϊκούς χρήστες στο αποθετήριο λεξικών LIWC (https://www.liwc.app/dictionaries).

3.7. Επεξεργασία φυσικής γλώσσας (Natural Language Processing)

Μη δομημένα δεδομένα, όπως το κείμενο, οι εικόνες και τα βίντεο περιέχουν πληθώρα πληροφοριών. Λόγω της εγγενούς πολυπλοκότητας στην επεξεργασία και την ανάλυση αυτών των δεδομένων, αλλά και της έλλειψης κατάλληλων εργαλείων (κατά τα πρώτα χρόνια των μελετών), η ανάλυση αυτών των μη δομημένων πηγών δεδομένων καθίσταται δύσκολη και περισσότερο χρονοβόρα. Η επεξεργασία φυσικής γλώσσας (ΕΦΓ) (Natural Language Processing) αφορά στην αξιοποίηση εργαλείων, τεχνικών και αλγορίθμων για την επεξεργασία και την κατανόηση δεδομένων φυσικής γλώσσας, όπως γραπτό κείμενο και ομιλία, τα οποία συνήθως δεν είναι δομημένα. Η επεξεργασία φυσικής γλώσσας συναντάται σε ευρεία γκάμα εφαρμογών όπως φαίνεται στην παρακάτω Εικόνα 7.

Search	Web	Documents	Autocomplete
Editing	Spelling	Grammar	Style
Dialog	Chatbot	Assistant	Scheduling
Writing	Index	Concordance	Table of contents
Email	Spam filter	Classification	Prioritization
Text mining	Summarization	Knowledge extraction	Medical diagnoses
Law	Legal inference	Precedent search	Subpoena classification
News	Event detection	Fact checking	Headline composition
Attribution	Plagiarism detection	Literary forensics	Style coaching
Sentiment analysis	Community morale monitoring	Product review triage	Customer care
Behavior prediction	Finance	Election forecasting	Marketing
Creative writing	Movie scripts	Poetry	Song lyrics

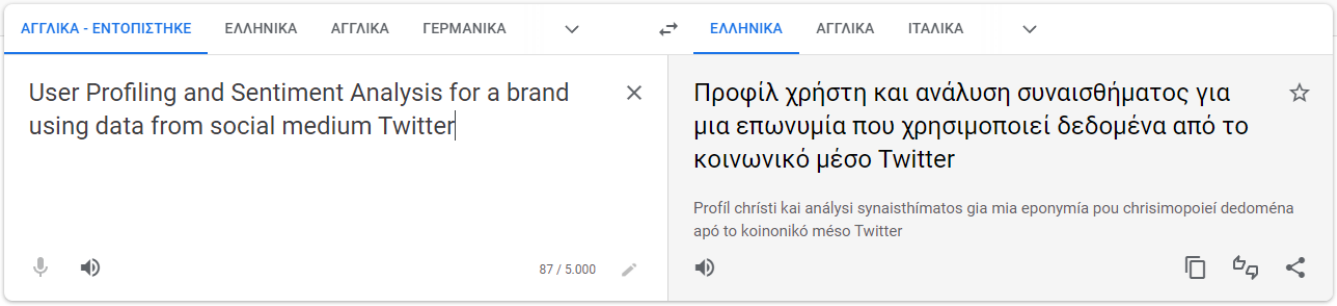
Εικόνα 7. Λίστα εφαρμογών επεξεργασίας φυσικής γλώσσας [107]

Η επεξεργασία φυσικής γλώσσας ορίζεται ως ένας εξειδικευμένος τομέας της επιστήμης των υπολογιστών και της μηχανικής και της τεχνητής νοημοσύνης με ρίζες στην υπολογιστική γλωσσολογία. Ασχολείται κυρίως με το σχεδιασμό και την κατασκευή εφαρμογών και συστημάτων που επιτρέπουν την αλληλεπίδραση μεταξύ μηχανών και φυσικών γλωσσών που δημιουργούνται από τον άνθρωπο, καθιστώντας με αυτό τον τρόπο σχετιζόμενη με τον τομέα αλληλεπίδρασης ανθρώπου-υπολογιστή (HCI - human-computer interaction). Οι τεχνικές της ΕΦΓ επιτρέπουν στους υπολογιστές να επεξεργάζονται και να κατανοούν την ανθρώπινη φυσική γλώσσα και να την χρησιμοποιούν περαιτέρω για να παρέχουν χρήσιμα συμπεράσματα. Στη συνέχεια, παρουσιάζονται μερικές από τις κύριες εφαρμογές της.

A. Machine Translation – Μηχανική μετάφραση

Η μηχανική μετάφραση είναι ίσως μία από τις πιο περιζήτητες εφαρμογές της NLP. Ορίζεται ως η τεχνική που βοηθά στην παροχή συντακτικών, γραμματικών και σημασιολογικά σωστών μεταφράσεων μεταξύ δύο γλωσσών. Αυτός ήταν ίσως ο πρώτος σημαντικός τομέας έρευνας και ανάπτυξης της NLP. Σε βασικό επίπεδο, η μηχανική μετάφραση είναι η μετάφραση της φυσικής γλώσσας που πραγματοποιείται από μια μηχανή. Από προεπιλογή, τα βασικά δομικά στοιχεία για τη διαδικασία μηχανικής μετάφρασης περιλαμβάνουν απλή αντικατάσταση λέξεων από τη μία γλώσσα στην άλλη, αλλά σε αυτή την περίπτωση αγνοούμε πράγματα όπως η γραμματική και η συνέπεια της δομής της διατύπωσης. Ως εκ τούτου, έχουν εξελιχθεί πιο περίπλοκες τεχνικές, που λαμβάνουν υπόψιν τους και συνδυάζουν δεδομένα από μεγάλα αποθετήρια κειμένων μαζί με στατιστικές και γλωσσικές τεχνικές. Ένα από τα πιο δημοφιλή συστήματα μηχανικής μετάφρασης είναι το Google Translate.

Με την πάροδο του χρόνου, τα συστήματα μηχανικής μετάφρασης βελτιώνονται και μπορούν πλέον να παρέχουν την δυνατότητα μεταφράσεων σε πραγματικό χρόνο καθώς μιλάμε ή γράφουμε στην εφαρμογή. Στην παρακάτω Εικόνα 8 δίνεται η μετάφραση του τίτλου της παρούσας διπλωματικής εργασίας.



Εικόνα 8. Μηχανική μετάφραση που πραγματοποιείται από το Google Translate

B. Συστήματα αναγνώρισης ομιλίας

Η αναγνώριση ομιλίας είναι από τις πιο δύσκολες και απαιτητικές εφαρμογές της NLP και των συστημάτων τεχνητής νοημοσύνης. Η πιο απαιτητική δοκιμασία για ανίχνευση πραγματικής νοημοσύνης στα συστήματα τεχνικής νοημοσύνης είναι η δοκιμασία του αγγλού μαθηματικού Alan Turing. Το 1950, σε ένα άρθρο του για την πιθανότητα ανάπτυξης τεχνητής νοημοσύνης, έχοντας ασχοληθεί επί μακρόν με το θέμα και καθώς ο ορισμός της νοημοσύνης αποτελούσε ένα περίπλοκο φιλοσοφικό ζήτημα, πρότεινε το ακόλουθο κριτήριο: εάν μια μηχανή καταφέρει να ξεγελάσει τους ανθρώπους

και να τους κάνει να πιστέψουν πως είναι άνθρωπος, τότε πρέπει να είναι τουλάχιστον εξίσου έξυπνη με έναν άνθρωπο [108]. Ο Turing μάλιστα είχε προβλέψει πως μέχρι το 2000 θα είχε αναπτυχθεί τεχνητή νοημοσύνη που θα μπορούσε να ξεγελάσει το 30% των ερωτώντων, έπειτα από πέντε λεπτά συζήτησης. Κατά τη διάρκεια των τελευταίων δεκαετιών, έχει σημειωθεί μεγάλη πρόοδος σε αυτόν τον τομέα χρησιμοποιώντας τεχνικές όπως η σύνθεση ομιλίας, η ανάλυση, η συντακτική ανάλυση και η συλλογιστική με βάση τα συμφραζόμενα. Ωστόσο, ένας βασικός περιορισμός για τα συστήματα αναγνώρισης ομιλίας εξακολουθεί να παραμένει το γεγονός ότι σε τέτοια συστήματα είναι πολύ συγκεκριμένος ο τομέας των ερωτημάτων και δεν λειτουργεί επαρκώς εάν ο χρήστης απομακρυνθεί έστω και λίγο από τις αναμενόμενες εισόδους (ερωτήματα) που απαιτούνται από το σύστημα. Τα συστήματα αναγνώρισης ομιλίας βρίσκουν στις μέρες μας εφαρμογή σε πολλούς τομείς, από τους υπολογιστές και τα κινητά τηλέφωνα, έως τα συστήματα εικονικής βοήθειας ή/και εξυπηρέτησης.

C. Συστήματα απάντησης ερωτήσεων (QAS – Question Answering Systems)

Τα συστήματα απάντησης ερωτήσεων χτίστηκαν πάνω στην αρχή της Απάντησης Ερωτήσεων, βασισμένα στη χρήση τεχνικών από την NLP και την ανάκτηση πληροφοριών (IR – Information Retrieval). Το QAS ασχολείται κυρίως με τη δημιουργία ισχυρών και κλιμακωτών συστημάτων που παρέχουν απαντήσεις σε ερωτήσεις που δίνονται από χρήστες σε μορφή φυσικής γλώσσας. Αυτή τη στιγμή, έχουν δημιουργηθεί κάποια αρκετά πετυχημένα μοντέλα εξατομικευμένης βοήθειας όπως η Siri της Apple και η Cortana της Microsoft, αλλά το πεδίο εφαρμογής τους εξακολουθεί να είναι περιορισμένο καθώς κατανοούν μόνο ένα υποσύνολο βασικών όρων και φράσεων από το σύνολο της ανθρώπινης φυσικής γλώσσας.

Για της δημιουργία ενός επιτυχημένου QAS, χρειάζεστε μια τεράστια γνωσιακή βάση που θα αποτελείται από δεδομένα σχετικά με διάφορους τομείς. Τα αποτελεσματικά συστήματα ερωτημάτων σε αυτήν τη γνωσιακή βάση θα αξιοποιηθούν από το QAS για να δώσουν απαντήσεις σε ερωτήσεις σε μορφή φυσικής γλώσσας. Η δημιουργία και η διατήρηση μιας τεράστιας γνωσιακής βάσης δεδομένων είναι εξαιρετικά δύσκολη. Ευρεία χρήση συστημάτων QAS παρατηρείτε όλο και πιο συχνά σε εξειδικευμένους τομείς όπως τα τρόφιμα, η υγειονομική περίθαλψη και το ηλεκτρονικό

εμπόριο. Μία από τις αναδυόμενες τάσεις που χρησιμοποιούν εκτενώς τέτοια συστήματα είναι και τα chatbots.

D. Αναγνώριση συμφραζόμενων και Ανάλυση Συσχέτισης

Η εφαρμογή αυτή καλύπτει έναν ευρύ τομέα στην κατανόηση της φυσικής γλώσσας, η οποία περιλαμβάνει συντακτική και σημασιολογική συλλογιστική. Η αποσαφήνιση του συναισθήματος μιας λέξης (contextual recognition) είναι μια δημοφιλής εφαρμογή, καθώς μέσα από αυτή την διαδικασία θέλουμε να βρούμε την χρήση μιας λέξης με βάση τα συμφραζόμενα σε μια δεδομένη πρόταση. Για παράδειγμα η λέξη «book», μπορεί να σημαίνει ένα αντικείμενο που περιέχει γνώσεις και πληροφορίες όταν χρησιμοποιείται ως ουσιαστικό και μπορεί επίσης να σημαίνει ότι «κλείνω» τραπέζι ή δωμάτιο όταν χρησιμοποιείται ως ρήμα. Ο εντοπισμός αυτών των διαφορών στις προτάσεις με βάση το πλαίσιο είναι η κύρια προϋπόθεση της αποσαφήνισης της έννοιας της λέξης και είναι μια δύσκολη εργασία.

Η ανάλυση της συσχέτισης είναι ένα άλλο πρόβλημα στη γλωσσολογία που προσπαθεί να αντιμετωπίσει η NLP. Εξ ορισμού, η συσχέτιση εμφανίζεται όταν δύο ή περισσότερες όροι/εκφράσεις σε ένα σώμα κειμένου αναφέρονται στην ίδια οντότητα (πρόσωπο ή πράγμα). Έτσι, λέμε ότι έχουν την ίδια αναφορά. Σκεφτείτε το παράδειγμα της πρότασης, «John just told me that he is going to the exam hall». Σε αυτή την πρόταση, η αντωνυμία «he» παραπέμπει στον «John». Η αντιστοίχιση αυτή των αντωνυμιών είναι μέρος της ανάλυσης συσχέτισης και γίνεται απαιτητική όταν έχουμε πολλαπλές αναφορές σε ένα κείμενο. Ένα τέτοιο παράδειγμα κειμένου θα ήταν, «John just talked with Jim. He told me we have a surprise test tomorrow». Σε αυτό το κείμενο, η αντωνυμία «he» θα μπορούσε να αναφέρεται είτε στον «John» είτε στον «Jim», καθιστώντας έτσι δύσκολο να προσδιοριστεί η ακριβής αναφορά.

E. Σύνοψη κειμένου (Text summarization)

Ο κύριος στόχος της σύνοψης κειμένου είναι να λάβει ένα σύνολο εγγράφων κειμένου, το οποίο θα μπορούσε να είναι μια συλλογή κειμένων, παραγράφων ή προτάσεων, και να μειώσει το περιεχόμενο κατάλληλα για τη δημιουργία μια περίληψης διατηρώντας τα βασικά σημεία των εγγράφων αυτών.

Η σύνοψη πραγματοποιείται εξετάζοντας τα διάφορα έγγραφα και αναζητώντας τις λέξεις-κλειδιά, τις φράσεις και τις προτάσεις που έχουν εξέχουσα θέση στα έγγραφα. Δύο

βασικές τεχνικές για την σύνοψη κειμένου είναι η σύνοψη βάσει εξαγωγής και η σύνοψη βάσει αφαίρεσης. Με την εμφάνιση τεράστιων ποσοτήτων κειμένων και μη δομημένων δεδομένων, η ανάγκη για την δημιουργία σύνοψης κειμένου, με στόχο την γρήγορη λήψη πολύτιμων πληροφοριών έχει μεγάλη ζήτηση.

Τα συστήματα σύνοψης κειμένου συνήθως εκτελούν δύο κύριους τύπους λειτουργιών. Η πρώτη είναι η γενική σύνοψη, η οποία προσπαθεί να παράσχει μια γενική περίληψη της συλλογής των εγγράφων υπό ανάλυση. Ο δεύτερος τύπος λειτουργίας είναι η σύνοψη βάσει ερωτήματος, η οποία παρέχει περιλήψεις κειμένου σχετικές με ερωτήματα όπου το σύνολο των εγγράφων φιλτράρεται περαιτέρω με βάση συγκεκριμένα ερωτήματα.

F. Κατηγοριοποίηση κειμένου (Text categorization)

Ο κύριος στόχος της κατηγοριοποίησης κειμένου είναι να προσδιορίσει σε ποια κατηγορία ή κλάση θα πρέπει να τοποθετηθεί ένα συγκεκριμένο έγγραφο με βάση το περιεχόμενο του εγγράφου. Αυτή είναι μια από τις πιο δημοφιλείς εφαρμογές της επεξεργασίας φυσικής γλώσσας και της μηχανικής μάθησης. Τόσο οι εποπτευόμενες όσο και οι μη επιτηρούμενες τεχνικές μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την επίλυση αυτού του προβλήματος και σε αρκετές περιπτώσεις χρησιμοποιείται ένας συνδυασμός και των δύο. Αυτό έχει βοηθήσει στη δημιουργία πολλών επιτυχημένων και πρακτικών εφαρμογών, συμπεριλαμβανομένων φίλτρων ανεπιθύμητης αλληλογραφίας και κατηγοριοποίησης άρθρων ειδήσεων.

G. Ανάλυση κειμένου (Text Analysis)

Η ανάλυση κειμένου, επίσης γνωστή ως εξόρυξη κειμένου, ορίζεται ως η μεθοδολογία και η διαδικασία που ακολουθείται για την εξαγωγή ποιοτικών και εφαρμόσιμων πληροφοριών και πληροφοριών από δεδομένα κειμένου. Αυτό περιλαμβάνει τη χρήση τεχνικών επεξεργασίας φυσικής γλώσσας, ανάκτησης πληροφοριών και μηχανικής μάθησης για την ανάλυση μη δομημένων δεδομένων κειμένου σε πιο δομημένες φόρμες και την εξαγωγή μοτίβων και πληροφοριών από αυτά που θα ήταν χρήσιμα για τον τελικό χρήστη. Η ανάλυση κειμένου περιλαμβάνει μια συλλογή τεχνικών μηχανικής μάθησης, γλωσσολογίας και στατιστικής που χρησιμοποιούνται για τη μοντελοποίηση και εξαγωγή πληροφοριών από το κείμενο που υπόκεινται σε ανάλυση, συμπεριλαμβανομένης της επιχειρηματικής ευφυΐας, της

διερευνητικής, της περιγραφικής και της προγνωστικής ανάλυσης. Ορισμένες από τις κύριες τεχνικές και λειτουργίες στην ανάλυση κειμένου είναι οι ακόλουθες:

- Ταξινόμηση κειμένου (Text classification)
- Ομαδοποίηση κειμένου (Text clustering)
- Σύνοψη κειμένου (Text summarization)
- Ανάλυση συναισθημάτων (Sentiment analysis)
- Εξαγωγή και αναγνώριση οντότητας (Entity extraction and recognition)
- Ανάλυση ομοιότητας και μοντελοποίηση σχέσεων (Similarity analysis and relation modeling)

Η ανάλυση κειμένου είναι μια αρκετά πιο πολύπλοκη διαδικασία σε σύγκριση με την κανονική στατιστική ανάλυση ή τη μηχανική μάθηση. Ο λόγος είναι ότι πριν εφαρμόσουμε μια τεχνική μάθησης ή έναν αλγόριθμο, πρέπει να μετατρέψουμε τα μη δομημένα δεδομένα κειμένου σε μορφή αποδεκτή από αυτούς τους αλγόριθμους. Εξ ορισμού, ένα σώμα κειμένου υπό ανάλυση συχνά ονομάζεται έγγραφο και εφαρμόζοντας διάφορες τεχνικές, το μετατρέπουμε σε διανύσματα λέξεων. Αυτός είναι συνήθως ένας αριθμητικός πίνακας του οποίου οι τιμές είναι συγκεκριμένα βάρη για κάθε λέξη, η οποία θα μπορούσε να είναι η συχνότητά της, η εμφάνισή της ή διάφορες άλλες απεικονίσεις. Συχνά το κείμενο πρέπει να καθαριστεί και να υποβληθεί σε επεξεργασία ώστε να αφαιρεθούν όροι και δεδομένα που δημιουργούν θόρυβο δυσκολεύοντας την ανάλυση. Αυτή η διαδικασία ονομάζεται προεπεξεργασία κειμένου. Μόλις έχουμε τα δεδομένα σε μορφή που μπορεί να αναγνώσει και να κατανοήσει ο ηλεκτρονικός υπολογιστής, μπορούμε να προχωρήσουμε στην εφαρμογή των σχετικών αλγορίθμων με βάση το πρόβλημα που πρέπει να επιλυθεί. Οι εφαρμογές ανάλυσης κειμένου είναι πολλαπλές, οι πιο δημοφιλείς από τις οποίες (ορισμένες από τις οποίες παρουσιάστηκαν παραπάνω) είναι οι εξής.

- Εντοπισμός ανεπιθύμητων μηνυμάτων (Spam detection)
- Κατηγοριοποίηση άρθρων ειδήσεων (News articles categorization)
- Ανάλυση και παρακολούθηση των κοινωνικών μέσων (Social media analysis and monitoring)
- Βιοϊατρική (Biomedical)
- Πληροφορίες ασφαλείας (Security intelligence)
- Μάρκετινγκ και CRM
- Ανάλυση συναισθημάτων (Sentiment analysis)

- Τοποθετήσεις διαφημίσεων (Ad placement)
- Chatbots
- Εικονικοί βοηθοί

Η. Μηχανική Μάθηση (Machine Learning - ML)

Η μηχανική μάθηση αποτελεί υποπεδίο της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Arthur Samuel ορίζει τη μηχανική μάθηση ως «Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί» [109] Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα [110] και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με την υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος.. Συνήθως ένας συνδυασμός NLP και ML είναι συχνά απαραίτητος για την επίλυση προβλημάτων του πραγματικού κόσμου, όπως κατηγοριοποίηση κειμένου, ομαδοποίηση και ούτω καθεξής. Οι τρεις κύριες κατηγορίες τεχνικών μηχανικής μάθησης περιλαμβάνουν εποπτευόμενους, μη επιτηρημένους αλγορίθμους, καθώς και αλγορίθμους ενισχυτικής μάθησης.

Ι. Βαθιά Μάθηση (Deep Learning - DL)

Ο τομέας της βαθιάς μάθησης είναι ένα υποπεδίο της μηχανικής μάθησης που ειδικεύεται σε μοντέλα και αλγορίθμους, οι οποίοι έχουν εμπνευστεί από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Πράγματι, το τεχνητό νευρωνικό δίκτυο (artificial neural network - ANN) ήταν το πρώτο μοντέλο που κατασκευάστηκε αντλώντας έμπνευση από τον ανθρώπινο εγκέφαλο. Αν και είμαστε αρκετά μακριά από την αναπαραγωγή του τρόπου λειτουργίας του εγκεφάλου, τα νευρωνικά δίκτυα είναι εξαιρετικά περίπλοκα, μη γραμμικά, μοντέλα, τα οποία είναι ικανά να μαθαίνουν αυτόματα ιεραρχικές αναπαραστάσεις δεδομένων. Η βαθιά μάθηση ή τα βαθιά

νευρωνικά δίκτυα συνήθως χρησιμοποιούν πολλαπλά στρώματα μη γραμμικών μονάδων επεξεργασίας, επίσης γνωστά ως νευρώνες, ή προτιμότερα «μονάδες επεξεργασίας». Από κάθε επίπεδο εξάγεται μια σειρά από γνωρίσματα, τα οποία επεξεργάζονται και μετασχηματίζονται από μόνη τους χρησιμοποιώντας την έξοδο από την προηγούμενη στρώση ως είσοδο. Ως εκ τούτου, κάθε επίπεδο καταλήγει να μαθαίνει ιεραρχικές αναπαραστάσεις των δεδομένων σε διαφορετικά αφαιρετικά επίπεδα. Μπορούμε να χρησιμοποιήσουμε αυτά τα μοντέλα για να λύσουμε τόσο εποπτευόμενα όσο και μη εποπτευόμενα προβλήματα. Πρόσφατα, η βαθιά μάθηση έχει δείξει πολλές υποσχέσεις όσον αφορά την επίλυση προβλημάτων επεξεργασίας φυσικής γλώσσας.

3.7.1.Γλωσσολογία

Η γλωσσολογία ορίζεται ως η επιστημονική μελέτη της γλώσσας, συμπεριλαμβανομένης της μορφής και της σύνταξης της γλώσσας, της έννοιας και της σημασιολογίας που απεικονίζονται από τη χρήση της γλώσσας και του πλαισίου χρήσης. Η προέλευση της γλωσσολογίας μπορεί να χρονολογηθεί από τον 4ο αιώνα π.Χ., όταν ο Ινδός λόγιος και γλωσσολόγος Panini επισημοποίησε την περιγραφή της σανσκριτικής γλώσσας.

Ο όρος γλωσσολογία χρησιμοποιήθηκε για πρώτη φορά το 1847, για να υποδηλώσει την επιστημονική μελέτη των γλωσσών. Οι διαφορετικοί τομείς μελέτης της γλωσσολογίας που χρησιμοποιούνται εκτενώς στην επεξεργασία φυσικής γλώσσας είναι οι εξής:

1. Φωνητική: Πρόκειται για τη μελέτη των ακουστικών ιδιοτήτων των ήχων που παράγονται από την ανθρώπινη φωνητική οδό κατά τη διάρκεια μιας ομιλίας.
2. Φωνολογία: Αυτή είναι η μελέτη των ηχητικών μοτίβων όπως ερμηνεύονται στο ανθρώπινο μυαλό και χρησιμοποιούνται για τη διάκριση μεταξύ διαφορετικών φωνημάτων.
3. Σύνταξη: Αυτή είναι συνήθως η μελέτη των προτάσεων, των φράσεων, των λέξεων και των δομών τους. Αυτό περιλαμβάνει την έρευνα του τρόπου με τον οποίο οι λέξεις συνδυάζονται γραμματικά για να σχηματίσουν φράσεις και προτάσεις. Συντακτική σειρά λέξεων που χρησιμοποιούνται σε μια φράση ή μια πρόταση, καθώς η σειρά μπορεί να αλλάξει εντελώς το νόημα.
4. Σημασιολογία: Αυτή περιλαμβάνει τη μελέτη του νοήματος στη γλώσσα και μπορεί να υποδιαιρεθεί περαιτέρω σε λεξική και συνθετική σημασιολογία.

- Λεξική σημασιολογία: Αυτή περιλαμβάνει τη μελέτη των εννοιών των λέξεων και των συμβόλων χρησιμοποιώντας μορφολογία και σύνταξη.
- Συνθετική σημασιολογία: Αυτή περιλαμβάνει τη μελέτη των σχέσεων μεταξύ των λέξεων και του συνδυασμού των λέξεων και την κατανόηση της έννοιας των φράσεων και των προτάσεων και του τρόπου με τον οποίο σχετίζονται.

6. Μορφολογία: Εξ ορισμού, ένα μόρφημα είναι η μικρότερη μονάδα της γλώσσας που έχει διακριτικό νόημα. Αυτό περιλαμβάνει πράγματα όπως λέξεις, προθέματα, επιθήματα και ούτω καθεξής, τα οποία έχουν τη δική τους ξεχωριστή σημασία. Η μορφολογία είναι η μελέτη της δομής και της σημασίας αυτών των διακριτικών μονάδων ή μορφωμάτων σε μια γλώσσα. Υπάρχουν συγκεκριμένοι κανόνες και συντάξεις που διέπουν τον τρόπο με τον οποίο μπορούν να συνδυαστούν τα μορφώματα.
7. Λεξικό: Αυτή είναι η μελέτη των ιδιοτήτων των λέξεων και των φράσεων που χρησιμοποιούνται σε μια γλώσσα και πώς χτίζουν το λεξιλόγιο της γλώσσας. Αυτά περιλαμβάνουν τα είδη των ήχων που συνδέονται με τις έννοιες για τις λέξεις, καθώς και τα μέρη του λόγου στα οποία ανήκουν οι λέξεις και τις μορφολογικές τους μορφές.
8. Πραγματολογία: Αυτή είναι η μελέτη του τρόπου με τον οποίο γλωσσικοί και μη γλωσσικοί παράγοντες όπως το πλαίσιο και το σενάριο μπορεί να επηρεάσουν την έννοια μιας έκφρασης ενός μηνύματος ή μιας εκφώνησης. Αυτό περιλαμβάνει την προσπάθεια να συμπεράνουμε εάν υπάρχουν κρυφές ή έμμεσες έννοιες στην επικοινωνία.
9. Ανάλυση λόγου: Αυτή αναλύει τη γλώσσα και την ανταλλαγή πληροφοριών με τη μορφή προτάσεων σε συνομιλίες μεταξύ ανθρώπων. Αυτές οι συνομιλίες θα μπορούσαν να εκφωνηθούν, να γραφτούν ή ακόμα και να υπογραφούν.
10. Υφολογία: Αυτή είναι η μελέτη της γλώσσας με έμφαση στο στυλ γραφής, συμπεριλαμβανομένου του τόνου, της προφοράς, του διαλόγου, της γραμματικής και του τύπου της φωνής.
11. Σημειολογία: Αυτή είναι η μελέτη των σημείων, των συμβόλων και των διαδικασιών των νοημάτων και του τρόπου με τον οποίο επικοινωνούν το νόημα. Πράγματα όπως αναλογίες, μεταφορές και συμβολισμοί καλύπτονται σε αυτόν τον τομέα.

4. Εξόρυξη προφίλ χρηστών (User profiling) και τύποι προσωπικότητας

4.1. Εισαγωγή

Από την αρχαιότητα έως και σήμερα, κάθε κοινωνία απαρτίζεται από κατηγορίες ανθρώπων με κοινά χαρακτηριστικά, αξίες, πεποιθήσεις και κοινωνικές συμπεριφορές, που συχνά έχουν και ψυχολογικές διαστάσεις. Στο σύγχρονο κόσμο, η εμφάνιση του διαδικτύου και των μέσων κοινωνικής δικτύωσης συνοδεύτηκε από την διακίνηση τεράστιου όγκου πληροφοριών και δεδομένων από τους χρήστες, δίνοντας στους ερευνητές διαφόρων επιστημονικών κλάδων, αλλά και σε επαγγελματίες του μάρκετινγκ, την δυνατότητα να πραγματοποιήσουν μελέτες με αντικείμενο την εξόρυξη του προφίλ των χρηστών αυτών και να δώσουν απαντήσεις σε προβλήματα ή/και σε ερωτήματα που σχετίζονται τόσο με τον τύπο την προσωπικότητά τους, όσο και με τις καταναλωτικές τους συνήθειες.

Για την ανίχνευση του τύπου προσωπικότητα, έχουν αναπτυχθεί διάφορα ψυχολογικά μοντέλα όπως είναι το Big 5 ή OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism), το οποίο περιγράφεται στη συνέχεια.

4.2. Εξόρυξη προφίλ χρηστών (User profiling)

Η εξόρυξη προφίλ χρηστών (User Profiling) αποτελεί υποπεδίο της Τεχνητής Νοημοσύνης (Artificial Intelligence, AI). Σύμφωνα με το αγγλικό λεξικό της Οξφόρφης υπάρχουν δύο ορισμοί για τον όρο «user profile» [111]

1ος. Είναι ο υπολογισμός των μοναδικών διαμορφώσεων, προτιμήσεων, ρυθμίσεων κ.λπ., που έχουν ρυθμιστεί για ή από χρήστη υπολογιστή, ειδικά όπως αποθηκεύονται σε έναν διακομιστή και είναι προσβάσιμες μέσω διαφόρων υπολογιστών του δικτύου.

2ος. Είναι η συλλογή πληροφοριών ή δεδομένων σχετικά με τις συνήθειες, τις προτιμήσεις κ.λπ, ενός χρήση, ιδίως για ένα προϊόν ή μια υπηρεσία.

Με βάση τους παραπάνω δύο ορισμούς προκύπτει ότι βασικός στόχος του user profiling είναι η κατανόηση σχετικά με έναν χρήστη και τις προτιμήσεις του με βάση τις πληροφορίες που λαμβάνονται σχετικά με αυτόν. Με άλλα λόγια, η εξόρυξη προφίλ χρήστη μπορεί να περιγραφεί ως μια διαδικασία απόκτησης δεδομένων, εκτέλεσης οποιασδήποτε απαιτούμενης επεξεργασίας σε αυτά για την παραγωγή ενός ολοκληρωμένου μοντέλου ή αναπαράστασης ενός χρήστη ή μιας ομάδας χρηστών[112]

4.2.1. Big 5 - OCEAN (Openess-Conscientiousness-Extraversion-Agreebleness-Neuroticism)

Η προσωπικότητα είναι ένα σύνολο ιδιαίτερων χαρακτηριστικών που διαφοροποιούν το ένα άτομο από το άλλο [113]. Αυτές οι διαφορές αντικατοπτρίζονται στην αντίληψή τους για τον περιβάλλοντα κόσμο, στις σκέψεις, τις πράξεις τους, αλλά και στις λέξεις που χρησιμοποιούν για να περιγράψουν τι αισθάνονται και τι σκέφτονται την κάθε δεδομένη στιγμή του χρόνου. Υπάρχουν διάφορες προσεγγίσεις ανάλυσης λέξεων για την πρόβλεψη των χαρακτηριστικών προσωπικότητας, αλλά το πιο ευρέως ερευνώμενο και περισσότερο χρησιμοποιούμενο σε προηγούμενες εργασίες είναι το μοντέλο Big Five ή OCEAN από τα αρχικά των λέξεων Openness (Ανοιχτός σε εμπειρίες), Conscientiousness (Ευσυνειδησία), Extraversion (Εξωστρέφεια), Agreeableness (Τερπνότητα) and Neuroticism (Νευρωτισμός).

Το μοντέλο των 5 παραγόντων, γνωστό και ως μεγάλη πεντάδα ή ως μοντέλο OCEAN, είναι μια ταξινόμηση για τα χαρακτηριστικά της προσωπικότητας[114]. Αυτό το μοντέλο ορίστηκε από διάφορες ανεξάρτητες ομάδες ερευνητών που χρησιμοποίησαν την ανάλυση παραγόντων των λεκτικών περιγραφέων (verbal descriptors) της ανθρώπινης συμπεριφοράς[115] Αυτοί οι ερευνητές ξεκίνησαν μελετώντας τις σχέσεις μεταξύ ενός μεγάλου αριθμού λεκτικών περιγραφέων που σχετίζονται με χαρακτηριστικά προσωπικότητας. Μείωσαν τους καταλόγους αυτών των περιγραφέων κατά 5-10 φορές και στη συνέχεια χρησιμοποίησαν την ανάλυση παραγόντων για να ομαδοποιήσουν τα υπόλοιπα χαρακτηριστικά (χρησιμοποιώντας δεδομένα που βασίζονται κυρίως στις εκτιμήσεις των ανθρώπων, σε ερωτηματολόγιο αυτοαναφοράς και αξιολογήσεις από συνεργάτες) προκειμένου να βρουν τους υποκείμενους παράγοντες της προσωπικότητας[116]–[120]

Το αρχικό μοντέλο προτάθηκε από τους Ernest Tupes και Raymond Christal το 1961 [119], αλλά απέτυχε να προσεγγίσει ένα ακαδημαϊκό κοινό μέχρι τη δεκαετία του 1980. Το 1990, ο J.M. Digman [121]πρότεινε το μοντέλο προσωπικότητας πέντε παραγόντων, το οποίο ο Lewis Goldberg επέκτεινε σε υψηλότερο επίπεδο οργάνωσης [122]. Αυτοί οι πέντε γενικοί παράγοντες έχουν βρεθεί ότι περιέχουν και ενσωματώνουν τα πιο γνωστά χαρακτηριστικά της προσωπικότητας και θεωρείται ότι αντιπροσωπεύουν τη βασική δομή πίσω από όλα τα χαρακτηριστικά της προσωπικότητας [123].

Οι μελέτες δείχνουν ότι τα χαρακτηριστικά του μοντέλου των 5 παραγόντων δεν είναι τόσο ισχυρά στην πρόβλεψη και την εξήγηση της πραγματικής συμπεριφοράς όσο

είναι οι πιο πολυάριθμες πτυχές ή τα κύρια χαρακτηριστικά [124] [125]

Σύμφωνα με την θεωρία υπάρχουν οι εξής 5 παράγοντες που αντιστοιχούν στους 5 διαφορετικούς τύπους προσωπικότητας:

1. **Δεκτικότητα σε εμπειρίες (**Openness) (εφευρετικός/περίεργος έναντι του συνεπής/καχύποπτος)
2. **Ευσυνειδησία** (Conscientiousness) (αποτελεσματικός/οργανωτικός έναντι του υπερβολικός/άμελοςς)
3. **Εξωστρέφεια** (Extraversion) (κοινωνικός/δραστήριος έναντι του μοναχικός/συνεσταλμένος)
4. **Τερπνότητα** (Agreeableness) (φιλικός/συμπονετικός έναντι του επικριτικός/λογικός)
5. **Νευρωτισμός** (Neuroticism) (ευαίσθητος/νευρικός έναντι του ανθεκτικός/με αυτοπεποίθηση) [126]

Κάθε ένα από τα χαρακτηριστικά της προσωπικότητας των Big Five περιέχει δύο ξεχωριστές, αλλά συσχετιζόμενες, πτυχές που αντικατοπτρίζουν ένα επίπεδο προσωπικότητας υπό μια συγκεκριμένη κατηγορία, αλλά πάνω σε μια σειρά από όψεις που αποτελούν επίσης μέρος των Big Five [123]. Οι πτυχές επισημαίνονται ως εξής:

- Διάνοια και ειλικρίνεια για την δεκτικότητα στην εμπειρία.
- Εργατικότητα και ευταξία για ευσυνειδησία.
- Ενθουσιασμός και διεκδικητικότητα για την εξωστρέφεια.
- Συμπόνια και ευγένεια για την τερπνότητα.
- Μεταβλητότητα και υπαναχώρηση για τον νευρωτισμό. [127]

Οι άνθρωποι που δεν παρουσιάζουν σαφή προδιάθεση για έναν μόνο παράγοντα σε κάθε παραπάνω διάσταση θεωρούνται ευπροσάρμοστοι, μετριοπαθείς και λογικοί, αλλά μπορούν επίσης να θεωρηθούν ως ανήθικοι, ακατανόητοι και ραδιούργοι. [128]

- Οι άνθρωποι που ανήκουν στην κατηγορία *Openness* είναι ανοιχτοί χαρακτήρες με διάθεση να βιώσουν νέες εμπειρίες, πρόθυμοι να δοκιμάσουν νέα πράγματα, διαθέτουν πλούσια φαντασία, είναι δημιουργικοί και περίεργοι, με καλή αισθητική (Farnadi et al., 2014)[129] και πιο συνειδητοποιημένοι για τα συναισθήματά τους σε σχέση με τους ανθρώπους των υπόλοιπων κατηγοριών.

- Οι άνθρωποι της κατηγορίας *Conscientiousness,* αγαπούν την εργασία, είναι οργανωτικοί, ειλικρινείς, αξιόπιστοι, τείνουν να κάνουν σχέδια και επικεντρώνονται στην επίτευξη των στόχων τους.

- Οι άνθρωποι με υψηλή *εξωστρέφεια (Extraversion)* σύμφωνα με τους Benet-

Martinez και ο John (1998) [130]είναι δραστήριοι και γεμάτοι ενέργεια, διαθέτουν κυριαρχία, είναι κοινωνικοί και διακατέχονται από θετικά συναισθήματα. Τα εξωστρεφή άτομα τείνουν να συνάπτουν φιλίες εύκολα, τους αρέσει να μιλούν και να είναι το κέντρο της προσοχής και επίσης συνήθως συμμετέχουν σε κοινωνικές δραστηριότητες.

- Χαρακτηριστικά όπως ο αλτρουισμός, η τρυφερότητα, η εμπιστοσύνη και η μετριοφροσύνη χαρακτηρίζουν τους ανθρώπους της κατηγορίας Agreeableness. Οι άνθρωποι αυτού του τύπου προσωπικότητας είναι ευχάριστα άτομα, και παρόμοια με τους εξωστρεφείς, τείνουν να εκπέμπουν θετικά συναισθήματα, αποφεύγοντας να εκφράσουν αρνητικότητα, ενώ αγαπούν να βοηθούν τους συνανθρώπους τους και να προσαρμόζονται στις ανάγκες τους.
- *Ο νευρωτισμός (Neuroticism)* συνδυάζει μια μεγάλη ποικιλία αρνητικών συναισθημάτων όπως το άγχος, η θλίψη, ευεραισθησία και τα νεύρα. Αυτός ο τύπος ανθρώπων τείνει να είναι καταθλιπτικός, να έχει απρόβλεπτη διάθεση και επίσης να χρησιμοποιεί λέξεις που αντανακλούν αρνητικές σκέψεις και συναισθήματα.

4.3. Εφαρμογές εξόρυξης προφίλ χρηστών

Οι κύριες εφαρμογές ενδιαφέροντος από την εξόρυξη προφίλ χρηστών είναι εκείνες που αφορούν δεδομένα καταναλωτών, καθώς επιφέρουν στους οργανισμούς και τις εταιρίες επιπλέον πληροφορίες για την επίτευξη των στόχων τους, όπως η δημιουργία εσόδων, η δημιουργία συνθηκών δέσμευσής των καταναλωτών με τα προϊόντα ή/και τις υπηρεσίες τους, αλλά και την ανάπτυξη στοχευμένων διαφημιστικών ενεργειών. Οι ηγέτες στον τομέα της έρευνας στο τομές της εξόρυξης προφίλ χρηστών είναι τα Διαδικτυακά Κοινωνικά Δίκτυα (Online Social Networks), όπως το Facebook, το Instagram, το Twitter, το LinkedIn και το YouTube. Η εγγενής πρόσβασή τους σε λεπτομερείς πληροφορίες διασύνδεσης χρηστών σε πραγματικό χρόνο τους επιτρέπει να αναπτύξουν καλά κατασκευασμένα και πλήρως εκπαιδευμένα μοντέλα εξόρυξης προφίλ χρηστών. Οι οργανισμοί που είναι οπλισμένοι με τέτοιες γνώσεις είναι σε θέση να παρέχουν υπηρεσίες που είναι πιο προσαρμοσμένες στις επιθυμίες και τις ανάγκες των καταναλωτών, χρησιμοποιώντας περισσότερο στοχευμένες και σχετικές διαφημίσεις.

4.4. Σχετικές εργασίες

Προηγούμενες εργασίες στην εξόρυξη προφίλ χρηστών και την ανίχνευση του τύπου προσωπικότητας έχουν καταλήξει στο συμπέρασμα ότι τα δεδομένα που συλλέγονται

από τα κοινωνικά μέσα, αποτελούν ουσιαστικά αντανάκλαση του ανθρώπου πίσω από τον λογαριασμό του. Ως εκ τούτου, τα τελευταία χρόνια, έχει διεξαχθεί μια μεγάλη ποικιλία μελετών σχετικά με την ανάλυση των μέσων κοινωνικής δικτύωσης και τη ομαδοποίηση/ταξινόμηση/κατηγοριοποίηση των χρηστών με βάση διαφορετικά χαρακτηριστικά, καθώς και την συσχέτιση μεταξύ του τύπου προσωπικότητας σύμφωνα με την κατηγοριοποίηση OCEAN και των γραπτό λόγο. Στη συνέχεια παρατίθενται ορισμένες από τις μελέτες αυτές συνοδευόμενες από μια σύντομη περιγραφή.

Σύμφωνα με την μελέτη του Pennebaker (1999) [131] γίνεται σαφές ότι οι κατηγορίες των λέξεων που χρησιμοποιούνται σε καθημερινές συζητήσεις συνδέονται με τα OCEAN χαρακτηριστικά προσωπικότητας. Εάν ένα άτομο τείνει να χρησιμοποιεί πολλές λέξεις και οι περισσότερες από αυτές είναι μεγαλύτερες από έξι γράμματα και ταυτόχρονα τείνει να αποφεύγει κατηγορίες όπως άρθρα, αιτιώδης και κοινωνικές λέξεις, αρνήσεις θα μπορούσε να ανήκει στον τύπο του εξωστρεφή (extraversion). Έχει παρατηρηθεί ότι τα λόγια των εξωστρεφών ανθρώπων εκφράζουν περισσότερα θετικά συναισθήματα από ότι αρνητικά, ενώ οι άνθρωποι με κατάθλιψη χρησιμοποιούν κυρίως αντωνυμίες πρώτου προσώπου και εκφράζουν αρνητικά συναισθήματα.

Οι Argamon et al. (2005)[132], επισήμαναν στην μελέτη τους την δυσκολία πρόβλεψης του τύπου της εξωστρέφειας, σε σχέση με τον τύπο του νευρωτισμού. Σύμφωνα τους συγγραφείς, οι εξωστρεφείς άνθρωποι επικεντρώνονται περισσότερο σε λέξεις σχετικά με την πληρότητα/έλλειψη, τη βεβαιότητα/ αβεβαιότητα, ενώ οι νευρωτικοί άνθρωποι ασχολούνται με τον εαυτό τους, κάτι που είναι εμφανές από την χρήση μεγάλου αριθμού αντωνυμιών πρώτου προσώπου στη καθημερινή ομιλία τους.

Στις εργασίας τους οι Tausczik και Pennebaker (2010b)[86] εξηγούν τον τρόπο με τον οποίο η καθημερινή χρήση των λέξεων μπορεί να χαρακτηρίσει ένα άτομο ως προς τον τρόπο σκέψης, την εστίαση της προσοχής του, την συναισθηματικότητα, τις κοινωνικές σχέσεις με τους άλλους. Έτσι, η ανάλυση αυτών των γλωσσικών σημάτων επιτρέπει μια κατά προσέγγιση πρόβλεψη της κατάστασης ψυχικής υγείας των ανθρώπων.

Ο Yarkoni (2010)[133] στην μελέτη του συμπέρανε την ύπαρξη συσχέτισης μεταξύ των προτιμήσεων στη χρήση λέξεων και τον τύπο προσωπικότητας ενός ατόμου. Προηγούμενες παρόμοιες έρευνες είχαν τρία βασικά ελαττώματα που τις καθιστούσαν ανακριβείς. Το πρώτο ήταν ότι βασίζονταν στην ανάλυση κειμένων που αφορούσαν θέματα που επέλεγαν οι συμμετέχοντες στην έρευνα. Το δεύτερο αφορούσε το σύντομο χρονικό διάστημα κατά το οποίο ελήφθησαν τα δείγματα ομιλίας, και τρίτον οι μελέτες

αυτές λάμβαναν υπόψιν μόνο τα βασικά χαρακτηριστικά του OCEAN. Έτσι, μία από τις κύριες διαφορές αυτής της μελέτης ήταν ότι σε αυτή μελετήθηκαν συσχετισμοί μεταξύ γλωσσικών αναφορών όχι μόνο με τα OCEAN χαρακτηριστικά, αλλά και με τις χαμηλού επιπέδου όψεις τους. Για την συλλογή πληροφοριών όπως η ηλικία, το φύλο, η προσωπικότητα των συμμετεχόντων χρησιμοποιήθηκαν ερωτηματολόγια, ενώ οι συμμετέχοντες που επιλέχθηκαν για το πείραμα ήταν μόνο εκείνοι οι bloggers που άφηναν την διεύθυνση ηλεκτρονικού ταχυδρομείου διαθέσιμη στο κοινό και μόνο εκείνοι που είχαν απαντήσει σε mail που του στάλθηκε. Έτσι, το γεγονός ότι ορισμένοι τύποι προσωπικότητας είναι πιο πιθανό να επικοινωνήσουν μέσω ηλεκτρονικού ταχυδρομείου και πιο πιθανό να ανταποκριθούν σε σχέση με άλλους, καθιστά αυτή τη μέθοδο επιλογής, και ως εκ τούτου, τα αποτελέσματα της έρευνας, όχι και τόσο ακριβή όσο ήταν θα ήταν επιθυμητό. Μετά την επιλογή των ιστολογίων, έγινε μια ανάλυση βάσει κατηγοριών, κατά τη διάρκεια της οποίας αναλύθηκαν 66 κατηγορίες του λεξικού LIWC και τα αποτελέσματα έδειξαν ισχυρές συσχετίσεις μεταξύ των OCEAN χαρακτηριστικών και τη συχνότητα χρήσης λέξεων από διαφορετικές κατηγορίες του λεξικού LIWC. Οι συγγραφείς κατέληξαν στο συμπέρασμα ότι μια προσωπικότητα είναι ένας σημαντικός παράγοντας που επηρεάζει είτε τη συμπεριφορά ενός ατόμου στον εικονικό κόσμο όσο και τη συμπεριφορά του στον πραγματικό κόσμο.

Ο D. Quercia et al. (2011) [134] πρότεινε μια μέθοδο για την ανάλυση των tweets ενός χρήστη του Twitter για να συμπεράνει τα OCEAN χαρακτηριστικά της προσωπικότητας του. Τα δεδομένα για τη μελέτη τους συλλέχθηκαν χρησιμοποιώντας το Twitter API και περιορίστηκαν σε μερικές εκατοντάδες χρήστες του Twitter που μοιράστηκαν τη βαθμολογία της προσωπικότητάς τους από μια εφαρμογή του Facebook που ονομάζεται MyPersonality.

Η μελέτη των Qiu et al. (2012) [135] είχε ως στόχο να μετρήσει τα OCEAN χαρακτηριστικά ερευνώντας τη σχέση μεταξύ αυτών και ορισμένων γλωσσικών χαρακτηριστικών που εμφανίζονται σε tweets. Η αξιολόγηση του τύπου προσωπικότητας των χρηστών του twitter, πραγματοποιήθηκε από κριτές που προσλήφθηκαν γι' αυτή την εργασία. Η γλωσσική ανάλυση για την πρόβλεψη της προσωπικότητας έγινε χρησιμοποιώντας την εφαρμογή LIWC2007. Τα αποτελέσματα έδειξαν ότι η *εξωστρέφεια* (*extraversion*) συνδέεται στενά με τη χρήση λέξεων που σχετίζονται με τις κοινωνικές διαδικασίες και τη χρήση θετικών λέξεων συναισθήματος και ταυτόχρονα συσχετίζεται αρνητικά με τη χρήση άρθρων. Επιπλέον, οι εξωστρεφείς άνθρωποι αποφεύγουν τη χρήση σύνθετων λεξικών δομών. Οι *ευχάριστοι άνθρωποι* (*agreeable)*

αποφεύγουν να χρησιμοποιούν αρνήσεις, οι *νευρωτικοί (*neurotic) τείνουν να επικεντρώνονται στον εαυτό τους, ενώ κατηγορία των ανοιχτών ανθρώπων (openness) συσχετίζεται αρνητικά με τις βρισιές, την επιρροή και τις λέξεις χωρίς ευχέρεια, αλλά συσχετίζεται έντονα με τη χρήση προθέσεων.

Ο R. Wald et al. (2012) [136] πρότεινε μια μέθοδο για την εξαγωγή χαρακτηριστικών προσωπικότητας ενός χρήστη με βάση τα δεδομένα του Facebook, χρησιμοποιώντας δημογραφικά και βασισμένα σε κείμενο χαρακτηριστικά που εξάγονται από τα προφίλ χρηστών του Facebook. Η μελέτη αυτή καταλήγει στην διαπίστωση ότι τα αποτελέσματα της έρευνας «επιτρέπουν στους διαφημιζόμενους και σε άλλες ομάδες να επικεντρωθούν σε ένα συγκεκριμένο υποσύνολο ατόμων με βάση τα χαρακτηριστικά της προσωπικότητάς τους».

Η προσέγγιση που περιγράφεται σε μια μελέτη των Schwartz et al. (2013)[137] επιτρέπει την εξαγωγή χαρακτηριστικών προσωπικότητας, γνωρίζοντας την ηλικία, την τοποθεσία και τα ψυχολογικά χαρακτηριστικά που αποκτώνται με την ανάλυση των δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης. Η μέθοδος που χρησιμοποιήθηκε στη μελέτη ονομάστηκε «ανάλυση του ανοικτού λεξιλογίου», επειδή το λεξικό βασίζεται στις λέξεις που χρησιμοποιούνται στις δημόσιες τοποθετήσεις των χρηστών και όχι σε προκαθορισμένες κατηγορίες λέξεων. Κατά την έρευνα αυτή αναλύθηκαν 15,4 εκατομμύρια μηνύματα στο Facebook από 75.000 χρήστες. Για την σύνδεση των κατηγοριών των λέξεων που αναφέρονται στις δημοσιεύσεις με τους τύπους προσωπικότητας και τα υπόλοιπα χαρακτηριστικά του χρήση έγινε χρήση της μεθόδου των ελαχίστων τετραγώνων (least squares regression).

LIWC	Gender		Age		Extraversion		Agreeableness		Conscientious.		Neuroticism		Openness	
Category	[34] *d*	our *β*	[30] *β*	our *β*	[27] *ρ*	our *β*	[27] *ρ*	our *β*	[27] *ρ*	our *β*	[27] *ρ*	our *β*	[27] *ρ*	our *β*
Total function words	-	-0.04	-	0.16	-	-0.04	-	0.02	-	0.02	-	0.03	-	0.09
Total pronouns	0.36	0.07	-	-0.02	*ns*	*ns*	0.11	*ns*	*ns*	-0.03	*ns*	0.04	-0.21	0.07
Personal pronouns	-	0.14	-	-0.08	-	*ns*	-	*ns*	-	-0.04	-	0.04	-	0.05
1st pers singular	0.17	0.13	-0.14	-0.22	*ns*	*ns*	*ns*	-0.03	*ns*	-0.06	0.12	0.05	-0.16	0.05
1st pers plural	*ns*	*ns*	-0.13	0.21	0.11	0.03	0.18	0.05	*ns*	0.05	*ns*	-0.04	-0.1	*ns*
2nd person	-0.06	0.05	-	0.04	0.16	*ns*	*ns*	0.02	*ns*	*ns*	-0.15	*ns*	-0.12	0.02
3rd pers singular	-	0.09	-	0.15	-	*ns*	-	*ns*	-	*ns*	-	0.02	-	*ns*
3rd pers plural	-	-0.05	-	0.26	-	-0.06	-	-0.04	-	*ns*	-	0.02	-	0.03
3rd pers overall	0.2	-	-	-	*ns*	-	*ns*	-	*ns*	-	*ns*	-	*ns*	-
Impersonal pronouns	-	-0.09	-	0.11	-	-0.05	-	*ns*	-	*ns*	-	0.02	-	0.08
Articles	-0.24	-0.24	-	0.28	*ns*	-0.05	*ns*	*ns*	0.09	0.02	-0.11	-0.02	0.2	0.13
Common verbs	-	0.04	-	0.02	-	-0.03	-	*ns*	-	*ns*	-	0.04	-	0.03
Auxiliary verbs	-	0.02	-	0.08	-	-0.06	-	*ns*	-	*ns*	-	0.05	-	0.07
Past tense	0.12	-0.03	-0.16	*ns*	*ns*	-0.04	0.1	0.02	*ns*	-0.02	*ns*	*ns*	-0.16	*ns*
Present tense	0.18	0.08	0.04	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	0.04	-0.16	0.03
Future tense	*ns*	-0.07	0.14	0.09	*ns*	-0.05	*ns*	*ns*	*ns*	*ns*	*ns*	0.03	*ns*	0.05
Adverbs	-	0.05	-	-0.07	-	-0.04	-	*ns*	-	*ns*	-	0.05	-	0.04
Prepositions	-0.17	-0.13	-	0.27	*ns*	-0.04	*ns*	0.03	*ns*	0.06	*ns*	*ns*	0.17	0.06
Conjunctions	-	0.03	-	0.12	-	-0.02	-	0.02	-	0.02	-	0.02	-	0.06
Negations	0.11	*ns*	-	-0.12	*ns*	-0.06	*ns*	-0.05	-0.17	-0.03	0.11	0.07	-0.13	0.02
Quantifiers	-	-0.09	-	0.24	-	-0.02	-	0.03	-	0.05	-	*ns*	-	0.05
Numbers	-0.15	-0.13	-	0.05	-0.12	-0.06	0.11	0.02	*ns*	0.02	*ns*	*ns*	-0.08	0.06
Swear words	-0.22	-0.21	-	-0.17	*ns*	*ns*	-0.21	-0.15	-0.14	-0.09	0.11	0.06	*ns*	*ns*
Social processes	-	0.08	-0.13	0.21	0.15	0.04	0.13	0.02	*ns*	*ns*	*ns*	*ns*	-0.14	*ns*
Family	0.12	0.22	-	0.28	0.09	0.03	0.19	0.03	*ns*	0.03	*ns*	*ns*	-0.17	-0.12
Friends	0.09	0.08	-	0.26	0.15	0.05	0.11	0.04	*ns*	0.02	-0.08	*ns*	*ns*	-0.04
Humans	*ns*	0.04	-	0.06	0.13	0.06	*ns*	-0.05	-0.12	*ns*	*ns*	*ns*	-0.09	*ns*
Affective processes	0.11	0.11	-	-0.05	0.09	0.07	*ns*	0.02	*ns*	*ns*	*ns*	*ns*	-0.12	-0.04
Positive emotion	*ns*	0.21	0.12	0.14	0.1	0.13	0.18	0.13	*ns*	0.1	*ns*	-0.08	-0.15	-0.07
Negative emotion	0.1	-0.12	-0.05	-0.31	*ns*	-0.07	-0.15	-0.17	-0.18	-0.13	0.16	0.15	*ns*	0.03
Anxiety	0.16	0.08	-	-0.13	*ns*	-0.04	*ns*	-0.02	*ns*	-0.02	0.17	0.06	*ns*	0.07
Anger	*ns*	-0.22	-	-0.25	*ns*	-0.05	-0.23	-0.19	-0.19	-0.12	0.13	0.11	*ns*	0.02
Sadness	0.1	0.08	-	-0.15	*ns*	-0.04	*ns*	-0.02	-0.11	-0.04	0.1	0.09	*ns*	*ns*
Cognitive processes	0.07	-0.03	0.07	0.1	*ns*	-0.05	*ns*	0.02	-0.11	*ns*	0.13	0.04	-0.09	0.1
Insight	0.09	-0.05	0.11	0.04	*ns*	-0.09	*ns*	*ns*	*ns*	-0.02	*ns*	0.05	*ns*	0.13
Causation	*ns*	-0.05	*ns*	-0.01	-0.09	-0.06	-0.11	-0.02	-0.12	*ns*	0.11	0.02	*ns*	0.08
Discrepancy	0.07	*ns*	-	0.02	*ns*	-0.05	*ns*	-0.02	-0.13	-0.03	0.13	0.07	-0.12	0.02
Tentative	*ns*	-0.12	-	0.07	-0.11	-0.08	*ns*	*ns*	-0.1	-0.03	0.12	0.06	*ns*	0.07
Certainty	0.14	*ns*	-	0.09	0.1	*ns*	*ns*	0.03	-0.1	0.04	0.13	*ns*	*ns*	0.06
Inhibition	-	0.03	-	0.09	-0.13	*ns*	*ns*	*ns*	*ns*	0.04	0.09	*ns*	*ns*	*ns*
Inclusive	*ns*	0.04	-	0.23	0.09	0.04	0.18	0.05	*ns*	0.05	*ns*	-0.02	0.11	0.06
Exclusive	*ns*	-0.05	*ns*	*ns*	*ns*	-0.07	*ns*	*ns*	-0.16	-0.03	0.1	0.05	*ns*	0.05
Perceptual Processes	0.12	*ns*	-	-0.06	0.09	-0.04	*ns*	*ns*	-0.1	-0.07	*ns*	0.03	-0.11	0.1
See	*ns*	*ns*	-	*ns*	*ns*	-0.02	0.09	*ns*	*ns*	-0.04	*ns*	*ns*	*ns*	0.04
Hear	0.1	-0.07	-	-0.1	0.12	-0.04	*ns*	*ns*	-0.12	-0.06	*ns*	0.02	-0.08	0.08
Feel	0.17	0.04	-	-0.07	*ns*	-0.02	0.1	*ns*	*ns*	-0.04	0.1	0.03	*ns*	0.05
Biological processes	*ns*	0.05	-	-0.06	0.14	0.04	0.09	-0.06	*ns*	-0.06	*ns*	0.05	-0.09	0.02
Body	-	-0.02	-	-0.14	0.1	*ns*	0.09	-0.09	*ns*	-0.09	*ns*	0.06	-0.04	0.04
Health	-	0.05	-	0.07	-	*ns*	-	*ns*	-	*ns*	-	0.06	-	*ns*
Sexual	*ns*	0.05	-	-0.14	0.17	0.1	0.08	-0.04	*ns*	-0.04	*ns*	*ns*	*ns*	*ns*
Ingestion	-	0.02	-	0.12	-	*ns*	-	-0.03	-	-0.03	-	*ns*	-	0.03
Relativity	-	-0.06	-	0.16	-	*ns*	-	0.05	-	0.08	-	-0.03	-	-0.03
Motion	0.07	*ns*	-	0.12	-	0.02	-	0.05	-	0.07	-	-0.04	-	-0.04
Space	*ns*	-0.18	-	0.21	*ns*	*ns*	0.16	*ns*	*ns*	0.02	-0.09	*ns*	-0.11	0.07
Time	*ns*	0.02	-0.19	0.08	*ns*	*ns*	0.12	0.06	0.09	0.09	*ns*	-0.03	-0.22	-0.07
Work	-0.12	-0.08	-	-0.02	-0.08	-0.05	*ns*	0.03	*ns*	0.1	*ns*	-0.03	*ns*	-0.02
Achievement	-	-0.17	-	0.16	-0.09	*ns*	*ns*	0.05	0.14	0.11	*ns*	-0.06	*ns*	-0.02
Leisure	*ns*	-0.08	-	0.03	0.08	0.06	0.15	0.04	*ns*	0.03	*ns*	-0.07	-0.17	*ns*
Home	0.15	0.19	-	0.18	*ns*	*ns*	0.19	0.03	*ns*	0.04	*ns*	-0.02	-0.2	-0.06
Money	-0.1	-0.12	-	0.24	*ns*	*ns*	-0.11	-0.04	*ns*	0.03	*ns*	*ns*	*ns*	0.03
Religion	-	-0.03	-	0.21	0.11	*ns*	*ns*	0.06	*ns*	0.04	*ns*	-0.04	*ns*	*ns*
Death	-	-0.18	-	-0.1	*ns*	-0.08	-0.13	-0.09	-0.12	-0.08	*ns*	0.08	0.15	0.09
Assent	-	0.07	-	-0.22	*ns*	0.05	*ns*	0.04	-0.09	*ns*	*ns*	-0.04	-0.11	-0.05
Nonfluencies	-	-0.03	-	0.02	-	*ns*	-	*ns*	-	*ns*	-	0.03	-	*ns*
Fillers	-	-0.02	-	-0.24	-	*ns*	-	-0.04	-	-0.08	-	0.03	-	0.04
participants (N)	*9,130*	*74,859*	*3,087*	*74,859*	*576*	*72,709*	*576*	*72,772*	*576*	*72,781*	*576*	*71,968*	*576*	*72,809*

Figure 2. Correlation values of *LIWC* categories with gender, age, and the five factor model of personality. [34] *d*: Effect size as Cohen's *d* values from Newman et al. 's recent study of gender (positive is female, *ns*= not significant at $p<.001$) [30]. *β*: Standardized linear regression coefficients adjusted for sex, writing/talking, and experimental condition from Pennebaker and Stone's study of age (*ns*= not significant at $p<.05$) [27]. *ρ*: Spearman correlations values from Yarkoni's recent study of personality (*ns*= not significant at $p<.05$). **our** *β*: Standardized multivariate regression coefficients adjusted for gender and age for this current study over Facebook (*ns*= not significant at Bonferroni-corrected $p<.001$).
doi:10.1371/journal.pone.0073791.g002

Εικόνα 9. Τιμές συσχέτισης των κατηγοριών LIWC με το φύλο, την ηλικία και το μοντέλο προσωπικότητας OCEAN

Οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν σε αυτή την έρευνα, ήταν κατηγορίες του λεξικού LIWC, ενώ τα χαρακτηριστικά προσωπικότητας χρησιμοποιήθηκαν ως εξαρτώμενες μεταβλητές. Η συχνότητα χρήσης μιας λέξης κάθε κατηγορίας υπολογίστηκε διαιρώντας τον αριθμό εμφανίσεων μιας λέξης από μια κατηγορία με τον συνολικό αριθμό λέξεων που χρησιμοποιεί ο κάθε συμμετέχων. Ο

συντελεστής της ανεξάρτητης μεταβλητής χρησίμευσε ως βάρος σε μια γραμμική συνάρτηση που συνέδεε τις ανεξάρτητες μεταβλητές με τις εξαρτημένες. Τα αποτελέσματα αυτής της εργασίας απέδειξαν ότι η προσέγγιση του ανοικτού λεξιλογίου παρέχει λεπτομερέστερες πληροφορίες από άλλα ερευνητικά μοντέλα όπου οι κατηγορίες λέξεων είναι προκαθορισμένες. Επίσης, δόθηκαν τιμές συσχέτισης μεταξύ ηλικίας, φύλου και προσωπικότητας (Εικόνα 9).

Οι Mahmud et al. (2014) [138] δημιούργησαν ένα «έξυπνο σύστημα συλλογής πληροφοριών» που παράγει ορισμένες ερωτήσεις για να λάβει τις επιθυμητές πληροφορίες, όπως για παράδειγμα ερωτήσεις σχετικά με τις εκδηλώσεις που επισκέφθηκε ο χρήστης ή σχετικά την ποιότητα κάποιων αγαθών. Ο σκοπός ήταν να επιλεγούν «οι κατάλληλοι χρήστες την κατάλληλη στιγμή» που είναι πιο πιθανό να δώσουν τις απαραίτητες πληροφορίες. Για να το καταφέρουν αυτό, το σύστημα αναλύει ροή μηνυμάτων του κοινωνικού δικτύου Twitter, επιλέγοντας τα tweets που περιέχουν τις πληροφορίες που ενδιαφέρουν τους ερευνητές και στη συνέχεια επεξεργάζεται τα tweets του χρονοδιαγράμματος (timeline) των συντακτών των επιλεγμένων tweets για να υπολογίσει τον τύπο προσωπικότητας του χρήστη. Για τον υπολογισμό των σχετικών χαρακτηριστικών για τον καθορισμό ενός τύπου προσωπικότητας σε αυτή την έρευνα οι συγγραφείς χρησιμοποίησαν την εφαρμογή LIWC-2001. Πρέπει να σημειωθεί ότι, τα retweets κατά τη διάρκεια αυτής της διαδικασίας αποκλείστηκαν. Μετά την εκτέλεση της διαδικασίας το σύστημα εμφάνιζε μια λίστα των συνιστώμενων χρηστών για να κάνει ερωτήσεις. Οι συγγραφείς υπέθεσαν ότι μόνο οι χρήστες με συγκεκριμένα χαρακτηριστικά προσωπικότητας, όπως η εξωστρέφεια και η φιλικότητα, ήταν πιο πιθανό να ανταποκριθούν.

Οι Matz et al. (2017) [139] πρότειναν την ανάλυση της προσωπικότητα των χρηστών του Facebook ως μέσο μαζικής πειθούς για την αποτελεσματικότερη ανάπτυξη της αγοράς. Στο έργο τους, οι χρήστες του Facebook έπρεπε να συμπληρώσουν ένα ερωτηματολόγιο για να καθορίσουν τα OCEAN χαρακτηριστικά της προσωπικότητας τους. Μόλις κατοχυρώνονταν τα χαρακτηριστικά ενός χρήστη, οι ερευνητές έτρεχαν διαφημίσεις προσαρμοσμένες στις προσωπικότητες των χρηστών σε μια προσπάθεια να συσχετίσουν τα χαρακτηριστικά της προσωπικότητας κάθε ατόμου με το προτιμώμενο στυλ και μορφή διαφήμισης.

Οι Z. Xu et al. (2011) [140] προσπάθησαν να αποκαλύψουν τα ενδιαφέροντα και την εξειδίκευση των χρηστών του Twitter, προτείνοντας ένα νέο πλαίσιο μοντελοποίησης θεμάτων για να το πράξουν. Κατασκεύασαν ένα σύνολο δεδομένων αποτελούμενο από

200 tweets, 200 retweets, 200 links[3] και 200 replies και χαρακτήρισαν με μη αυτόματο τρόπο καθένα από αυτά είτε ως (i) σχετικό με το θέμα, είτε ως (ii) μη σχετικό με το θέμα. Για να αποκαλύψουν τα υποκείμενα θέματα ενδιαφέροντος στα δεδομένα του Twitter, χρησιμοποίησαν μια εκτεταμένη παραλλαγή του αλγόριθμου Latent Dirichlet Allocation (LDA), ο οποίος ενσωματώνει πληροφορίες του συγγραφέα στα θεματικά μοντέλα[4] [141].

[3] Με την έννοια ότι τα αντίστοιχα 200 tweets περιλάμβαναν διευθύνσεις URL.

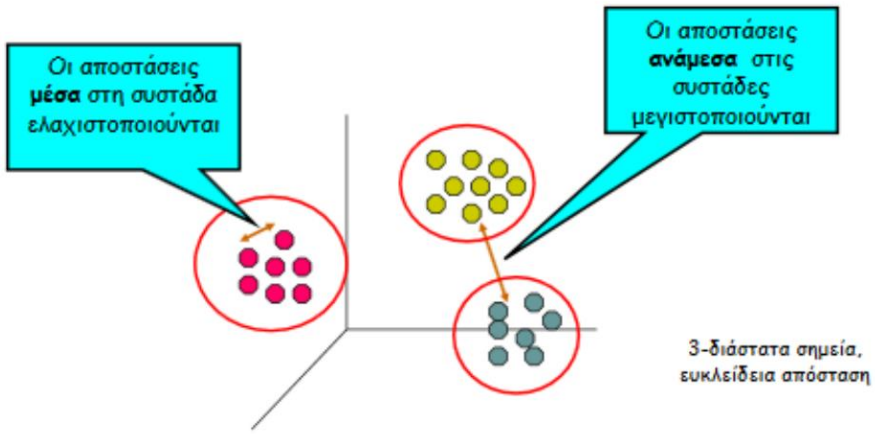
[4] Η μοντελοποίηση θεμάτων είναι μια κοινώς χρησιμοποιούμενη τεχνική Natural Language Processing και Machine Learning, η οποία χρησιμοποιεί στατιστική μοντελοποίηση για να ανακαλύψει θέματα που συμβαίνουν σε ένα σώμα εγγράφων. Χρησιμοποιείται συνήθως για την ανακάλυψη κρυφών σημασιολογικών δομών σε ένα κείμενο, όπως είναι το αντικείμενο των τεχνικών όπως η ενσωμάτωση λέξεων.

4.5. Μέθοδοι κατηγοριοποίησης

4.5.1. Συσταδοποίηση (Clustering)

Η εξόρυξη προφίλ χρηστών είναι μια διαδικασία που έχει ως βασικό στόχο την συσταδοποίηση/ομαδοποίηση/κατηγοριοποίηση (clustering) των χρηστών των μέσων κοινωνικής δικτύωσης.

Η συσταδοποίηση (clustering) είναι μια μη επιτηρούμενη διαδικασία κατηγοριοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες (clusters). Στόχος της συσταδοποίησης είναι η παραγωγή ενός συνόλου από ομάδες με υψηλή ομοιότητα εντός των ομάδων (intra-cluster similarity), ενώ παράλληλα θα πρέπει να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων (inter-cluster similarity) (Εικόνα 10).

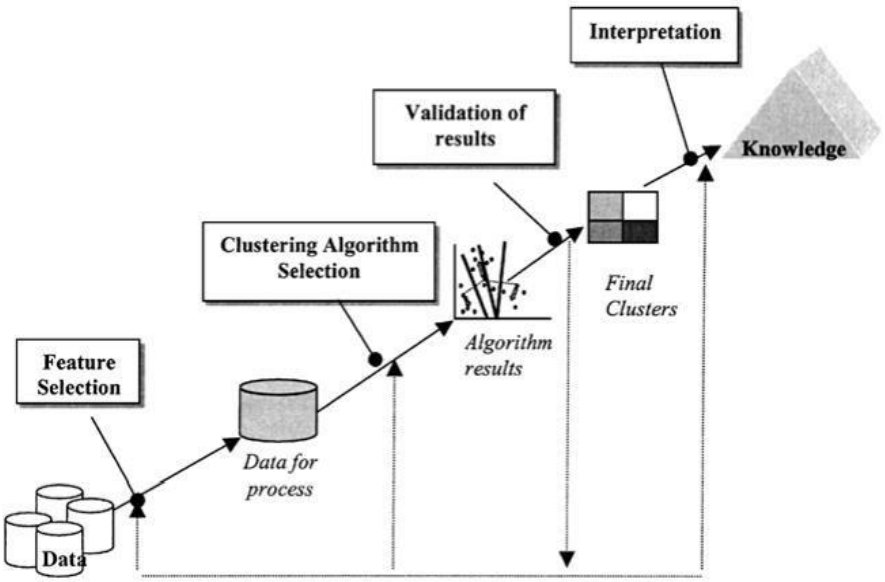


Εικόνα 10. Αναπαράσταση διαδικασίας συσταδοποίησης

(Πηγή:http://archive.eclass.uth.gr/eclass/modules/document/file.php/DIB263/ΔΙΑΛΕΞΕΙΣ/PR%20-%202.pdf)

4.5.1.1. Βήματα συσταδοποίησης

Η συσταδοποίηση είναι μία από τις πιο χρήσιμες εργασίες στη διαδικασία εξόρυξης δεδομένων για την ανακάλυψη ομάδων και τον εντοπισμό χρήσιμων και ενδιαφερόντων μοτίβων στα υποκείμενα δεδομένα. Τα βασικά βήματα της διαδικασίας συσταδοποίησης παρουσιάζονται στην Εικόνα 11 και είναι τα εξής (Μαρία Χαλκίδη, 2001)[142]:



Εικόνα 11. Βήματα της διαδικασίας συσταδοποίησης

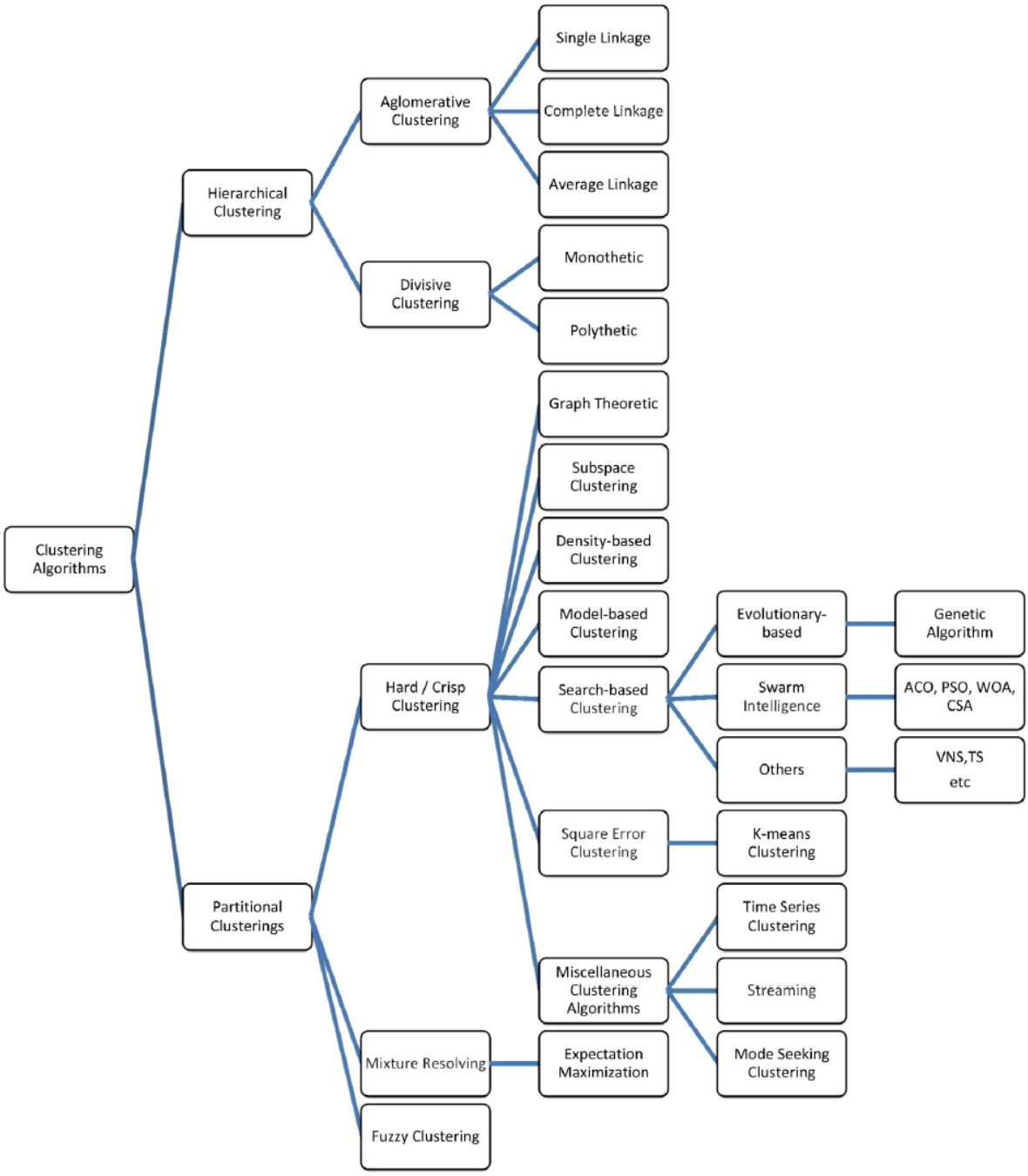
- **Επιλογή χαρακτηριστικών γνωρισμάτων:** Στόχος είναι να επιλεγούν τα κατάλληλα χαρακτηριστικά γνωρίσματα με βάση τα οποία πρόκειται να εκτελεσθεί με επιτυχία η συσταδοποίηση. Η διαδικασία της προεπεξεργασία των δεδομένων κρίνεται απαραίτητη σε αυτό το βήμα, ώστε τα δεδομένα να αναπαρίστανται με τη μορφή διανυσμάτων.
- **Αλγόριθμος συσταδοποίησης**: Σε αυτό το βήμα γίνεται η επιλογή του κατάλληλου αλγορίθμου συσταδοποίησης. Η επιλογή αυτή εξαρτάται από τα μορφή των δεδομένων που πρόκειται να συσταδοποιηθούν και τις ανάγκες τις εκάστοτε εφαρμογής. Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης είναι αυτά που κυρίως χαρακτηρίζουν έναν αλγόριθμο συσταδοποίησης.
    1. Με το μέτρο γειτνίασης, υπολογίζεται η ομοιότητα μεταξύ των στοιχείων.
    2. Το κριτήριο συσταδοποίησης, εκφράζεται συνήθως μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων.
- **Επικύρωση των αποτελεσμάτων:** Η διαδικασία αξιολόγησης των αποτελεσμάτων ενός αλγορίθμου συσταδοποίησης είναι γνωστή με τον όρο εγκυρότητα συσταδοποίησης (cluster validity). Σε γενικές γραμμές, υπάρχουν τρεις προσεγγίσεις για τη διερεύνηση της εγκυρότητας της συσταδοποίησης :
    - ✓ ***Εξωτερικά κριτήρια***: Σε αυτήν την προσέγγιση η βασική ιδέα είναι να ελεγχθεί εάν τα σημεία του συνόλου δεδομένων είναι τυχαία δομημένα ή όχι. Τέτοια εξωτερικά κριτήρια είναι ο δείκτης Rand, ο συντελεστής Jaccard, η εντροπία και η καθαρότητα.

- ✓ ***Τα εσωτερικά κριτήρια*** αξιολογούν το αποτέλεσμα σε σχέση με τις πληροφορίες που είναι εγγενείς μόνο στα δεδομένα. Ο δείκτης Silhouette, ο δείκτης Davies-Bouldin (DB), ο δείκτης Calinski-Harabasz (CH) και ο δείκτης Dunn είναι τα πιο διάσημα κριτήρια σε αυτή την κατηγορία (Eréndira Rendón, 2011)[143].
- ✓ ***Τα σχετικά κριτήρια*** αξιολογούν την ποιότητα μιας συστάδας συγκρίνοντας την με άλλα σχήματα συσταδοποίησης, που προκύπτουν από τον ίδιο αλγόριθμο αλλά με διαφορετικές τιμές παραμέτρων.

- **Ερμηνεία των αποτελεσμάτων**: Σε πολλές περιπτώσεις, οι ειδικοί στον τομέα της εφαρμογής αυτών των αλγορίθμων, πρέπει να ενοποιήσουν τα αποτελέσματα της συσταδοποίησης με άλλα πειραματικά στοιχεία και αναλύσεις προκειμένου να καταλήξουν στα σωστά συμπεράσματα.

4.5.1.2. Κατηγοριες και αλγοριθμοι συσταδοποιησης

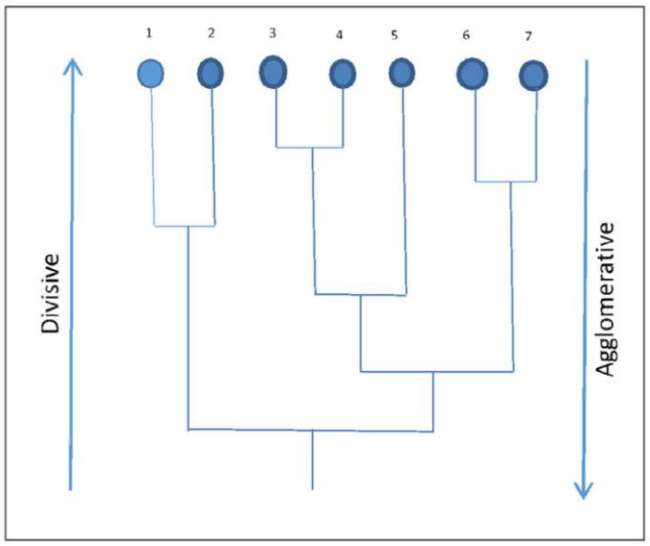
Μία ταξινόμηση των αλγορίθμων συσταδοποίησης εμφανίζονται στην παρακάτω Εικόνα 12.



Εικόνα 12. Ταξινόμηση αλγορίθμων συσταδοποίησης (A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects Absalom E. Ezugwu a,∗, Abiodun M. Ikotun a, Olaide O. Oyelade a, Laith Abualigah b,c,∗∗, Jeffery O. Agushaka a, Christopher I. Eke d, Andronicus A. Akinyelu e)

Οι αλγόριθμοι συσταδοποίησης μπορούν να ταξινομηθούν ευρέως στις εξής κατηγορίες:

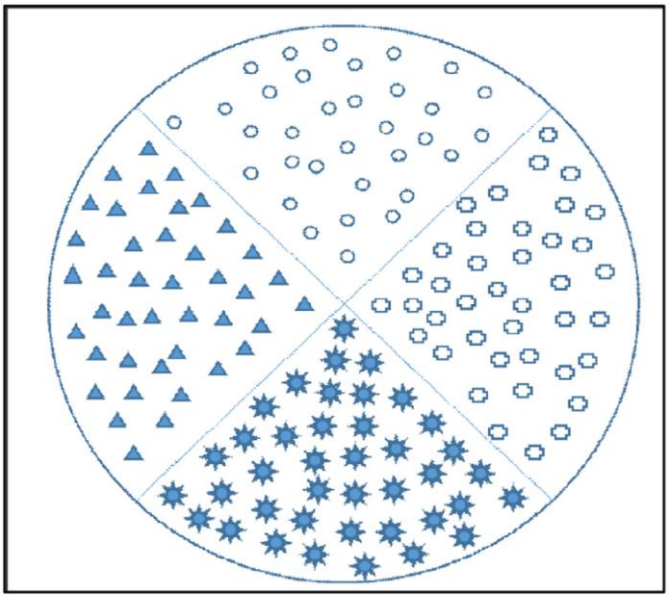
- ✓ ***Ιεραρχική συσταδοποίηση (hierarchical clustering):*** Κατά την διάρκεια αυτής παράγονται νέες συστάδες, είτε από την συγχώνευση μικρότερων συστάδων είτε από την διαίρεση μεγαλύτερων. Το αποτέλεσμα του αλγορίθμου είναι ένα δέντρο συστάδων, που ονομάζεται δενδροδιάγραμμα, το οποίο δείχνει πώς σχετίζονται οι συστάδες (Εικόνα 13). «Κόβοντας» το δενδρόγραμμα στο επιθυμητό επίπεδο, λαμβάνεται μια συσταδοποίηση των δεδομένων σε ασύνδετες μεταξύ τους ομάδες.



Εικόνα 13. Αναπαράσταση δενδροδιαγράμματος για την ιεραρχική συσταδοποίηση των αντικειμενικών δεδομένων 1,2,3,4,5,6,7…

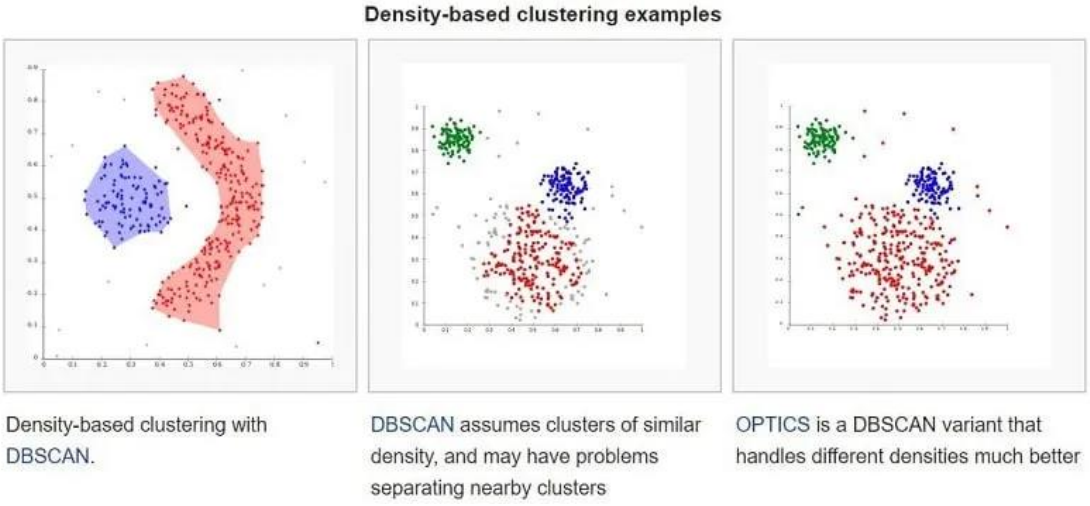
- ✓ ***Διαχωριστική συσταδοποίηση (partinional clustering):*** Ο αλγόριθμος δημιουργεί μόνο ένα σύνολο συστάδων. Οι διαχωριστικοί αλγόριθμοι, δηλαδή, διαιρούν το σύνολο των δεδομένων σε μη επικαλυπτόμενα υποσύνολα (συστάδες) τέτοια ώστε κάθε αντικείμενο να ανήκει ακριβώς σε ένα υποσύνολο. Ο συνολικός αριθμός των επιθυμητών παραγόμενων συστάδων αποτελεί είσοδο για ένα διαχωριστικό αλγόριθμο σε αντίθεση με οποιονδήποτε ιεραρχικό. Σε αυτή την κατηγορία ανήκουν οι αλγόριθμοι K-Means, K-medoids or PAM (Partitioning Around Medoids) και CLARA algorithm (Clustering Large Applications), που είναι μια επέκταση του PAM για μεγαλύτερα σύνολα δεδομένων.

  Στην Εικόνα 14 δίνεται η αναπαράσταση του μοτίβου ομαδοποίησης της μέθοδος διαμερισματικής συσταδοποίησης.



Εικόνα 14. Διαμερισματική συσταδοποίηση

- ✓ ***Συσταδοποίηση με βάση την πυκνότητα (Density-based clustering):*** Η βασική ιδέα αυτού του αλγορίθμου είναι η ομαδοποίηση πολύ πυκνών περιοχών αποτελούμενη από γειτονικά αντικείμενα ενός συνόλου δεδομένων σε συστάδες με βάση τις συνθήκες πύκνωσης. Ένας ευρέως γνωστός αλγόριθμος αυτής της κατηγορίας είναι η Χωρική Συσταδοποίηση βάσει πυκνότητας (DBSCAN - Density Based Spatial Clustering) (Εικόνα 15).



Εικόνα 15. Παραδείγματα Χωρικής Συσταδοποίησης βάσει πυκνότητας

- ✓ ***Συσταδοποίηση βασισμένη σε πλέγμα (grid-based clustering):*** Η συσταδοποίηση αυτή προτείνεται κυρίως για την εξόρυξη χωρικών δεδομένων. Το κύριο χαρακτηριστικό τους είναι ότι κβαντοποιεί τις περιοχές αντικειμένων σε έναν πεπερασμένο αριθμό κελιών που σχηματίζουν μια δομή πλέγματος στην

οποία υλοποιούνται όλες οι λειτουργίες για συσταδοποίηση. Το πλεονέκτημα της μεθόδου αυτής είναι ο γρήγορος χρόνος επεξεργασίας της - ο οποίος είναι γενικά ανεξάρτητος από τον αριθμό των δεδομένων - που εξακολουθεί να εξαρτάται μόνο από τα πολλαπλά κελιά σε κάθε διάσταση στον κβαντικό χώρο. Στην κατηγορία αυτή ανήκουν ο αλγόριθμος STING, ο οποίος διερευνάς στατιστικά δεδομένα που είναι αποθηκευμένα στα κύτταρα του πλέγματος, ο WaveCluster, ο οποίος ομαδοποιεί αντικείμενα χρησιμοποιώντας μια προσέγγιση μετασχηματισμού κυματιδίων, και το CLIQUE, το οποίο ορίζει μια προσέγγιση βάσει πλέγματος και πυκνότητας για ομαδοποίηση σε χώρο δεδομένων υψηλών διαστάσεων.

- ✓ ***Ασαφής συσταδοποίηση (fuzzy clustering):*** Η μέθοδος αυτή χρησιμοποιεί ασαφείς τεχνικές για τη συσταδοποίηση δεδομένων, θεωρώντας ότι ένα αντικείμενο μπορεί να ταξινομηθεί σε περισσότερα από μία συστάδα. Αυτός ο τύπος αλγορίθμων οδηγεί σε σχήματα συσταδοποίησης που είναι συμβατά με την εμπειρία της καθημερινής ζωής καθώς χειρίζονται την αβεβαιότητα των δεδομένων της πραγματικότητας. Ο πιο σημαντικός αλγόριθμος ασαφούς συσταδοποίησης ο Fuzzy C-Means.
- ✓ ***Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (Kohonen net clustering):*** Η συσταδοποίηση αυτή βασίζεται στις έννοιες των νευρικών δικτύων. Τα νευρωνικά δίκτυα Kohonen παρέχουν έναν τρόπο κατηγοριοποίησης των δεδομένων μέσω αυτό-οργανωμένων (self-organizing) δικτύων τεχνητών νευρώνων. Δύο βασικές έννοιες που κυριαρχούν στα δίκτυα Kohonen η ανταγωνιστική μάθηση και η αυτό-οργάνωση. Ο όρος ανταγωνιστική μάθηση αφορά στην εύρεση ενός νευρώνα ο οποίος προσεγγίζει περισσότερο το πρότυπο εισόδου. Το δίκτυο στη συνέχεια τροποποιεί αυτό τον νευρώνα και τους γειτονικούς του (ανταγωνιστική μάθηση με αυτό-οργάνωση) έτσι ώστε να μοιάζουν περισσότερο με το πρότυπο.

4.5.2.Συσταδοποίηση Δύο Βημάτων (TwoStep Cluster Analysis)

Η συσταδοποίηση δύο βημάτων είναι ένα διερευνητικό εργαλείο που έχει σχεδιαστεί για να αποκαλύπτει φυσικές ομαδοποιήσεις (ή συστάδες) μέσα σε ένα σύνολο δεδομένων που διαφορετικά δεν θα ήταν εμφανείς. Ο αλγόριθμος που χρησιμοποιείται από αυτή τη

διαδικασία έχει αρκετά επιθυμητά χαρακτηριστικά που τον διαφοροποιούν από τις παραδοσιακές τεχνικές ομαδοποίησης, όπως είναι:

1. Η δυνατότητα δημιουργίας συστάδων με βάση τόσο κατηγορικές όσο και συνεχείς μεταβλητές.
2. Αυτόματη επιλογή του αριθμού των συστάδων.
3. Η δυνατότητα αποτελεσματικής ανάλυσης μεγάλων αρχείων δεδομένων.

Για το χειρισμό των κατηγορικών και συνεχών μεταβλητών, η συσταδοποίηση δύο βημάτων χρησιμοποιεί μια μέτρηση απόστασης πιθανότητας που υποθέτει ότι οι μεταβλητές στο μοντέλο συστάδας είναι ανεξάρτητες. Επιπλέον, κάθε συνεχής μεταβλητή θεωρείται ότι έχει μια κανονική (Γκαουσιανή - Gaussian) κατανομή και κάθε κατηγορική μεταβλητή θεωρείται ότι έχει μια πολυωνυμική κατανομή. Οι εμπειρικές εσωτερικές δοκιμές έχουν δείξει ότι η διαδικασία είναι αρκετά ισχυρή σε παραβιάσεις τόσο της υπόθεσης της ανεξαρτησίας όσο και των υποθέσεων κατανομής, αλλά θα πρέπει σε κάθε περίπτωση εφαρμογής της να γίνεται έλεγχος για το αν πληρούνται αυτές οι παραδοχές. να προσπαθήσετε να γνωρίζετε πόσο καλά πληρούνται αυτές οι υποθέσεις.

Τα δύο βήματα του αλγορίθμου της συσταδοποίησης δύο βημάτων μπορούν να συνοψιστούν ως εξής:

Βήμα 1. Η διαδικασία ξεκινά με την κατασκευή ενός Δέντρου Χαρακτηριστικών Συστάδων (Cluster Features Tree). Το δέντρο ξεκινά τοποθετώντας την πρώτη περίπτωση στη ρίζα του δέντρου σε έναν κόμβο φύλλων που περιέχει πληροφορίες των μεταβλητών σχετικά με αυτή την περίπτωση. Κάθε διαδοχική περίπτωση προστίθεται στη συνέχεια σε έναν υπάρχοντα κόμβο ή σχηματίζει έναν νέο κόμβο, με βάση την ομοιότητά του με τους υπάρχοντες κόμβους και χρησιμοποιώντας την μέτρηση της απόστασης ως κριτήριο ομοιότητας. Ένας κόμβος που περιέχει πολλές περιπτώσεις περιέχει μια σύνοψη των πληροφοριών των μεταβλητών σχετικά με αυτές τις περιπτώσεις. Έτσι, το Δέντρο Χαρακτηριστικών Συστάδων παρέχει μια περίληψη του αρχείου δεδομένων.

Βήμα 2. Οι κόμβοι φύλλων του Δέντρου Χαρακτηριστικών Συστάδων ομαδοποιούνται στη συνέχεια χρησιμοποιώντας έναν αλγόριθμο συσσωματούμενης συσταδοποίησης. Η συσσωματούμενη συσταδοποίηση μπορεί να χρησιμοποιηθεί για την παραγωγή μιας σειράς λύσεων. Για να προσδιοριστεί ποιος αριθμός συστάδων είναι «καλύτερος», κάθε μία από αυτές τις λύσεις συστάδων αξιολογείται χρησιμοποιώντας το Κριτήριο Schwarz's Bayesian (Schwarz's Bayesian Criterion - BIC) ή το Κριτήριο Πληροφοριών του Akaike (Akaike Information Criterion - AIC) ως κριτήριο συσταδοποίησης. [144]

5. Μεθοδολογία

Για την ανάλυση συναισθήματος, βασισμένοι στα tweets των χρηστών του κοινωνικού μέσου Twitter επιχειρήσαμε να εξάγουμε το συναίσθημα που εκφράζεται μέσα από αυτά με τη βοήθεια της γλώσσας προγραμματισμού Python και την χρήση κατάλληλων βιβλιοθηκών επεξεργασίας φυσικής γλώσσας. Πριν την ανάλυση προηγήθηκε η εκκαθάριση και κανονικοποίηση των δεδομένων, δύο προαπαιτούμενα στάδια, ώστε το κείμενο του κάθε tweet να μπορεί να «διαβαστεί» από τον αλγόριθμο ανάλυσης συναισθήματος.

Στο δεύτερο κομμάτι της διπλωματικής, αξιοποιώντας τα δεδομένα της ανάλυσης που είχε προηγηθεί στα tweets από την εφαρμογή LIWC2007, εφαρμόσαμε τον αλγόριθμο Two-Step Cluster μέσω της προγράμματος στατιστικής ανάλυσης SPSS, προκειμένου να εξάγουμε συμπεράσματα για το προφίλ των χρηστών, κατατάσσοντάς τους σε κάποια από τις κατηγορίες χρηστών σύμφωνα με τα 7 επίπεδα της «Σκάλας των κοινωνικών τεχνολογικών συμπεριφορών» (βάση της διαδικτυακής έρευνας της Forrester Research, 2010) και το πρότυπο OCEAN.

5.1. Περιγραφή δεδομένων

Το σύνολο των δεδομένων που χρησιμοποιήθηκαν στην παρούσα διπλωματική, αφορούν τον κλάδο της αυτοκινητοβιομηχανίας και συλλέχθηκαν από το κοινωνικό μέσο Twitter κατά την περίοδο διεξαγωγής του Super Bowl τον Φεβρουάριο του 2014 στο στάδιο MetLife του αθλητικού συγκροτήματος Meadowlands της κομητείας Bergen του New Jersey. Το Super Bowl είναι ο σημαντικότερος και δημοφιλέστερος ετήσιος αγώνας του πρωταθλήματος του αμερικανικού ποδοσφαίρου και είναι το πρώτο σε τηλεθέαση γεγονός κάθε χρόνο στις ΗΠΑ και δεύτερο παγκοσμίως σε ετήσια αθλητικά γεγονότα μετά τον τελικό του ΟΥΕΦΑ Τσάμπιονς Λιγκ.

Λόγω της μεγάλης δημοσιότητάς του το Super Bowl αποτελεί για τις εταιρίες σημαντικό βήμα προβολής νέων προϊόντων με σκοπό την προσέλκυση νέων αγοραστών και την αύξηση των πωλήσεών τους. Το 2014, μια έρευνα της εταιρείας τεχνολογίας BrandAds σε 37440 Αμερικανούς καταναλωτές διαπίστωσε ότι η μέση διαφήμιση του Super Bowl αύξησε την πιθανότητα των θεατών να αγοράσουν το προϊόν κατά 6.6 %. Ο αριθμός ήταν πολύ υψηλότερος για τις πιο δημοφιλείς θέσεις, με τη Hyundai να σημειώνει άνοδο 39,5% και την Budweiser να σημειώνει αύξηση 37,8%. Μόνο το 16% των επωνυμιών δημιούργησαν αρνητικό ενδιαφέρον με τις

διαφημίσεις τους. Ομοίως, μια ανάλυση του 2012 της προβολής διαφημίσεων του Super Bowl από την Kantar Media διαπίστωσε ότι, συνολικά, οι διαφημίσεις δημιούργησαν δημοσιότητα αξίας 11 εκατομμυρίων δολαρίων για τους διαφημιζόμενους, με τις 10 κορυφαίες θέσεις να αντιπροσωπεύουν 8.6 εκατομμύρια δολάρια του συνόλου.[145], [146]

Κάθε χρόνο, ένα μεγάλο μέρος του διαθέσιμου διαφημιστικού χρόνου αγοράζετε από εταιρείες του χώρου της αυτοκινητοβιομηχανίας, όπως η Audi, η Chrysler, η KIA, η Volkswagen κ.α. είτε για την παρουσίαση νέων μοντέλων, είτε/και για την παρουσίαση νέων πρωτοβουλιών και στρατηγικών προς όφελος των καταναλωτών (π.χ. δημοσίευση νέων μειωμένων τιμών αγοράς, επενδύσεις σε φιλικά προς το περιβάλλον καύσιμα, κ.α).

Τα tweets που συλλέχθηκαν αφορούσαν τις εξής αυτοκινητοβιομηχανίες:

1. Audi
2. Chevrolet
3. Chrysler
4. KIA
5. Volkswagen

Μετά την συλλογή των δεδομένων ακολούθησε η ανάλυσή τους με την εφαρμογή LIWC2007, κατά την διάρκεια της οποίας, για κάθε λέξη του κειμένου του κάθε tweet (target words) έγινε αναζήτηση και αντιστοίχιση σε μία λέξη λεξικού (dictionary words), αυξάνονταν την βαθμολογία της κάθε μιας από τις 64 μεταβλητές εξόδου. Τα αποτελέσματα της ανάλυσης αποθηκεύτηκαν στη συνέχεια σε αρχεία csv (comma-separated values). Για την περαιτέρω επεξεργασία των δεδομένων έγινε μετατροπή των αρχείων csv σε αρχεία excel, ενώ ταυτόχρονα μετασχηματίστηκαν σε μορφή κατάλληλη για την ανάλυση που ακολούθησε.

5.2. Εργαλεία υλοποίησης της ανάλυσης

5.2.1. Python

Η Python είναι μια απλή αλλά ταυτόχρονα ισχυρή γλώσσα προγραμματισμού με λειτουργίες που ξεχωρίζουν στο αντικείμενο της επεξεργασίας γλωσσικών δεδομένων, ενώ χαρακτηρίζεται από την ευκολία της εκμάθησής της. Η σύνταξη και η σημασιολογία της είναι διαφανείς, ενώ παρουσιάζει καλή λειτουργικότητα στον χειρισμό συμβολοσειρών. Επιπλέον είναι εξαιρετικά ευανάγνωστη, καθιστώντας αρκετά εύκολο το να μαντέψει κάποιος τι κάνει αυτό το πρόγραμμα ακόμα κι αν δεν έχετε γράψει ποτέ ένα πρόγραμμα πριν.

Η Python χρησιμοποιείται σε μεγάλο βαθμό στη βιομηχανία, την επιστημονική έρευνα και την εκπαίδευση σε όλο τον κόσμο, διευκολύνοντας την παραγωγικότητα, την ποιότητα και την συντηρησιμότητα του λογισμικού[5].

Ένα απλό πρόγραμμα γραμμένο σε Python, το οποίο «διαβάζει» το αρχείο file.txt και εκτυπώνει στη συνέχεια όλες τις λέξεις που έχουν κατάληξη «ing» είναι το παρακάτω:

```
>>> for line in open("file.txt"):
...   for word in line.split():
...       if word.endswith('ing'):
... print word
```

Αυτό το πρόγραμμα απεικονίζει μερικά από τα κύρια χαρακτηριστικά της Python. Πρώτον, η εσοχή σε κάθε επόμενη γραμμή του κώδικα χρησιμοποιείται για την ένθεση (φώλιασμα) γραμμών κώδικα, έτσι ώστε η γραμμή που αρχίζει με το **if** εμπίπτει στο πεδίο εφαρμογής της προηγούμενης γραμμής που αρχίζει με **for**. Αυτό εξασφαλίζει ότι ο έλεγχος για την ύπαρξη της κατάληξης «ing» εκτελείται για κάθε λέξη. Δεύτερον, η Python είναι αντικειμενοστραφής γλώσσα. Κάθε μεταβλητή είναι μια οντότητα που έχει ορισμένα καθορισμένα χαρακτηριστικά και μεθόδους. Για παράδειγμα, η τιμή της μεταβλητής **line** είναι κάτι περισσότερο από μια ακολουθία χαρακτήρων. Είναι ένα αντικείμενο συμβολοσειράς (string) που έχει μια «μέθοδο» (ή λειτουργία) που ονομάζεται **split()** που μπορούμε να χρησιμοποιήσουμε για να σπάσουμε μια γραμμή στις λέξεις της. Για να εφαρμόσουμε μια μέθοδο σε ένα αντικείμενο, γράφουμε το όνομα

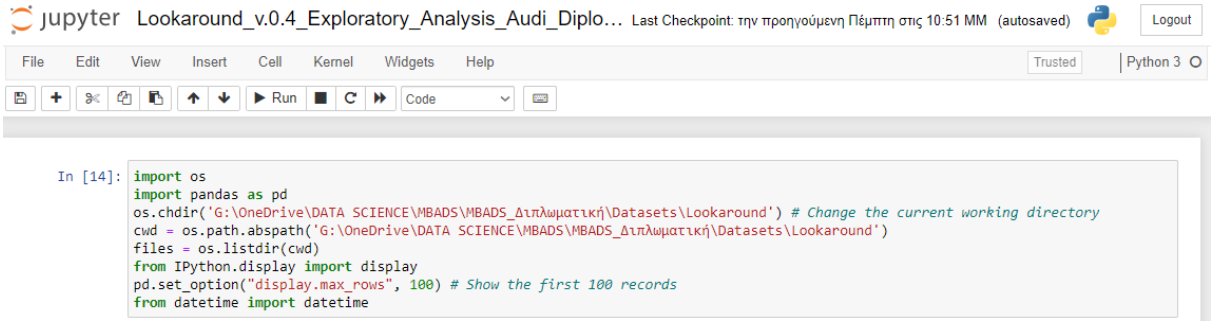
[5] Συντηρησιμότητα είναι η ευκολία με την οποία ένας κώδικας μπορεί να καταστεί άμεσα λειτουργικός, μετά από αντικατάσταση ελαττωματικών τμημάτων του, χωρίς να χρειάζεται μετατροπή των υπολοίπων τμημάτων του κώδικα που λειτουργούν χωρίς πρόβλημα.

του αντικειμένου, ακολουθούμενο από μια τελεία, ακολουθούμενο από το όνομα της μεθόδου, δηλαδή, **line.split()**. Τρίτον, οι μέθοδοι έχουν *επιχειρήματα* που εκφράζονται μέσα σε παρενθέσεις. Στο παράδειγμα, το **word.endswith('ing')** έχει το επιχείρημα 'ing' για να δείξει ότι θέλουμε λέξεις που τελειώνουν με «ing» και όχι κάτι άλλο.

Ως ερμηνευτική γλώσσα, η Python διευκολύνει τη διαδραστική εξερεύνηση. Ως αντικειμενοστρεφής γλώσσα, επιτρέπει την ενθυλάκωση δεδομένων και μεθόδων και την εύκολη επαναχρησιμοποίησή τους. Ως δυναμική γλώσσα, επιτρέπει την προσθήκη χαρακτηριστικών σε αντικείμενα και την δυναμική προσθήκη νέων μεταβλητών, διευκολύνοντας την ταχεία ανάπτυξη του κώδικα. Με την εγκατάσταση της Python, εγκαθίσταται και μια εκτεταμένη βασική βιβλιοθήκη που περιέχει στοιχεία (components) για γραφικό προγραμματισμό, αριθμητική επεξεργασία και συνδεσιμότητα ιστού[107] . Ένα χαρακτηριστικό της ισχύος της Python είναι πως έχουν αναπτυχθεί πολυάριθμες βιβλιοθήκες για κάθε είδους προγραμματιστική εργασία. Έτσι ο προγραμματιστής της Python για να κάνει μια συγκεκριμένη εργασία επιλέγει και εγκαθιστά και την αντίστοιχη βιβλιοθήκη (για την έκδοση της γλώσσας με την οποία εργάζεται).

Άλλες γλώσσες προγραμματισμού έχουν επίσης ισχυρές βιβλιοθήκες, αλλά η διαφορά είναι ότι οι βιβλιοθήκες της Python τείνουν να είναι πολύ εύκολο να εγκατασταθούν, να εισαχθούν και να αναπτυχθούν. Επιπλέον, η μεγάλη κοινότητα προγραμματιστών ανοιχτού κώδικα της Python συντηρεί και αναπτύσσει διαρκώς χρήσιμες βιβλιοθήκες, γεγονός που καθιστά την Python ιδιαίτερα δημοφιλής στον τομέα της επιστήμης των δεδομένων, της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας[147].

Για την παρούσα διπλωματική χρησιμοποιήθηκε η έκδοση 3.8 της Python, ενώ για την επεξεργασία των δεδομένων και την ανάπτυξη του κώδικα της ανάλυσης συναισθήματος χρησιμοποιήθηκε το Jupyter Notebook, που είναι ένα διαδραστικό περιβάλλον ανάπτυξης κώδικα (Interactice Development Environments – IDE) που χαρακτηρίζεται από ευελιξία και ευκολία στη χρήση του (Εικόνα 16).



Εικόνα 16

Το Jupyter Notebook αποτελεί κομμάτι της διανομής Python Anaconda η οποία είναι μια ελεύθερη και ανοιχτού κώδικα διανομή για την επιστημονική υπολογιστική (scientific computing) και την επιστήμη των δεδομένων. Η Python Anaconda διαθέτει μια σειρά από πολύ χρήσιμες βιβλιοθήκες και επιτρέπει την εύκολη διαχείριση/εγκατάσταση/επεγκατάσταση επιπλέον εφαρμογών και βιβλιοθηκών. [148]

5.2.1.1. Βιβλιοθήκες της Python

Για την επεξεργασία της αρχικής βάσης δεδομένων και την διεξαγωγή της ανάλυσης συναισθήματος εγκαταστάθηκαν και χρησιμοποιήθηκαν ένα σύνολο από βιβλιοθήκες της Python, οι οποίες περιγράφονται στη συνέχεια.

- **NumPy:** Είναι το θεμελιώδες πακέτο για επιστημονικούς υπολογισμούς στην Python. Πρόκειται για μια βιβλιοθήκη που παρέχει την δυνατότητα δημιουργίας πολυδιάστατων πινάκων, μεγάλη ποικιλία ρουτίνων για γρήγορες λειτουργίες σε πίνακες, συμπεριλαμβανομένων μαθηματικών και λογικών πράξεων, βασική γραμμική άλγεβρα, βασικές στατιστικές πράξεις κ.α. Αυτός είναι ένας από τους λόγους που το NLTK, το scikit-learn, το pandas και άλλες βιβλιοθήκες χρησιμοποιούν το NumPy ως βάση για να υλοποιήσουν μερικούς από τους αλγόριθμους.Επιπλέον παρέχει μερικές από τις εξαιρετικά βελτιστοποιημένες δομές δεδομένων, τα n-darrays.[149]

- **pandas:** Είναι μια βιβλιοθήκη ανοιχτού κώδικα για ανάλυση και χειρισμό δεδομένων, που βασίζεται στην NumPy. Δημιουργήθηκε κυρίως για εργασίες με σχεσιακά ή επισημασμένα δεδομένα, ώστε αυτές να εκτελούνται τόσο εύκολα όσο και διαισθητικά. Παρέχει διάφορες δομές δεδομένων και λειτουργίες για το χειρισμό αριθμητικών δεδομένων και χρονοσειρών. Περιέχει/ορίζει ειδικές μονοδιάστατες δομές που ονομάζονται series και δισδιάστατες δομές δεδομένων που ονομάζονται Dataframes. Τα Dataframes μπορούν να αποθηκεύσουν δεδομένα διαφορετικών τύπων (χαρακτήρες, ακέραιους, δεκαδικούς) σε στήλες, παρόμοια με τα υπολογιστικά φύλλα Excel. Υποστηρίζει την αυτόματη εισαγωγή και μετατροπή δεδομένων από διαφορετικά μοντέλα μορφοποίησης (xls, html, SQL, json) σε Dataframes για περεταίρω επεξεργασία, καθώς και την επανεξαγωγή τους μετά το πέρας της επεξεργασίας. Μερικά από τα βασικά χαρακτηριστικά της είναι:
    - ✓ Γρήγορη και αποτελεσματική για το χειρισμό και την ανάλυση δεδομένων.

- Εύκολος χειρισμός των δεδομένων που λείπουν (που αντιπροσωπεύονται ως NaN) με κινητή υποδιαστολή καθώς και σε δεδομένα μη κινητής υποδιαστολής
- ✓ Μέγεθος μεταβλητότητας: οι στήλες μπορούν να εισαχθούν και να διαγραφούν από το DataFrame και από αντικείμενα υψηλότερων διαστάσεων
- ✓ Συγχώνευση και σύνδεση συνόλου δεδομένων.
- ✓ Ευέλικτη αναδιαμόρφωση και περιστροφή συνόλων δεδομένων
- ✓ Ισχυρή ομαδοποίηση κατά λειτουργικότητα για την εκτέλεση λειτουργιών διαχωρισμού-εφαρμογής-συνδυασμού σε σύνολα δεδομένων. [150]

- **scikit-learn:** Είναι μια ανοιχτού κώδικα βιβλιοθήκη μηχανική μάθησης που υποστηρίζει αλγόριθμους τόσο επιβλεπόμενης όσο και μη επιβλεπόμενη μάθησης, όπως για παράδειγμα για ταξινόμηση, παλινδρόμηση, συσταδοποίηση κλπ., καθώς και εργαλεία για προεπεξεργασία δεδομένων κειμένου, εξαγωγή χαρακτηριστικών και κανονικοποίηση. Έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy. [151]

- **NLTK:** Είναι μια ανοιχτού κώδικα βιβλιοθήκη για τη δημιουργία προγραμμάτων που επεξεργάζονται δεδομένα ανθρώπινης γλώσσας. Παρέχει εύχρηστες διεπαφές σε περισσότερα από 50 συλλογικά έργα και λεξιλογικούς πόρους όπως το WordNet, μαζί με μια σουίτα βιβλιοθηκών επεξεργασίας κειμένου για ταξινόμηση (classification), διακριτοποίηση (tokenization), αποκοπή καταλήξεων (stemming), τοποθέτηση ετικετών (tagging), συντακική/γραμματική ανάλυση (parsing) και σημασιολογικό συλλογισμό (semantic reasoning). [152]

- **re (RegEx – Regular Expression):** Μία κανονική έκφραση (regular expression), είναι μια ακολουθία χαρακτήρων που σχηματίζει ένα μοτίβο αναζήτησης (matching pattern) και χρησιμοποιείται για την ευέλικτη αναζήτηση και «ταίριασμα» (matching) κειμένου σύμφωνα με το αυτό. Με την βιβλιοθήκη re ελέγχεται εάν μια συμβολοσειρά (πχ. κείμενο) περιέχει το καθορισμένο μοτίβο αναζήτησης, ενώ δίνεται και η δυνατότητα εκτέλεσης ρουτινών (ανάκτηση κειμένου, μετατροπή, αντικατάσταση κειμένου κ.λ.π) με βάση τα μοτίβα αυτά. Τα μοτίβα αναζήτησης καθορίζονται με την βοήθεια των μετασυμβόλων (metacharacters) . ^ $ * + ? { } [ ] \ | ( ). Κάθε μετασύμβολο έχει μια καθορισμένη λειτουργία, ενώ ο συνδυασμός τους μπορεί να οδηγήσει στον εντοπισμό πολύ συγκεκριμένων λέξεων και εκφράσεων. [153], [154]

- **TextBlob:** Είναι μια βιβλιοθήκη για την επεξεργασία δεδομένων κειμένου. Παρέχει μία απλή διεπαφή προγραμματισμού εφαρμογών (API) μέσω της οποίας μπορούν να πραγματοποιηθούν κοινές εργασίες επεξεργασίας φυσικής γλώσσας (NLP), όπως προσθήκη μέρους-του-λόγου ετικετών (part-of-speech tagging), εξαγωγή ονοματικής φράσης [6], ανάλυση συναισθήματος, ταξινόμηση, μετάφραση και πολλά άλλα. [155]

- **spaCy:** Το spaCy είναι μια δωρεάν βιβλιοθήκη ανοιχτού κώδικα για προηγμένη επεξεργασία φυσικής γλώσσας (NLP) στην Python. Επιτρέπει την δημιουργία εφαρμογών που μπορούν επεξεργαστούν και να «κατανοήσουν» μεγάλο όγκο κειμένων, ενώ μπορεί να χρησιμοποιηθεί και για τη δημιουργία συστημάτων εξαγωγής πληροφοριών ή κατανόησης φυσικής γλώσσας ή για την προεπεξεργασία κειμένου για βαθιά μάθηση.[156]

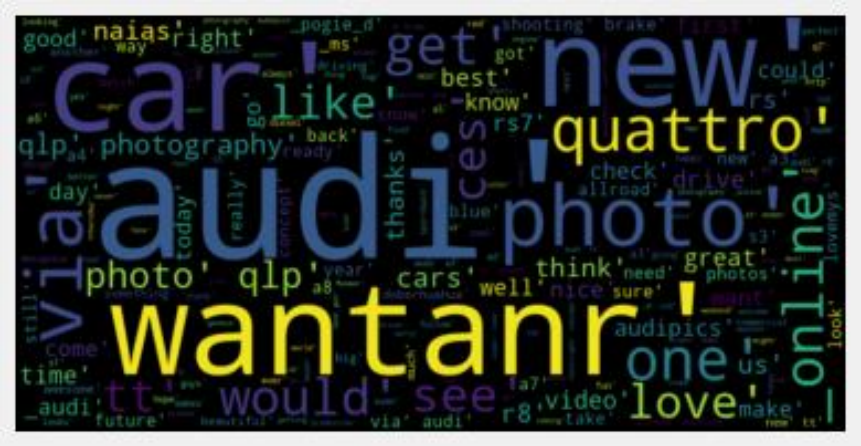
- **matplotlib:** Είναι βασική βιβλιοθήκη για την δημιουργία γραφημάτων στην Python, ενώ αποτελεί και ένα ισχυρό εργαλείο για την εκτέλεση μιας ποικιλίας εργασιών, μεταξύ των οποίων είναι η δημιουργία διαφορετικών τύπων αναφορών απεικόνισης, όπως γραφήματα γραμμών, γραφήματα διασποράς, ιστογράμματα, γραφήματα ράβδων, γραφήματα πίτας, γραφήματα πλαισίων, καθώς επίσης και υποστήριξη τρισδιάστατης σχεδίασης.[157]

- **seaborn:** Είναι μια βιβλιοθήκη απεικόνισης δεδομένων που βασίζεται στο matplotlib. Παρέχει μια διεπαφή υψηλού επιπέδου για τη σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών. Το Seaborn βοηθάει στην εξερεύνηση και την κατανόηση των δεδομένων. Οι συναρτήσεις σχεδίασης λειτουργούν σε πλαίσια δεδομένων και πινάκων που περιέχουν ολόκληρα σύνολα δεδομένων και εκτελούν εσωτερικά την απαραίτητη σημασιολογική χαρτογράφηση και στατιστική συνάθροιση για την παραγωγή ενημερωτικών γραφημάτων. Το API του είναι προσανατολισμένο στο σύνολο δεδομένων επιτρέποντας στον χρήστη να εστιάσει στο τι σημαίνουν τα διάφορα στοιχεία των γραφημάτων και όχι στις λεπτομέρειες του τρόπου με τον οποίο μπορεί να τα σχεδιάστηκαν.[158]

- **Wordcloud:** Με την βιβλιοθήκη αυτή δίνεται η δυνατότητα απεικόνισης των λέξεων που εμφανίζονται μέσα σε ένα κείμενο, σε δομές που μοιάζουν με «σύννεφο». Οι

[6] Ονοματική φράση είναι είτε μια αντωνυμία είτε οποιαδήποτε ομάδα λέξεων που μπορούν να αντικατασταθούν από μια αντωνυμία.

λέξεις απεικονίζονται με διαφορετικά χρώματα και μεγέθη, ανάλογα με τη συχνότητά τους ή τη σημαντικότητά τους (Εικόνα 17).[159]. Είναι ένας γρήγορος τρόπος να αποκτήσουμε μια εικόνα του περιεχομένου ενός κειμένου.



Εικόνα 17. Σύννεφο λέξεων από την εφαρμογή του Wordcloud σε tweets αναφορικά με την εταιρεία Audi

- **string:** Η βιβλιοθήκη αυτή περιέχει λειτουργίες που αφορούν τις συμβολοσειρές. Συμβολοσειρά είναι μια ακολουθία συμβόλων. Με την βιβλιοθήκη αυτή δίνεται η δυνατότητα εντοπισμού σημείων στίξης με την λειτουργία string.punctuation, αριθμητικών ψηφίων με την λειτουργεία string.digits, και άλλες λειτουργίες που διευκολύνουν την επεξεργασία φυσικής γλώσσας. [160]

5.2.2. SPSS (Statistical Package for the Social Sciences)

Το SPSS είναι λογισμικό για την επεξεργασία και την στατιστική ανάλυση όλων των ειδών δεδομένων. Όπως δηλώνουν και τα αρχικά του πρόκειται για ένα "Στατιστικό Πακέτο για τις Κοινωνικές Επιστήμες" και ξεκίνησε για πρώτη φορά το 1968. Δεδομένου ότι το SPSS εξαγοράστηκε από την IBM το 2009, είναι επίσημα γνωστό ως IBM SPSS Statistics, αλλά οι περισσότεροι χρήστες εξακολουθούν να αναφέρονται σε αυτό ως "SPSS".

Τα δεδομένα που μπορεί να επεξεργαστεί μπορεί να προέρχονται από οποιαδήποτε πηγή: επιστημονική έρευνα, βάση δεδομένων πελατών, Google Analytics ή ακόμα και τα αρχεία καταγραφής διακομιστή ενός ιστότοπου. Το SPSS μπορεί να ανοίξει όλες τις μορφές αρχείων που χρησιμοποιούνται συνήθως για δομημένα δεδομένα, όπως υπολογιστικά φύλλα από το MS Excel ή το OpenOffice, αρχεία απλού κειμένου (.txt ή .csv), σχεσιακές βάσεις δεδομένων (SQL), αρχεία Stata και SAS.

Το SPSS χρησιμοποιείται σε πολλούς τομείς, όπως η υγειονομική περίθαλψη, το μάρκετινγκ και η εκπαιδευτική έρευνα, την έρευνα αγοράς, την εξόρυξη δεδομένων, ενώ αποτελεί εργαλείο μελέτης για τους ερευνητές υγείας και εκπαίδευσης, τις εταιρείες ερευνών, τις κυβερνήσεις, τους οργανισμούς μάρκετινγκ κ.α.

Παρέχει την δυνατότητας ανάλυσης δεδομένων για περιγραφική στατιστική, προβλέψεις αριθμητικών αποτελεσμάτων και προσδιορισμό ομάδων (ομαδοποίηση). Επιπλέον, υποστηρίζει λειτουργίες όπως μετασχηματισμό δεδομένων, δημιουργία γραφημάτων, διαχείριση δεδομένων (επιλογή περιπτώσεων, αναδιαμόρφωση αρχείων, δημιουργία παράγωγων δεδομένων) και τεκμηρίωση δεδομένων (ένα λεξικό μεταδεδομένων αποθηκεύεται στο αρχείο δεδομένων).

Τα χαρακτηριστικά του SPSS Statistics είναι προσβάσιμα μέσω αναπτυσσόμενων μενού ή μπορούν να προγραμματιστούν με μια ιδιόκτητη γλώσσα σύνταξης εντολών 4GL. Ο προγραμματισμός σύνταξης εντολών έχει τα πλεονεκτήματα της αναπαραγώγιμης εξόδου, της απλοποίησης επαναλαμβανόμενων εργασιών και του χειρισμού σύνθετων χειρισμών και αναλύσεων δεδομένων. Επιπλέον, ορισμένες σύνθετες εφαρμογές μπορούν να προγραμματιστούν μόνο στη σύνταξη και δεν είναι προσβάσιμες μέσω της δομής του μενού.

Η διεπαφή αναπτυσσόμενου μενού δημιουργεί επίσης σύνταξη εντολών, η οποία μπορεί να εμφανιστεί στην έξοδο, αν και οι προεπιλεγμένες ρυθμίσεις πρέπει να αλλάξουν για να καταστεί η σύνταξη ορατή στον χρήστη. Μπορούν επίσης να επικολληθούν σε ένα αρχείο σύνταξης χρησιμοποιώντας το κουμπί «Επικόλληση» που υπάρχει σε κάθε μενού. Τα προγράμματα μπορούν να εκτελούνται αλληλεπιδραστικά ή χωρίς επίβλεψη, χρησιμοποιώντας την παρεχόμενη λειτουργία παραγωγής εργασίας. (Production Job Facility).

Επιπλέον, μια επέκταση δυνατότητας προγραμματισμού Python μπορεί να έχει πρόσβαση στις πληροφορίες του λεξικού δεδομένων και στα δεδομένα και να δημιουργεί δυναμικά προγράμματα σύνταξης εντολών. Η επέκταση δυνατότητας προγραμματισμού Python, παρουσιάστηκε στην έκδοση 14 και επιτρέπει στο SPSS να «τρέξει» οποιοδήποτε από τα στατιστικά στοιχεία του πακέτου ελεύθερου λογισμικού R. Από την έκδοση 14 και μετά, το SPSS μπορεί να οδηγηθεί εξωτερικά από μια Python ή ένα πρόγραμμα VB.NET χρησιμοποιώντας τις παρεχόμενες «προσθήκες»[7].

[7] Από την έκδοση SPSS 20 και μετά, αυτές οι δύο δυνατότητες δέσμης ενεργειών, καθώς και πολλές δέσμες ενεργειών, περιλαμβάνονται στο μέσο εγκατάστασης και συνήθως εγκαθίστανται από προεπιλογή.

Το SPSS Statistics θέτει περιορισμούς στην εσωτερική δομή αρχείων, τους τύπους δεδομένων, την επεξεργασία δεδομένων και την αντιστοίχιση αρχείων, ο συνδυασμός των οποίων απλοποιεί σημαντικά τον προγραμματισμό. Τα σύνολα δεδομένων SPSS έχουν μια δισδιάστατη δομή πίνακα, όπου οι γραμμές συνήθως αντιπροσωπεύουν περιπτώσεις/εγγραφές (cases) και οι στήλες αντιπροσωπεύουν μετρήσεις ή μεταβλητές (variables) (όπως η ηλικία, το φύλο ή το εισόδημα του νοικοκυριού). Ορίζονται μόνο δύο τύποι δεδομένων: αριθμητικός και κείμενο (ή συμβολοσειρά). Η επεξεργασία των δεδομένων πραγματοποιείται διαδοχικά κατά περίπτωση μέσω του αρχείου (σύνολο δεδομένων). Τα αρχεία μπορούν να αντιστοιχιστούν ένα προς ένα και ένα προς πολλά, αλλά όχι πολλά προς πολλά. Εκτός από αυτή τη δομή και την επεξεργασία κατά περίπτωση, υπάρχει μια ξεχωριστή συνεδρία Matrix όπου μπορεί κανείς να επεξεργαστεί δεδομένα ως πίνακες χρησιμοποιώντας λειτουργίες μήτρας και γραμμικής άλγεβρας.

Η γραφική διεπαφή χρήστη έχει δύο προβολές που μπορούν να ενεργοποιηθούν κάνοντας κλικ σε μία από τις δύο καρτέλες στο κάτω αριστερό μέρος του παραθύρου SPSS Statistics. Η «Προβολή δεδομένων» (Data View) εμφανίζει μια προβολή υπολογιστικού φύλλου των περιπτώσεων (γραμμών) και των μεταβλητών (στήλες). Σε αντίθεση με τα υπολογιστικά φύλλα, τα κελιά δεδομένων μπορούν να περιέχουν μόνο αριθμούς ή κείμενο και οι τύποι δεν μπορούν να αποθηκευτούν σε αυτά τα κελιά. Η «Προβολή μεταβλητής» εμφανίζει το λεξικό μεταδεδομένων όπου κάθε γραμμή αντιπροσωπεύει μια μεταβλητή και εμφανίζει το όνομα της μεταβλητής, την ετικέτα μεταβλητής, τις ετικέτες τιμών, το πλάτος εκτύπωσης, τον τύπο μέτρησης και μια ποικιλία άλλων χαρακτηριστικών. Τα κελιά και στις δύο προβολές μπορούν να επεξεργαστούν με μη αυτόματο τρόπο, ορίζοντας τη δομή του αρχείου και επιτρέποντας την εισαγωγή δεδομένων χωρίς τη χρήση σύνταξης εντολής. Αυτό μπορεί να είναι αρκετό για μικρά σύνολα δεδομένων. Μεγαλύτερα σύνολα δεδομένων, όπως οι στατιστικές έρευνες, δημιουργούνται συχνότερα σε λογισμικό εισαγωγής δεδομένων ή εισάγονται κατά τη διάρκεια προσωπικών συνεντεύξεων με τη βοήθεια υπολογιστή, με σάρωση και χρήση λογισμικού οπτικής αναγνώρισης χαρακτήρων και οπτικής αναγνώρισης σημάτων ή με απευθείας λήψη από διαδικτυακά ερωτηματολόγια. Στη συνέχεια, αυτά τα σύνολα δεδομένων διαβάζονται στο SPSS.

Το SPSS Statistics μπορεί να διαβάσει και να γράψει δεδομένα από αρχεία κειμένου ASCII (συμπεριλαμβανομένων ιεραρχικών αρχείων), άλλα πακέτα στατιστικών,

υπολογιστικά φύλλα και βάσεις δεδομένων, ενώ μπορεί να διαβάσει και να γράψει σε εξωτερικούς σχεσιακούς πίνακες βάσεων δεδομένων μέσω ODBC8 και SQL9.

Τα αρχεία εξαγωγής της στατιστικής ανάλυσης είναι σε μια ειδική μορφή αρχείου (αρχείο*.spv, supporting pivot tables - υποστηρικτικοί συγκεντρωτικοί πίνακες) για την οποία, εκτός από το πρόγραμμα προβολής εντός του πακέτου, υπάρχει δυνατότητα μεταφόρτωσης σε αυτόνομο αναγνώστη. Τα αποτελέσματα της ανάλυσης μπορούν να εξαχθούν σε κείμενο ή Microsoft Word, PDF, Excel και άλλες μορφές.

Το SPSS Statistics κυκλοφόρησε την έκδοση 25 στις 08 Αυγούστου 2017. Το SPSS v.25 προσθέτει νέα και προηγμένα στατιστικά στοιχεία, όπως αποτελέσματα λύσεων τυχαίων εφέ (GENLINMIXED), ισχυρά τυπικά σφάλματα (GLM/UNIANOVA) και γραφήματα προφίλ με γραμμές σφάλματος στο πρόσθετο «Προηγμένα στατιστικά στοιχεία» (Advanced Statistics) και «Προσαρμοσμένοι πίνακες» (Custom Tables). Η έκδοση 25 περιλαμβάνει επίσης νέες δυνατότητες Bayesian στατιστικής, μια μέθοδο στατιστικών συμπερασμάτων και έτοιμων για δημοσίευση γραφημάτων, όπως επίσης νέες ισχυρές δυνατότητες γραφημάτων, συμπεριλαμβανομένων νέων προεπιλεγμένων προτύπων και τη δυνατότητα κοινής χρήσης με εφαρμογές του Microsoft Office. [161]–[166]

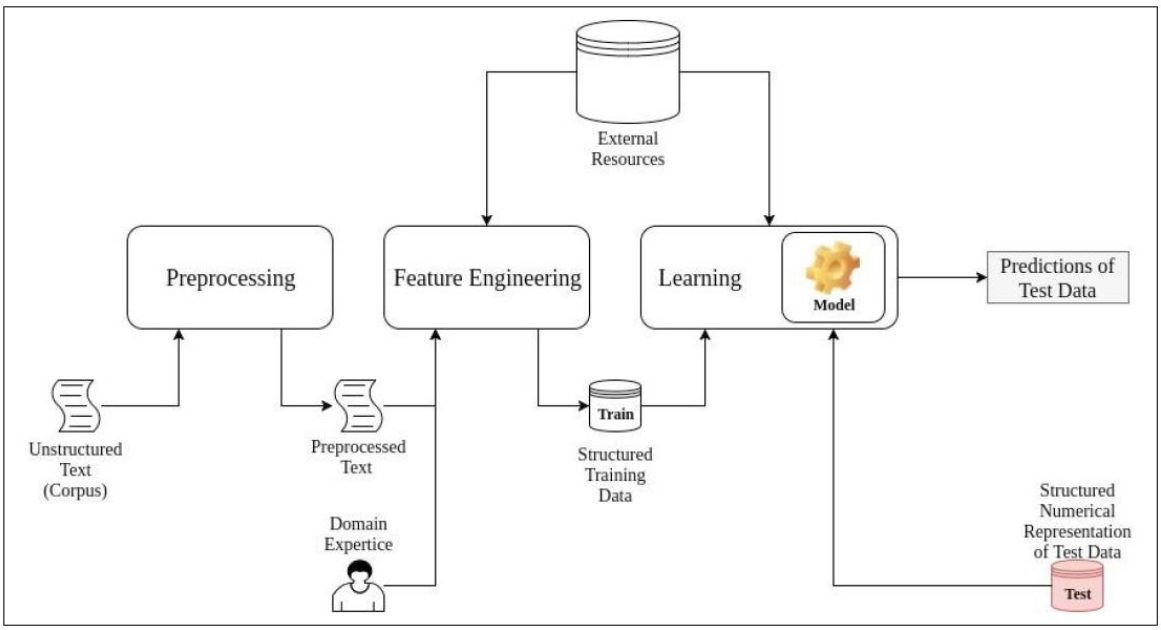
[8] Στην πληροφορική, το Open Database Connectivity (ODBC) είναι μια τυπική διεπαφή προγραμματισμού εφαρμογών (API) για την πρόσβαση σε συστήματα διαχείρισης βάσεων δεδομένων (DBMS)

[9] **SQL** (**Structured Query Language**) είναι μια γλώσσα συγκεκριμένου τομέα που χρησιμοποιείται στον προγραμματισμό και έχει σχεδιαστεί για τη διαχείριση δεδομένων που διατηρούνται σε ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων (RDBMS) ή για την επεξεργασία ροής σε ένα σχεσιακό σύστημα διαχείρισης ροής δεδομένων (RDSMS). Είναι ιδιαίτερα χρήσιμο στο χειρισμό δομημένων δεδομένων, δηλαδή δεδομένων που ενσωματώνουν σχέσεις μεταξύ οντοτήτων και μεταβλητών.

6. Ερευνητική διαδικασία

6.1. Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος των tweets των χρηστών του Twitter, εμπίπτει όπως έχουν αναφέρει, στο πεδίο έρευνας της επεξεργασίας φυσικής γλώσσας. Έτσι, για να πραγματοποιηθεί μια τέτοια ανάλυση, καθίσταται απαραίτητη η εφαρμογή μιας σειράς βημάτων που εμπλέκονται στην κατασκευή ενός μοντέλου NLP, που είναι γνωστή ως αγωγός (pipeline). Τα κύρια συστατικά ενός τυπικού αγωγού για τη σύγχρονη ανάπτυξη συστημάτων επεξεργασίας φυσικής γλώσσας φαίνονται στην παρακάτω Εικόνα 18.



Εικόνα 18. Στάδια επεξεργασίας φυσικής γλώσσας

6.1.1. Βήματα Επεξεργασίας Φυσικής Γλώσσας (NLP Pipeline)

Τα επόμενα στάδια της επεξεργασίας πραγματοποιήθηκαν με την χρήση της Python μέσο του περιβάλλοντος Jupyter Notebook. Παρακάτω εμφανίζονται τα βήματα που ακολουθήθηκαν για τα δεδομένα που αφορούσαν τα tweets χρηστών της αυτοκινητοβιομηχανίας Audi. Η ίδια διαδικασία επαναλήφθηκε στη συνέχεια και για τα δεδομένα των υπόλοιπων αυτοκινητοβιομηχανιών.

6.1.1.1. Συλλογή/εξόρυξη δεδομένων και καθαρισμός

Η συλλογή των δεδομένων πραγματοποιήθηκε με τη βοήθεια του Twitter API. Στη συνέχεια τα αρχεία επεξεργάστηκαν από την εφαρμογή LIWC, προκειμένου να αναλυθούν και να εξαχθούν για το κάθε tweet οι βαθμολογίες που αντιστοιχούν στην κάθε μεταβλητή του λεξικού του LIWC. Μετά την ανάλυση τα δεδομένα εξήχθησαν σε

αρχεία csv, τα οποία στη συνέχεια μετατράπηκαν σε αρχεία excel για την περαιτέρω επεξεργασία τους στην Python και το SPSS.

Στην αρχική τους μορφή τα δεδομένα που συλλέχθηκαν είχαν άναρχη δομή, καθώς περιείχαν διπλές εγγραφές του ίδιου tweet από τον ίδιο χρήστη και εγγραφές χωρίς κατάλληλο περιεχόμενο προς επεξεργασία. Τα βήματα που ακολουθήσαμε προκειμένου να φέρουμε την βάση δεδομένων μας στην κατάλληλη μορφή ώστε να πραγματοποιηθεί στη συνέχεια η ανάλυση συναισθήματος περιγράφονται παρακάτω.

- Εισαγάγαμε την βιβλιοθήκη pandas προκειμένου να επεξεργαστούμε τον πίνακα δεδομένων (Εικόνα 19).

```
In [2]: # Εισαγωγή της βιβλιοθήκης pandas για επεξεργασία του πίνακα δεδομένων
        import pandas as pd
```

Εικόνα 19. Εισαγωγή της βιβλιοθήκης pandas

- Εισαγάγαμε τα δεδομένα του αρχείου "Audi_SuperBowl_Ready.xlsx" που περιλαμβάνει τα tweets που αφορούν την αυτοκινητοβιομηχανία Audi (Εικόνα 20).

```
In [3]: # Εισαγωγή δεδομένων του αρχείου "Audi_SuperBowl_Ready.xlsx" που περιλαμβάνει τα tweets που αφορούν την αυτοκινητοβιομηχανία Audi
        df = pd.read_excel(r"G:\OneDrive\DATA SCIENCE\MBADS\MBADS_Διπλωματική\Datasets_Excel\Unclear_Datasets\Audi_SuperBowl_Ready.xlsx")
```

Εικόνα 20. Εισαγωγή δεδομένων για την αυτοκινητοβιομηχανία Audi

- Αφαιρέσαμε τις εγγραφές που δεν περιείχαν καθόλου δεδομένα (αναφέρονται ως NaN – Not a Number) (Εικόνα 21).

```
In [6]: # Αφαίρεση των NAN (Not a number) τιμών
        df.dropna(inplace = True)

In [7]: df
```

Out[7]:

	Timestamp	Tweet	TextLenght	Funct	Pronoun	Ppron	I	We	You	SheHe	...	Achiev	Leisure	Home	Money	Relig	Death	Assent	Nonflu	Filler	UserID
0	fri feb 07 00:00:21 +0000 2014	rt @audi: 4 amazing race tracks. 21 hours. 1 g...	18	11.11	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	105855418
1	fri feb 07 00:00:30 +0000 2014	#throwbackthursday #audi #july2013 @princetona...	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	57731606
2	fri feb 07 00:00:37 +0000 2014	rt @audi: zero to 60 in under 4.6 seconds. #au...	14	28.57	0	0	0	0	0	0	...	0	7.14	0	0	0	0	0	0	0	105855418
3	fri feb 07 00:00:39 +0000 2014	rt @audi: we certainly do. #audinaias mt @naia...	12	41.67	16.67	16.67	0	8.33	8.33	0	...	0	0	0	0	0	0	8.33	0	0	105855418
4	fri feb 07 00:00:46 +0000 2014	rt @audi: stadler: for almost 20 years, the au...	17	41.18	5.88	5.88	0	5.88	0	0	...	5.88	5.88	5.88	0	0	0	0	0	0	105855418
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
51689	wed mar 05 23:52:15 +0000 2014	rt @audi: appropriate plates. rt @balllllnoutta...	17	35.29	5.88	5.88	5.88	0	0	0	...	0	0	0	0	0	0	0	0	0	353981470
51690	wed mar 05 23:54:38 +0000 2014	rt @audi: appropriate plates. rt @balllllnoutta...	17	35.29	5.88	5.88	5.88	0	0	0	...	0	0	0	0	0	0	0	0	0	348730081
51691	wed mar 05 23:57:55 +0000 2014	@audi "truth in engineering""but not in servi...	14	50	14.29	14.29	14.29	0	0	0	...	7.14	0	0	0	0	0	0	0	0	335781593
51692	wed mar 05 23:58:24 +0000 2014	@audi what made spencer's mom in mrs dilaretti...	11	27.27	9.09	0	0	0	0	0	...	0	0	9.09	0	0	0	0	0	0	2203842577
51693	wed mar 05 23:58:45 +0000 2014	@audi i'm looking for a q5 but this one must b...	14	57.14	7.14	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1384547971

51694 rows × 68 columns

Εικόνα 21. Αφαίρεση των εγγραφών χωρίς περιεχόμενο

Για την ανάλυση συναισθήματος θα χρησιμοποιήσουμε τα δεδομένα κειμένου των tweets. Έτσι προχωρήσαμε στην δημιουργία ενός νέου dataframe που περιείχε μόνο την στήλη ***Tweet*** (Εικόνα 22).

```
In [31]: # Επιλογή στήλης "Tweet"

         df = df[['Tweet']]
         df
```

Out[31]:

	Tweet
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...
4	@yansarazin @adammcnamara well we know that won't happen @audi
...	...
51687	rt @audi: #tbt rt @cnettv: the audi tt is one of the most recognizable sports cars on the road. xcar looks back at its past... http://t.co/â€¦
51688	rt @audi: "too hot to hold, too much to handle." rt @jermainedupri: 2015 audi tt: audi debuts the 3rd generation of the audi tt http://t.coâ€¦
51689	rt @audi: irrefutable evidence that ghosts are real. rt @superstaro: @audi #50platesofaudi #50statesofaudi the ghost http://t.co/mabqohppg2
51690	@audi good cars
51691	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?

51692 rows × 1 columns

Εικόνα 22. Επιλογή της στήλης Tweet από τον αρχικό πίνακα δεδομένων

- Στη συνέχεια προχωρήσαμε στην αφαίρεση των πολλαπλών εγγραφών των ίδιων tweets. Οι πολλαπλές εγγραφές είτε είναι tweets που συλλέχτηκαν πολλές φορές με μια μικρή χρονική καθυστέρηση, είτε είναι retweets διαφορετικών χρηστών, που όμως δεν γνωρίζουμε αν η επαναδημοσίευση τους αντιπροσωπεύει απόλυτα τον ίδιο τον χρήστη, και κατά συνέπεια το συναίσθημα που εκφράζει. (Εικόνα 23)

```
In [10]: # Καταμέτρηση των διπλότυπων εγγραφών

         df['Tweet'].duplicated().sum()
```

Out[10]: 31151

```
In [11]: # Αφαίρεση διπλότυπων εγγραφών

         df = df.drop_duplicates(subset=['Tweet'], keep='first')
```

```
In [12]: # Εκ νέου καταμέτρηση των διπλότυπων Tweets

         df['Tweet'].duplicated().sum()
```

Out[12]: 0

```
In [13]: pd.set_option("display.max_colwidth", None)
```

```
In [14]: df.reset_index(drop=True)
```

Out[14]:

	Tweet
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...
4	@yansarazin @adammcnamara well we know that won't happen @audi
...	...
20536	get out of the whip and into the chopperðŸ˜ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy
20539	@audi good cars
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?

20541 rows × 1 columns

Εικόνα 23. Καταμέτρηση και αφαίρεση διπλότυπων εγγραφών

Όπως παρατηρούμε ο αριθμός των διαθέσιμων tweets προς ανάλυση μειώθηκε από 51692 σε 20541 tweets. Η εντολή pd.set_option ("display.max_colwidth", None) δίνει την δυνατότητα εμφάνισης ολόκληρου του tweet της κάθε εγγραφής. Μετά την αφαίρεση των διπλότυπων εγγραφών κάθε tweet συνοδευόταν από τον αρχικό αριθμητικό του δείκτη (index). Με την εντολή df.reset_index(drop=True) έγινε επαναπροσδιορισμός της αρίθμησης, έτσι ώστε να εμφανίζονται τα εναπομείναντα tweets με διαδοχική αρίθμηση. Η ανάλυση συναισθήματος που θα ακολουθήσει στη συνέχεια θα πραγματοποιηθεί όσο στο αρχικό δείγμα, όσο και στο δείγμα μετά την αφαίρεση των πολλαπλών εγγραφών, ώστε να γίνει στη συνέχεια η αποτίμηση του αποτελέσματος και το κατά πόσο επηρεάστηκε από αυτή την μείωση του δείγματος.

- ***Αφαίρεση κανονικών εκφράσεων:*** Τα δεδομένα που συλλέχθηκαν αποτελούνταν από tweets και retweets, τα οποία είχαν μη δομημένη μορφή, καθώς στο σύνολό τους, εκτός από λέξεις, περιείχαν και συνδέσμους που παρέπεμπαν σε άλλα site (hyper links), emoticons, hashtags (#), αναφορές (@) κ.α. Σε αυτό το στάδιο εισαγάγαμε την βιβλιοθήκη re, ώστε να χρησιμοποιήσουμε τις κανονικές εκφράσεις μέσα στη συνάρτηση που δημιουργήσαμε για τον καθαρισμό των tweets από τα παραπάνω στοιχεία (Εικόνες 24).

```
In [15]: # Εισαγωγή της βιβλιοθήκης re
         import re
```

Εικόνα 24. Εισαγωγή βιβλιοθήκης re

Δημιουργήσαμε στη συνέχεια μία συνάρτηση που περιλαμβάνει τα βήματα καθαρισμού των tweets από τις κανονικές εκφράσεις (Εικόνα 25).

```
In [16]: # Δημιουργία συνάρτησης καθαρισμού των tweets

def cleanText(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text) # Αφαίρεση αναφορών (@mentions)
    text = re.sub(r'#', '', text) # Αφαίρεση συμβόλων hashtag (#)
    text = re.sub(r'rt[\s]+', '', text) # Αφαίρεση προθέματος rt των retweets
    text = re.sub(r'https?:\/\/\S+', '', text) # Αφαίρεση υπερσυνδέσμων

    return text
```

Εικόνα 25. Δημιουργία συνάρτησης καθαρισμού των tweets

Η εφαρμογή της συνάρτησης cleanText στα tweets (Εικόνα 26) έδωσε τα αποτελέσματα που φαίνονται στην στήλη ***Tweet_clean*** (Εικόνα 27).

```
In [17]: # Εκκαθάριση δεδομένων

         df['Tweet_clean']=df['Tweet'].apply(cleanText)
```

Εικόνα 26. Εφαρμογή συνάρτησης cleanText στα δεδομένα της στήλης Tweet

```
In [17]: # Εκκαθάριση δεδομένων

         df['Tweet_clean']=df['Tweet'].apply(cleanText)

In [18]: # Εμφάνιση αποτελεσμάτων μετά την εκκαθάριση

         pd.set_option("display.max_colwidth", None)
         df
```

Out[18]:

	Tweet	Tweet_clean
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"	: lol good job ": new ad done perfectly right! sochi
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry	i don't concur. i want to talk with my brands. seriously. i'd dig chatting w , , ,
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.	should add a kindle book reader so you can read while you're driving.
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...	i hope they build it, but i will settle for an rs6avant if you want to loan me one...
4	@yansarazin @adammcnamara well we know that won't happen @audi	well we know that won't happen
...	...	...
20536	get out of the whip and into the chopperðŸ˜Ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc	get out of the whip and into the chopperðŸ˜Ž thatlife money power respect
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?	just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy	audi r8 with custom painted calipers and _porn
20539	@audi good cars	good cars
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?	": drove an to and from pretoria! mmmh one day is one day lol" which audi model?

20541 rows × 2 columns

Εικόνα 27. Εκκαθάριση δεδομένων και εμφάνιση αποτελεσμάτων (Tweet_clean)

6.1.1.2.Προεπεξεργασία

Μετά την ολοκλήρωση των παραπάνω εργασιών ακολούθησε η προεπεξεργασία του κειμένου των tweets, η οποία πραγματοποιήθηκε με την βοήθεια της βιβλιοθήκης nltk της python. Η προεπεξεργασία αποτελείται συνήθως από τις εξής διαδικασίες:

- **Διακριτοποίηση (Tokenization):** Είναι η διαδικασία διαχωρισμού μιας πρότασης στις λεκτικές μονάδες (token) που την αποτελούν. Κάθε λεκτική μονάδα φέρει μια σημασιολογική έννοια που σχετίζεται με αυτή και μπορεί να είναι λέξη, αριθμός, σημείο στίξης, σύμβολο και, μερικές φορές, συνδυασμός των τελευταίων με αποτέλεσμα την δημιουργία των λεγόμενων emoticons. Η διακριτοποίηση μπορεί να θεωρηθεί ως μια τεχνική τμηματοποίησης κατά την οποία μεγάλα τμήματα κειμένου αναλύονται σε μικρότερα που έχουν λεκτικό νόημα. [147].

⊙ Εισαγάγαμε την βιβλιοθήκη nltk στο περιβάλλον εργασίας μας, καθώς και την λειτουργία TweetTokenizer για την διακριτοποίηση των tweets (Εικόνα 28).

```
In [21]: # Εισαγωγή της βιβλιοθήκης nltk για την επεξεργασία φυσικής γλώσσας και της λειτουργίας TweetTokenizer για την διακριτοποίηση των tweets

         import nltk
         from nltk.tokenize import TweetTokenizer
```

Εικόνα 28. Εισαγωγή βιβλιοθήκης nltk και λειτουργίας TweetTokenizer

Η λειτουργία TweetTokenizer, αναλύει την άτυπη γλώσσα των μηνυμάτων του κοινωνικού μέσου Twitter (ετικέτες, emoticons, hashtags, σύντομες εκφράσεις κ.α) με στόχο να την μετατρέψει σε όσο το δυνατόν πιο φυσιολογική και περισσότερο κατανοητή μορφή.

- Στη συνέχεια εφαρμόσαμε την λειτουργία TweetTokenizer στα tweets της στήλης Tweet_clean, εξάγοντας τα αποτελέσματα της διακριτοποίησης σε νέα στήλη με το όνομα ***Tokens***. Τα στοιχεία (elements) του κάθε tweet είναι αποθηκευμένα πλέον σε μια νέα μορφή που στην Python ονομάζεται λίστα (list). Σε αυτή την μορφή είναι εύκολο να εφαρμόσουμε και τις υπόλοιπες διαδικασίες της προεπεξεργασίας. (Εικόνα 29)

```
In [25]: # Διακριτοποίηση (tokenization) των tweets

df_clean['Tokens'] = df_clean['Tweet_clean'].apply(TweetTokenizer().tokenize)
df_clean.head()
```

Out[25]:

	Tweet_clean	Tokens
0	: lol good job ": new ad done perfectly right! sochi	[:, lol, good, job, ", :, new, ad, done, perfectly, right, !, sochi]
1	i don't concur. i want to talk with my brands. seriously. i'd dig chatting w , , ,	[i, don't, concur, ., i, want, to, talk, with, my, brands, ., seriously, ., i'd, dig, chatting, w, ,, ,, ,]
2	should add a kindle book reader so you can read while you're driving.	[should, add, a, kindle, book, reader, so, you, can, read, while, you're, driving, .]
3	i hope they build it, but i will settle for an rs6avant if you want to loan me one...	[i, hope, they, build, it, ,, but, i, will, settle, for, an, rs6avant, if, you, want, to, loan, me, one, ...]
4	well we know that won't happen	[well, we, know, that, won't, happen]

Εικόνα 29. Διακριτοποίηση των tweets και αποτελέσματα (Tokens)

- ✓ **Αφαίρεση λέξεων διακοπής (Stopwords removal):** Ορισμένες από τις λέξεις που χρησιμοποιούνται συχνά στα αγγλικά, όπως a, an, the, of, in, κ.λπ., δεν συμβάλλουν στην έννοια μιας πρότασης για τους σκοπούς της επεξεργασίας φυσικής γλώσσας, καθώς δεν φέρουν κανένα περιεχόμενο από μόνες τους. Τέτοιες λέξεις ονομάζονται λέξεις διακοπής και συνήθως αφαιρούνται από την περαιτέρω ανάλυση. Ωστόσο, δεν υπάρχει κάποια τυποποιημένη λίστα λέξεων διακοπής για τα αγγλικά, καθώς αυτές οι λίστες είναι δυναμικές και καθορίζονται με βάση και το περιεχόμενο του κειμένου στο οποίο γίνεται η ανίχνευση τους. Υπάρχουν μερικές δημοφιλείς λίστες, όπως αυτή που περιλαμβάνει η βιβλιοθήκη NLTK, οι λέξεις τις οποίας εμφανίζονται στην παρακάτω Εικόνα 30.[167], [168]

```
In [25]: from nltk.corpus import stopwords
nltk.download('stopwords')
language = "english"
stop_words = set (stopwords.words(language))
print(stop_words)
```

```
{'only', "she's", 'before', "doesn't", "shan't", "isn't", 'theirs', "shouldn't", "it's", 'mustn', 'these', 'because', 'any', 'own', 'they', 'down', 'against', "weren't", 'all', "hadn't", 'ai
n', "didn't", 'nor', 'doesn', 'same', 'while', 'not', 'in', 'further', 'o', 'didn', 'again', 'if', 'of', 'under', 'other', 'am', 'those', 'is', 'below', 't', 'won', 'you', 'the', 'having', "h
asn't", 'isn', 'hers', 'been', 'too', 'our', 'than', "couldn't", 'aren', 'we', 'himself', 'where', 'by', 'has', 'and', 'couldn', 'at', 'him', 'are', 'which', 'then', "you've", 'most', "are
n't", 'were', 'to', 'there', 'ourselves', 'out', 'doing', 'as', "mightn't", 'can', 'myself', "you'd", 'no', 'on', 'during', 'does', "haven't", 'being', 'weren', 'yours', "you'll", 'do', 'ha
d', 'ma', 'between', 'why', 'wouldn', 'it', 'm', 'each', 'when', 'over', 'both', 'yourself', 'how', 'their', 'me', 'into', 'y', 'for', 'll', 'now', 'shouldn', "don't", 'once', 'about', 'had
n', 've', 'hasn', "mustn't", 'ours', 'whom', 'my', 'this', 'your', 'd', 'just', 'through', 'have', 'shan', 'more', 'such', 're', 'itself', 'mightn', 'that', 'don', 'themselves', 'or', 'was',
'very', 'did', 'so', "you're", 'its', 'above', 'from', 'until', 'but', 'he', 'an', 'needn', 'wasn', 'some', 's', 'she', "won't", 'her', 'with', "that'll", 'what', "wouldn't", 'his', 'few', 'a
fter', 'i', 'up', 'off', "should've", "needn't", 'a', 'here', 'haven', "wasn't", 'yourselves', 'be', 'will', 'them', 'should', 'who', 'herself'}
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\petro\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Εικόνα 30. Εισαγωγή της λειτουργίας stopwords και εμφάνιση των λέξεων διακοπής

⊙ Μετά την εισαγωγή και την εφαρμογή της λειτουργίας stopwords, εξαγάγαμε τα αποτελέσματα στη στήλη Tweet_no_sw (Εικόνα 31).

```
In [25]: # Αφαίρεση λέξεων διακοπής από τα tweets

stopwords = nltk.corpus.stopwords.words('english')
df_clean['Tweet_no_sw'] = df_clean['Tokens'].apply(lambda x: [i for i in x if i.lower() not in stopwords])

df_clean.head(10)
```

Out[25]:

	Tweet_clean	Tokens	Tweet_no_sw
0	: lol good job ": new ad done perfectly right! sochi	[:, lol, good, job, ", :, new, ad, done, perfectly, right, !, sochi]	[:, lol, good, job, ", :, new, ad, done, perfectly, right, !, sochi]
1	i don't concur. i want to talk with my brands. seriously. i'd dig chatting w , , ,	[i, don't, concur, ., i, want, to, talk, with, my, brands, ., seriously, ., i'd, dig, chatting, w, ,, ,, ,]	[concur, ., want, talk, brands, ., seriously, ., i'd, dig, chatting, w, ,, ,, ,]
2	should add a kindle book reader so you can read while you're driving.	[should, add, a, kindle, book, reader, so, you, can, read, while, you're, driving, .]	[add, kindle, book, reader, read, driving, .]
3	i hope they build it, but i will settle for an rs6avant if you want to loan me one...	[i, hope, they, build, it, ,, but, i, will, settle, for, an, rs6avant, if, you, want, to, loan, me, one, ...]	[hope, build, ,, settle, rs6avant, want, loan, one, ...]
4	well we know that won't happen	[well, we, know, that, won't, happen]	[well, know, happen]
5	: no computer-generated animals were harmed making this big game commercial	[:, no, computer-generated, animals, were, harmed, making, this, big, game, commercial, ...]	[:, computer-generated, animals, harmed, making, big, game, commercial, ...]
6	hereâ€™s the original image used in the fake sochi four rings ad.	[hereâ, €, ™, s, the, original, image, used, in, the, fake, sochi, four, rings, ad, .]	[hereâ, €, ™, original, image, used, fake, sochi, four, rings, ad, .]
7	speaking of throwback websites, check out this sweet flash app i made in 1999 for the s4:	[speaking, of, throwback, websites, ,, check, out, this, sweet, flash, app, i, made, in, 1999, for, the, s4, :]	[speaking, throwback, websites, ,, check, sweet, flash, app, made, 1999, s4, :]
8	_pixel: amazing : the marketing people at need to get a raise for this pronto sochiâ€¦	[_pixel, :, amazing, :, the, marketing, people, at, need, to, get, a, raise, for, this, pronto, sochiâ, €, ¦]	[_pixel, :, amazing, :, marketing, people, need, get, raise, pronto, sochiâ, €, ¦]
9	golf r *finally* available in the u.s. ...but 4-door only, so basically pointless. wtf	[golf, r, *, finally, *, available, in, the, u, ., s, ..., but, 4, -, door, only, ,, so, basically, pointless, ., wtf]	[golf, r, *, finally, *, available, u, ., ..., 4, -, door, ,, basically, pointless, ., wtf]

Εικόνα 31. Αφαίρεση των λέξεων διακοπής και εμφάνιση αποτελεσμάτων (Tweet_no_sw)(Audi)

✓ **Αφαίρεση ειδικών χαρακτήρων (Remove special words):** Στην Python μια ακολουθία από σύμβολα ονομάζεται συμβολοσειρά (string). Μια συμβολοσειρά είναι συνήθως μια λέξη, η οποία μπορεί να συνοδεύεται και από άλλα σύμβολα που ονομάζονται ειδικοί χαρακτήρες και στα πλαίσια μια πρότασης χρησιμεύουν στην σωστή σύνταξη, στην κατανόηση νοημάτων, στην έκφραση συναισθημάτων κ.α. Οι χαρακτήρες αυτοί μπορεί να είναι σημεία στίξης (punctuation), ψηφία (digits), χαρακτήρες ASCII κ.α. Η εισαγωγή της βιβλιοθήκης που περιέχει τις λειτουργίες για τις συμβολοσειρές, η αποθήκευση των σημείων στίξης και η εμφάνισή τους έπειτα φαίνονται στην παρακάτω Εικόνα 32.[169]

```
In [26]: # Εισαγωγή λειτουργιών βιβλιοθήκης συμβολοσειρών (string)
import string

# Αποθήκευση των σημείων στίξης στην μεταβλητή result
result = string.punctuation

# Εμφάνιση της λίστας των σημείων στίξης
print(result)
```

```
!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~
```

Εικόνα 32. Εισαγωγή λειτουργιών συμβολοσειρών και εμφάνιση των σημείων στίξης

- Τα σημεία στίξης και τα ψηφία δεν αποτελούν χρήσιμες πληροφορίες για την ανάλυση συναισθήματος στην οποία αποσκοπούμε και έτσι προχωρήσαμε στην εκκαθάριση των tweets από αυτά, όπως φαίνεται στην παρακάτω Εικόνα 33. Οι στήλες punctuation και digits περιέχουν τα tweets μετά την αφαίρεση των σημείων στίξης και των ψηφίων αντίστοιχα, ενώ τα tweets στην μορφή που θα χρησιμοποιηθούν για την ανάλυση συναισθήματος αποθηκεύτηκαν στην στήλη ***Tweet_semifinal***, όπου έγινε επιλογή εκείνων με αριθμό λέξεων μεγαλύτερο του ενός, ώστε να έχει νόημα η ανάλυσή του.[169]

In [27]:
```
# Αποθήκευση των σημείων στίξης στην μεταβλητή punctuations
punctuations = list(string.punctuation)

# Αφαίρεση σημείων στίξης από την στήλη Tweet_no_sw
df_clean['punctuation'] = df_clean['Tweet_no_sw'].apply(lambda x: [i for i in x if i not in punctuations])


# Αφαίρεση ψηφίων (αριθμών) από την στήλη punctuation
df_clean['digits'] = df_clean['punctuation'].apply(lambda x: [i for i in x if i[0] not in list(string.digits)])


# Επιλογή των tweets με αριθμό λέξεων μεγαλύτερο του 1
df_clean['Tweet_semifinal'] = df_clean['digits'].apply(lambda x: [i for i in x if len(i) > 1])

df_clean.head()
```

Out[27]:

	Tweet_clean	Tokens	Tweet_no_sw	punctuation	digits	Tweet_semifinal
0	: lol good job ": new ad done perfectly right! sochi	[:, lol, good, job, ", :, new, ad, done, perfectly, right, !, sochi]	[:, lol, good, job, ", :, new, ad, done, perfectly, right, !, sochi]	[lol, good, job, new, ad, done, perfectly, right, sochi]	[lol, good, job, new, ad, done, perfectly, right, sochi]	[lol, good, job, new, ad, done, perfectly, right, sochi]
1	i don't concur. i want to talk with my brands. seriously. i'd dig chatting w , , ,	[i, don't, concur, ., i, want, to, talk, with, my, brands, ., seriously, ., i'd, dig, chatting, w, ,, ,, ,]	[concur, ., want, talk, brands, ., seriously, ., i'd, dig, chatting, w, ,, ,, ,]	[concur, want, talk, brands, seriously, i'd, dig, chatting, w]	[concur, want, talk, brands, seriously, i'd, dig, chatting, w]	[concur, want, talk, brands, seriously, i'd, dig, chatting]
2	should add a kindle book reader so you can read while you're driving.	[should, add, a, kindle, book, reader, so, you, can, read, while, you're, driving, .]	[add, kindle, book, reader, read, driving, .]	[add, kindle, book, reader, read, driving]	[add, kindle, book, reader, read, driving]	[add, kindle, book, reader, read, driving]
3	i hope they build it, but i will settle for an rs6avant if you want to loan me one...	[i, hope, they, build, it, ,, but, i, will, settle, for, an, rs6avant, if, you, want, to, loan, me, one, ...]	[hope, build, ,, settle, rs6avant, want, loan, one, ...]	[hope, build, settle, rs6avant, want, loan, one, ...]	[hope, build, settle, rs6avant, want, loan, one, ...]	[hope, build, settle, rs6avant, want, loan, one, ...]
4	well we know that won't happen	[well, we, know, that, won't, happen]	[well, know, happen]	[well, know, happen]	[well, know, happen]	[well, know, happen]

Εικόνα 33. Αφαίρεση σημείων στίξης, αριθμών και επιλογή tweets με περισσότερες από 1 λέξεις (Audi)

- ✓ **Στελέχωση ή αποκοπή καταλήξεων (Stemming):** Η στελέχωση αναφέρεται στη διαδικασία αφαίρεσης των καταλήξεων και σμίκρυνση μιας λέξης σε κάποια βασική μορφή έτσι ώστε όλες οι διαφορετικές παραλλαγές αυτής της λέξης να μπορούν να αναπαρασταθούν με την ίδια μορφή. Για παράδειγμα οι λέξεις *computer*, *computerization*, και *computerize* αντικαθίστανται και οι τρεις από την λέξη *compute*. Αυτό επιτυγχάνεται με την εφαρμογή ενός σταθερού συνόλου κανόνων (π.χ. εάν η λέξη τελειώνει σε "-es", αφαιρείτε το "-es"). Παρόλο που τέτοιοι κανόνες μπορεί να μην καταλήγουν πάντα σε μια γλωσσικά σωστή βασική μορφή, η στελέχωση χρησιμοποιείται συνήθως στις μηχανές αναζήτησης για την αντιστοίχιση των ερωτημάτων των χρηστών με τα σχετικά έγγραφα και στην ταξινόμηση κειμένου για τη μείωση του χώρου των χαρακτηριστικών για την εκπαίδευση μοντέλων μηχανικής μάθησης[168]. Στην ανάλυση συναισθήματος που ακολουθεί αρχικά εφαρμόστηκε η διαδικασία της στελέχωσης, η οποία όπως θα δούμε στη συνέχεια

μας οδήγησε σε διαφορετικά αποτελέσματα σε σύγκριση με εκείνα χωρίς την εφαρμογή αυτής.

- Για την συγκεκριμένη διαδικασία εισαγάγαμε τον αλγόριθμο PorterStemmer και στη συνέχεια τον εφαρμόσαμε στα tweets της στήλης ***Tweet_semifinal***. Τα αποτελέσματα αποθηκεύτηκαν στην στήλη ***Tweet_final*** (Εικόνα 34).

```
In [29]: # Εισαγωγή της λειτουργίας στελέχωσης

from nltk.stem.porter import *
stemmer = PorterStemmer()
```

```
In [30]: # Στελέχωση των tweets

df_final['Tweet_final'] = df_final['Tweet_semifinal'].apply(lambda x: [stemmer.stem(i) for i in x])
df_final
```

```
<ipython-input-30-ab81702adb1c>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_final['Tweet_final'] = df_final['Tweet_semifinal'].apply(lambda x: [stemmer.stem(i) for i in x])
```

Out[30]:

	Tweet_semifinal	Tweet_final
0	[lol, good, job, new, ad, done, perfectly, right, sochi]	[lol, good, job, new, ad, done, perfectli, right, sochi]
1	[concur, want, talk, brands, seriously, i'd, dig, chatting]	[concur, want, talk, brand, serious, i'd, dig, chat]
2	[add, kindle, book, reader, read, driving]	[add, kindl, book, reader, read, drive]
3	[hope, build, settle, rs6avant, want, loan, one, ...]	[hope, build, settl, rs6avant, want, loan, one, ...]
4	[well, know, happen]	[well, know, happen]
...	...	...
20536	[get, whip, chopperðÿ, thatlife, money, power, respect]	[get, whip, chopperðÿ, thatlif, money, power, respect]
20537	[got, two, half, year, old, audi, a6, avant, line, love, mpg, expect]	[got, two, half, year, old, audi, a6, avant, line, love, mpg, expect]
20538	[audi, r8, custom, painted, calipers, _porn]	[audi, r8, custom, paint, calip, _porn]
20539	[good, cars]	[good, car]
20540	[drove, pretoria, mmmh, one, day, one, day, lol, audi, model]	[drove, pretoria, mmmh, one, day, one, day, lol, audi, model]

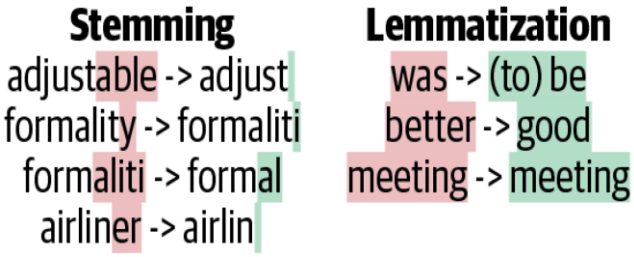
20541 rows × 2 columns

Εικόνα 34. Εισαγωγή λειτουργίας στελέχωσης, εφαρμογή και εμφάνιση αποτελεσμάτων (Tweet_final)

Μία επιπλέον διεργασία της προεπεξεργασίας των δεδομένων, που συνήθως όμως δεν είναι απαραίτητη στην περίπτωση ανάλυσης συναισθήματος μηνυμάτων των κοινωνικών μέσων είναι η λημματοποίηση (lemmatization), η οποία περιγράφεται παρακάτω.

- ✓ **Λημματοποίηση (lemmatizaion):** Είναι η διαδικασία αντιστοίχισης όλων των διαφορετικών μορφών μιας λέξης στη βασική της λέξη, ή *λήμμα (lemma)*. Η λημματοποίηση σε αντίθεση με την στελέχωση, δεν έχει ως στόχο την σμίκρυνση μιας λέξης, αλλά την αντικατάστασή της με μία άλλη που την αντιπροσωπεύει καλύτερα και διευκολύνει την επεξεργασία φυσικής γλώσσας. Για παράδειγμα, το επίθετο “better”, όταν υποστεί στελέχωση, παραμένει το ίδιο. Κατά την λημματοποίησή τους όμως θα πρέπει να αντικατασταθεί από την λέξη “good”.

Κάποια παραδείγματα στελέχωσης και λημματοποίησης μέσα από τα οποία αναδεικνύεται η διαφορά τους δίνονται στην παρακάτω Εικόνα 35. [168]



Εικόνα 35. Διαφορά μεταξύ στελέχωσης και λημματοποίησης [170]

Η λημματοποίηση απαιτεί περισσότερες γλωσσικές γνώσεις και η μοντελοποίηση και η ανάπτυξη αποτελεσματικών λημματοποιητών παραμένει ένα ανοιχτό πρόβλημα στην έρευνα της επεξεργασίας φυσικής γλώσσας ακόμη και σήμερα.

- Πριν προχωρήσουμε στο επόμενο βήμα και προκειμένου να έχουμε μια εικόνα των λέξεων που εμφανίζονται συχνότερα στα tweets των χρηστών για την αυτοκινητοβιομηχανία Audi, δημιουργήσαμε ένα σύννεφο λέξεων με την βοήθεια της βιβλιοθήκης Wordcloud. Αρχικά έγινε εισαγωγή της διαδικασίας WordCloud της βιβλιοθήκης wordcloud, και της βιβλιοθήκης matplotlib για την σχεδίαση του σύννεφου. Η διαδικασία που ακολουθήσαμε, καθώς και το αποτέλεσμα φαίνονται στις παρακάτω Εικόνες 36 και 37.

In [31]:
```
# Εισαγωγή διαδικασίας WordCloud από την βιβλιοθήκη wordcloud για την δημιουργία σύννεφου λέξεων
from wordcloud import WordCloud

# Εισαγωγή βιβλιοθήκης matplotlib για σχεδίαση γραφικών
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
```

In [32]:
```
# Δημιουργία σύννεφου λέξεων με βάση την συχνότητα και την σημαντικότητα αυτών

words = ','.join(str(w) for w in df_final.Tweet_final)
wordCloud = WordCloud(width=800, height=500).generate(words)

plt.imshow(wordCloud, interpolation = 'bilinear')
plt.axis('off')
plt.show()
```

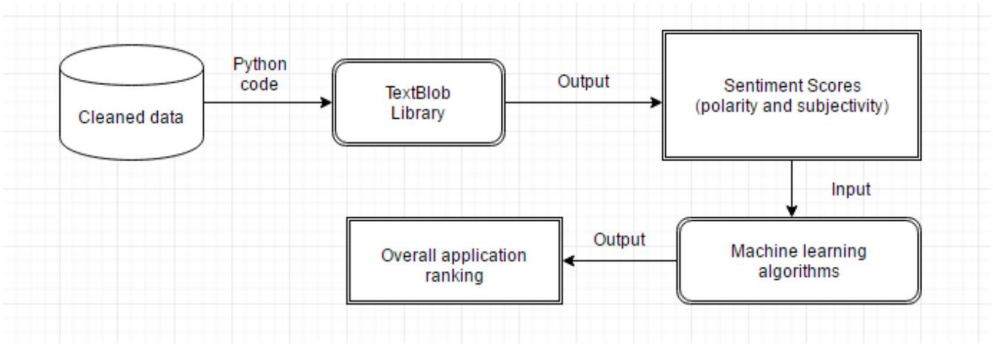
Εικόνα 36. Εισαγωγή λειτουργίας WordCloud και δημιουργία σύννεφου λέξεων

Εικόνα 37. Σύννεφο λέξεων αυτοκινητοβιομηχανίας Audi

Οι λέξεις που εμφανίζονται περισσότερο στα συγκεκριμένα tweets είναι η επωνυμία audi και η λέξη car, κάτι που είναι αναμενόμενο, καθώς τα συγκεκριμένα μηνύματα συλλέχθηκαν με σκοπό την σφυγμομέτρηση των χρηστών σχετικά με την συγκεκριμένη αυτοκινητοβιομηχανία και τα μοντέλα της. Επίσης παρατηρούμε την ύπαρξη πολλών θετικών λέξεων, όπως like, love, thank, best, good, cool, new, nice, want, awesome κ.α, γεγονός που μας οδηγεί σε ένα πρώτο συμπέρασμα σχετικά με το συναίσθημα που εκφράζεται από τους χρήστες για την συγκεκριμένη εταιρεία. Τέλος, παρατηρούμε και λέξεις που αναφέρονται σε συγκεκριμένα μοντέλα αυτοκινήτου της εταιρίας όπως tt, quatro, r8, a3 κ.α.

6.2. **Ανάλυση συναισθήματος**

Για την ανάλυση συναισθήματος χρησιμοποιήσαμε την βιβλιοθήκη TextBlob της Python. Ο αναλυτής του Textblob επιστρέφει δύο ιδιότητες για ένα δεδομένο κείμενο προς ανάλυση, την πολικότητα (polarity) και την υποκειμενικότητα (subjectivity). Η πολικότητα είναι μια βαθμολογία που παίρνει τιμές μεταξύ [-1,1], όπου -1 να δηλώνει αρνητικό συναίσθημα και +1 θετικό συναίσθημα. Η υποκειμενικότητα από την άλλη είναι μια βαθμολογία της αντικειμενικότητας ή υποκειμενικότητας μιας δήλωσης και παίρνει τις σε ένα εύρος [0,1], όπου το 0 σημαίνει ότι η δήλωση είναι αντικειμενική και το 1 ότι η δήλωση περιέχει προσωπική γνώμη, συναίσθημα ή κρίση. Το Textblob αγνοεί τις λέξεις που δεν γνωρίζει και εξετάζει λέξεις και φράσεις στις οποίες μπορεί να εκχωρήσει πολικότητα και μέσους όρους για να εξάγει την τελική βαθμολογία (Εικόνα 38).[147], [171]



Εικόνα 38. Βήματα λειτουργίας της βιβλιοθήκης TextBlob

- Για την προεπεξεργασία των δεδομένων μετατρέψαμε το κάθε tweet σε μία λίστα στοιχείων προκειμένου να εφαρμόσουμε τις διαδικασίες της διακριτοποίησης, της αφαίρεσης ειδικών χαρακτήρων και της στελέχωσης. Σε αυτό το στάδιο και προκειμένου να διενεργήσουμε την ανάλυση συναισθήματος επαναφέραμε τα tweets στην αρχική τους μορφή, δηλαδή σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis) (Εικόνα 39).

In [33]:
```
# Μετατροπή των tweets από μορφή λίστας στοιχείων σε συμβολοσειρά και αποθήκευση στη στήλη Tweet_RFSA

def listToStr(df):

    return ' '.join([str(elem) for elem in df])

df_final['Tweet_RFSA'] = df_final['Tweet_final'].apply(listToStr)

df_final
```

```
<ipython-input-33-677708c78087>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_final['Tweet_RFSA'] = df_final['Tweet_final'].apply(listToStr)
```

Out[33]:

	Tweet	Tweet_semifinal	Tweet_final	Tweet_RFSA
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"	[lol, good, job, new, ad, done, perfectly, right, sochi]	[lol, good, job, new, ad, done, perfectli, right, sochi]	lol good job new ad done perfectli right sochi
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry	[concur, want, talk, brands, seriously, i'd, dig, chatting]	[concur, want, talk, brand, serious, i'd, dig, chat]	concur want talk brand serious i'd dig chat
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.	[add, kindle, book, reader, read, driving]	[add, kindl, book, reader, read, drive]	add kindl book reader read drive
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...	[hope, build, settle, rs6avant, want, loan, one, ...]	[hope, build, settl, rs6avant, want, loan, one, ...]	hope build settl rs6avant want loan one ...
4	@yansarazin @adammcnamara well we know that won't happen @audi	[well, know, happen]	[well, know, happen]	well know happen
...	...	...	...	...
20536	get out of the whip and into the chopperðŸ˜ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc	[get, whip, chopperðÿ, thatlife, money, power, respect]	[get, whip, chopperðÿ, thatlif, money, power, respect]	get whip chopperðÿ thatlif money power respect
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?	[got, two, half, year, old, audi, a6, avant, line, love, mpg, expect]	[got, two, half, year, old, audi, a6, avant, line, love, mpg, expect]	got two half year old audi a6 avant line love mpg expect
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy	[audi, r8, custom, painted, calipers, _porn]	[audi, r8, custom, paint, calip, _porn]	audi r8 custom paint calip _porn
20539	@audi good cars	[good, cars]	[good, car]	good car
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?	[drove, pretoria, mmmh, one, day, one, day, lol, audi, model]	[drove, pretoria, mmmh, one, day, one, day, lol, audi, model]	drove pretoria mmmh one day one day lol audi model

20541 rows × 4 columns

Εικόνα 39. Μετατροπή των tweets από λίστα στοιχείων σε μορφή συμβολοσειρά (Tweet_RFSA)

- Αποθηκεύσαμε σε νέο dataframe με το όνομα **df_analysis** τις στήλες ***Tweet*** και ***Tweet_RFSA,*** ώστε να το χρησιμοποιήσουμε στην ανάλυση συναισθήματος. (Εικόνα 40).

```
In [34]: # Επιλογή των στηλών Tweet και Tweet_RFSA και αποθήκευση ως df_analysis

df_analysis = df_final[['Tweet','Tweet_RFSA']]
df_analysis
```

Out[34]:

	Tweet	Tweet_RFSA
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"	lol good job new ad done perfectli right sochi
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry	concur want talk brand serious i'd dig chat
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.	add kindl book reader read drive
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...	hope build settl rs6avant want loan one ...
4	@yansarazin @adammcnamara well we know that won't happen @audi	well know happen
...	...	...
20536	get out of the whip and into the chopperðÿ˜ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc	get whip chopperðÿ thatlif money power respect
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?	got two half year old audi a6 avant line love mpg expect
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy	audi r8 custom paint calip _porn
20539	@audi good cars	good car
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?	drove pretoria mmmh one day one day lol audi model

20541 rows × 2 columns

Εικόνα 40. Δημιουργία πίνακα με τις στήλες Tweet και Tweet_RFSA

- Εισαγάγαμε την λειτουργία TextBlob από την βιβλιοθήκη textblob (Εικόνα 41).

```
In [35]: # Εισαγωγή της βιβλιοθήκης TextBlob

from textblob import TextBlob
```

Εικόνα 41. Εισαγωγή της λειτουργίας Textblob

- Στη συνέχεια δημιουργήσαμε τις συναρτήσεις getPolarity και getSubjectivity με τις οποίες στη συνέχεια θα υπολογίσουμε την πολικότητα και την υποκειμενικότητα (Εικόνα 42).

```
In [36]: # Δημιουργία συνάρτησηςη για λήψη της πολικότητας (polarity)
def getPolarity(text):
    return TextBlob(text).sentiment.polarity

# Δημιουργία συνάρτησηςη για λήψη της υποκειμενικότητα (subjectivity)
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity
```

Εικόνα 42. Δημιουργία συναρτήσεων για τον υπολογισμό της πολικότητα και υποκειμενικότητας των tweets

- Εφαρμόσαμε τις συναρτήσεις τόσο στα αρχικά tweets της στήλης Tweet, όσο και στα επεξεργασμένα της στήλης Tweet_RFSA, ώστε να υπολογίσουμε την πολικότητα και την υποκειμενικότητα του κάθε tweet. Τα αποτελέσματα αποθηκεύτηκαν στις στήλες Polarity_Tweet, Subjectivity_Tweet και Polarity_Tweet_RFSA, Subjectivity_Tweet_RFSA (Εικόνα 43).

In [37]:
```
# Εφαρμογή συναρτήσεων getPolarity και getSubjectivity στα αρχικά tweets της στήλης Tweet
# και αποθήκευση αποτελεσμάτων ανάλυσης στις στήλες Polarity_Tweet και Subjectivity_Tweet

df_analysis['Polarity_Tweet'] = df_analysis['Tweet'].apply(getPolarity)
df_analysis['Subjectivity_Tweet'] = df_analysis['Tweet'].apply(getSubjectivity)


# Εφαρμογή συναρτήσεων getPolarity και getSubjectivity στα επεξεργασμένα tweets της στήλης Tweet_RFSA (Ready For Sentiment Analysis)
# και αποθήκευση αποτελεσμάτων ανάλυσης στις στήλες Polarity_Tweet_RFSA και Subjectivity_Tweet_RFSA

df_analysis['Polarity_Tweet_RFSA'] = df_analysis['Tweet_RFSA'].apply(getPolarity)
df_analysis['Subjectivity_Tweet_RFSA'] = df_analysis['Tweet_RFSA'].apply(getSubjectivity)


# Εμφάνιση αποτελεσμάτων
df_analysis
```

```
<ipython-input-37-30637ce72475>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_analysis['Polarity_Tweet'] = df_analysis['Tweet'].apply(getPolarity)
```

Out[37]:

	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"	lol good job new ad done perfectli right sochi	0.498377	0.572565	0.480519	0.572565
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry	concur want talk brand serious i'd dig chat	-0.333333	0.666667	-0.333333	0.666667
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.	add kindl book reader read drive	0.000000	0.000000	0.000000	0.000000
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...	hope build settl rs6avant want loan one ...	0.000000	0.000000	0.000000	0.000000
4	@yansarazin @adammcnamara well we know that won't happen @audi	well know happen	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...
20536	get out of the whip and into the chopperðÿ˜Ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc	get whip chopperðÿ thatlif money power respect	0.000000	0.000000	0.000000	0.000000
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?	got two half year old audi a6 avant line love mpg expect	0.144444	0.322222	0.144444	0.322222
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy	audi r8 custom paint calip _porn	0.000000	0.000000	0.000000	0.000000
20539	@audi good cars	good car	0.700000	0.600000	0.700000	0.600000
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?	drove pretoria mmmh one day one day lol audi model	0.800000	0.700000	0.800000	0.700000

20541 rows × 6 columns

Εικόνα 43. Εφαρμογή των συναρτήσεων getPolarity και getSubjectivity και εμφάνιση αποτελεσμάτων

- Στη συνέχεια δημιουργήσαμε την συνάρτηση getAnalysis με την οποία θα διακρίνουμε το κάθε tweet σε θετικό, αρνητικό ή ουδέτερο ανάλογα με την τιμή της πολικότητάς του (Εικόνα 44).

In [38]:
```
# Δημιουργία συνάρτησης για ανάλυση θετικού, αρνητικού ή ουδέτερου συναισθήματος

def getAnalysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'
```

Εικόνα 44. Δημιουργία συνάρτησης για καθορισμό θετικού, αρνητικού ή ουδέτερου συναισθήματος

- Εφαρμόσαμε την συνάρτηση getAnalysis στα tweets της στήλης Polarity_Tweet και Polarity_Tweet_RFSA και αποθηκεύσαμε τα αποτελέσματα στις στήλες Analysis_Tweet και Analysis_Tweet_RFSA αντίστοιχα (Εικόνα 45). Η εφαρμογή της συνάρτησης έγινε τόσο για τα αρχικά tweets (Polarity_Tweet) όσο και για τα τελικά (Polarity_Tweet_RFSA) προκειμένου να γίνει η σύγκριση των αποτελεσμάτων της ανάλυσης συναισθήματος πριν και μετά την επεξεργασία των tweets.

In [39]:
```
# Εφαρμογή της συνάρτησης getAnalysis στα tweets της λίστας Polarity_Tweet

df_analysis['Analysis_Tweet'] = df_analysis['Polarity_Tweet'].apply(getAnalysis)
df_analysis


# Εφαρμογή της συνάρτησης getAnalysis στα tweets της λίστας Polarity_Tweet_RFSA

df_analysis['Analysis_Tweet_RFSA'] = df_analysis['Polarity_Tweet_RFSA'].apply(getAnalysis)
df_analysis
```

Out[39]:

	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA	Analysis_Tweet	Analysis_Tweet_RFSA
0	rt @giodelgado: lol good job "@jadenrdn: new @audi ad done perfectly right! #sochi http://t.co/ard3lhxjmu"	lol good job new ad done perfectli right sochi	0.498377	0.572565	0.480519	0.572565	Positive	Positive
1	@startupljackson i don't concur. i want to talk with my brands. seriously. i'd dig chatting w @audi, @factionskis @sonos, @apple, @burberry	concur want talk brand serious i'd dig chat	-0.333333	0.666667	-0.333333	0.666667	Negative	Negative
2	@sonnydickson @audi should add a kindle book reader so you can read while you're driving.	add kindl book reader read drive	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
3	@adammcnamara i hope they build it, but i will settle for an #rs6avant @audi if you want to loan me one...	hope build settl rs6avant want loan one ...	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
4	@yansarazin @adammcnamara well we know that won't happen @audi	well know happen	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
...	...	...	...	...	...	...	...	...
20536	get out of the whip and into the chopperðŸ˜Ž @audi @audiuk @thedailymillion #thatlife #money #power #respect http://t.co/seb2rwvvgc	get whip chopperðÿ thatlif money power respect	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
20537	@audi just got a two and a half year old audi a6 avant s line and love it what mpg can i expect?	got two half year old audi a6 avant line love mpg expect	0.144444	0.322222	0.144444	0.322222	Positive	Positive
20538	audi r8 with custom painted calipers and @vossenwheels @audi @caranddriver @auto_porn http://t.co/vqjmzonrhy	audi r8 custom paint calip _porn	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
20539	@audi good cars	good car	0.700000	0.600000	0.700000	0.600000	Positive	Positive
20540	"@mphoputini: drove an @audi to and from pretoria! mmmh one day is one day lol" which audi model?	drove pretoria mmmh one day one day lol audi model	0.800000	0.700000	0.800000	0.700000	Positive	Positive

20541 rows × 8 columns

Εικόνα 45. Εφαρμογή της συνάρτησης getAnalysis και εμφάνιση αποτελεσμάτων ανάλυσης

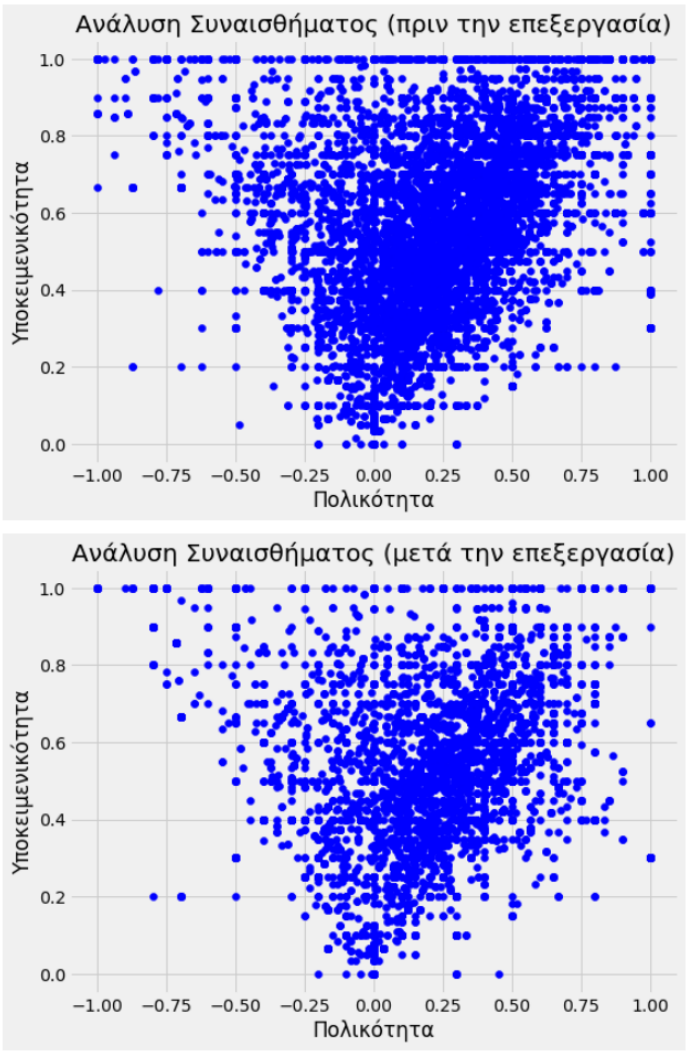
- Ένα διάγραμμα διασποράς είναι από τους πιο παραστατικούς τρόπους ανάδειξης της θετικότητας, αρνητικότητας ή ουδετερότητας ενός δείγματος tweets. Γι' αυτό το λόγο προχωρήσαμε στη σχεδίαση διαγραμμάτων διασποράς της πολικότητας συναρτήσει της υποκειμενικότητας για το σύνολο των tweets. Σε προηγούμενο βήμα είχαμε εισαγάγει την βιβλιοθήκη matplotlib που μας επιτρέπει την δημιουργία τέτοιων διαγραμμάτων. Η εντολή plt.figure δημιουργεί το αντίστοιχο γράφημα, ενώ η εντολή plt.scatter εξάγει το διάγραμμα διασποράς παίρνοντας τις τιμές της πολικότητας και της υποκειμενικότητας από τις στήλες Polariry_Tweet και Subjectivity_Tweet (Εικόνα 46).

In [40]:
```
# Σχεδίαση διαγράμματος διασποράς πολικότητας - υποκειμενικότητας για τα αρχικά tweets

plt.figure(figsize=(8,6))
for i in range(0, df_analysis.shape[0]):
    plt.scatter(df_analysis['Polarity_Tweet'][i], df_analysis['Subjectivity_Tweet'][i], color = 'Blue')

plt.title('Sentiment Analysis (Before)')
plt.xlabel('Polarity')
plt.ylabel('Subjectivity')

# Σχεδίαση διαγράμματος διασποράς πολικότητας - υποκειμενικότητας για τα επεξεργασμένα tweets

plt.figure(figsize=(8,6))
for i in range(0, df_analysis.shape[0]):
    plt.scatter(df_analysis['Polarity_Tweet_RFSA'][i], df_analysis['Subjectivity_Tweet_RFSA'][i], color = 'Blue')

plt.title('Sentiment Analysis (After)')
plt.xlabel('Polarity')
plt.ylabel('Subjectivity')
```

Out[40]:
```
Text(0, 0.5, 'Subjectivity')
```

Εικόνα 46. Σχεδίαση διαγράμματος διασποράς πολικότητας - υποκειμενικότητας

Κι εδώ εξαγάγαμε δύο διαγράμματα για πριν και μετά την επεξεργασία των δεδομένων (Εικόνα 47). Όπως παρατηρούμε στο αρχικό γράφημα το πλήθος των κουκίδων είναι μεγαλύτερο, καθώς το δείγμα περιλαμβάνει όλα τα αρχικά tweets και retweets χωρίς την αφαίρεση των διπλοεγγραφών. Ωστόσο, μπορούμε να διακρίνουμε και στα δύο διαγράμματα ότι το μεγαλύτερο πλήθος των κουκκίδων είναι μετατοπισμένο προς το θετικό τμήμα του άξονα της πολικότητας και προς το αντίστοιχο πάνω μέρος του άξονα της υποκειμενικότητας. Μπορούμε να συμπεράνουμε λοιπόν ότι τα περισσότερα tweets εκφράζουν θετική γνώμη για την αυτοκινητοβιομηχανία Audi, ενώ όπως ήταν αναμενόμενο το μεγαλύτερο μέρος αυτών εμπεριέχουν υψηλή υποκειμενικότητα καθώς εκφράζουν προσωπικές γνώμες και απόψεις. Ωστόσο, αξίζει να σημειώσουμε ότι στο δεύτερο διάγραμμα το πλήθος των κουκίδων που έχουν υψηλή υποκειμενικότητα (τιμή 1.0) έχει μειωθεί σημαντικά σε σχέση με το πρώτο διάγραμμα, γεγονός που οφείλεται τόσο στην απομάκρυνση διπλότυπων εγγραφών, όσο και στην αφαίρεση των προσωπικών αντωνυμιών I και we, μέσω της διαδικασίας της εκκαθάρισης των λέξεων διακοπής. Παρόμοια εικόνα βλέπουμε και στην περίπτωση των πολύ θετικών μηνυμάτων (πολικότητα με τιμή 1.00) που στο πρώτο διάγραμμα είναι περισσότερα απ’ ότι στο δεύτερο. Η αλλαγή αυτή πιθανότατα οφείλεται μόνο στην αφαίρεση των διπλότυπων εγγραφών.



Εικόνα 47. Διάγραμμα διασποράς πολικότητας – υποκειμενικότητας των tweets για την Audi

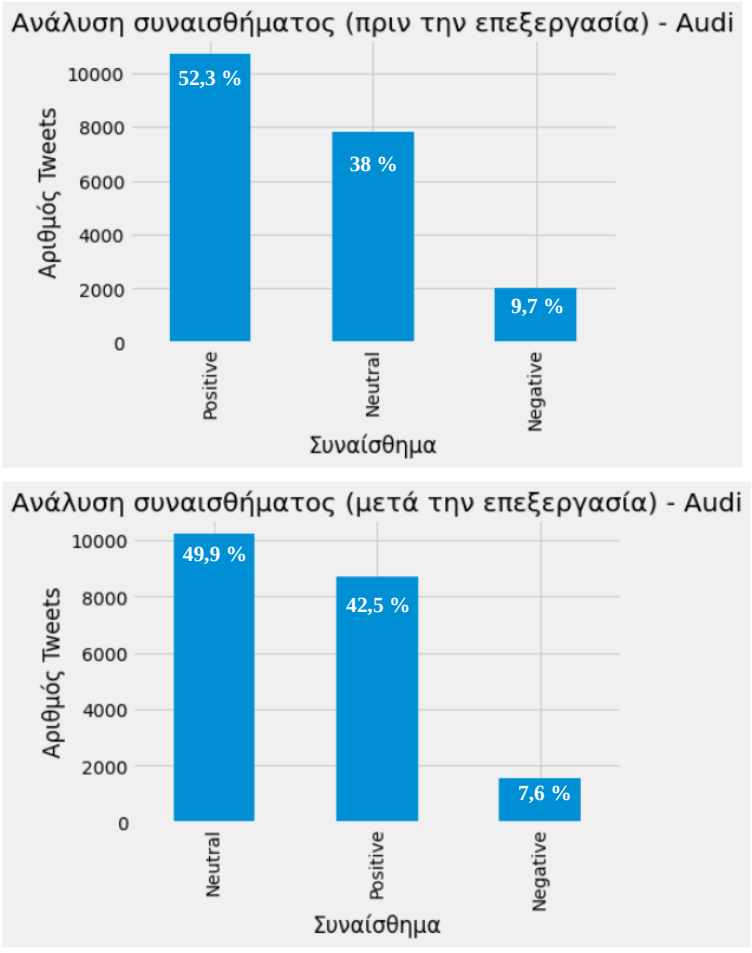
- Ο υπολογισμός των θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία μας έδωσε τα παρακάτω αποτελέσματα (Εικόνα 48), τα οποία στη συνέχεια παρουσιάζονται στα ραβδογράμματα της Εικόνας 49 .

```
In [41]: # Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

         df_analysis['Analysis_Tweet'].value_counts()

Out[41]: Positive    10734
         Neutral      7808
         Negative     1999
         Name: Analysis_Tweet, dtype: int64

In [42]: # Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

         df_analysis['Analysis_Tweet_RFSA'].value_counts()

Out[42]: Neutral     10248
         Positive     8723
         Negative     1570
         Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 48. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (με στελέχωση)(Audi)

Από την πρώτη ανάγνωση του αριθμού των tweets για κάθε συναίσθημα παρατηρούμε μια μεταστροφή στον αριθμό των θετικών και ουδέτερων tweets, ενώ υπάρχει και μια σημαντική μείωση στον αριθμό των αρνητικών μηνυμάτων.

```
In [50]: # Απεικόνιση αποτελεσμάτων ανάλυσης συναισθήματος αρχικών tweets (Analysis_Tweet)

plt.title('Ανάλυση συναισθήματος (πριν την επεξεργασία) - Audi')
plt.xlabel('Συναίσθημα')
plt.ylabel('Αριθμός Tweets')
df_analysis['Analysis_Tweet'].value_counts().plot(kind='bar')
plt.show()


# Απεικόνιση αποτελεσμάτων ανάλυσης συναισθήματος επεξεργασμένων tweets (Analysis_Tweet_RFSA)

plt.title('Ανάλυση συναισθήματος (μετά την επεξεργασία) - Audi')
plt.xlabel('Συναίσθημα')
plt.ylabel('Αριθμός Tweets')
df_analysis['Analysis_Tweet_RFSA'].value_counts().plot(kind='bar')
plt.show()
```



Εικόνα 49. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Audi (με στελέχωση)

- Στη συνέχεια υπολογίσαμε το ποσοστό των θετικών, αρνητικών και ουδέτερων tweets του δείγματος (Εικόνα 50) και τα συγκεντρωτικά αποτελέσματα για πριν και μετά την επεξεργασία των tweets δίνονται στον πίνακα της Εικόνας

```
In [50]: # Υπολογισμός ποσοστού θετικών tweets από τα αρχικά δεδομένα (Analysis_Tweet)

         b_ptweets = df_analysis[df_analysis.Analysis_Tweet == 'Positive']
         b_ptweets = b_ptweets['Tweet']

         b_pos = round((b_ptweets.shape[0] / df_analysis.shape[0])*100, 1)
         b_pos

Out[50]: 52.3

In [51]: # Υπολογισμός ποσοστού αρνητικών tweets από τα αρχικά δεδομένα (Analysis_Tweet)

         b_ntweets = df_analysis[df_analysis.Analysis_Tweet == 'Negative']
         b_ntweets = b_ntweets['Tweet']

         b_n = round((b_ntweets.shape[0] / df_analysis.shape[0])*100, 1)
         b_n

Out[51]: 9.7

In [52]: # Υπολογισμός ποσοστού ουδέτερων tweets από τα αρχικά δεδομένα (Analysis_Tweet)

         b_neutweets = df_analysis[df_analysis.Analysis_Tweet == 'Neutral']
         b_neutweets = b_neutweets['Tweet']

         b_neu = round((b_neutweets.shape[0] / df_analysis.shape[0])*100, 1)
         b_neu

Out[52]: 38.0

In [53]: # Υπολογισμός ποσοστού θετικών tweets από τα επεξεργασμένα δεδομένα (Analysis_Tweet_RFSA)

         a_ptweets = df_analysis[df_analysis.Analysis_Tweet_RFSA == 'Positive']
         a_ptweets = a_ptweets['Tweet_RFSA']

         a_pos = round((a_ptweets.shape[0] / df_analysis.shape[0])*100, 1)
         a_pos

Out[53]: 42.5

In [54]: # Υπολογισμός ποσοστού αρνητικών tweets από τα επεξεργασμένα δεδομένα (Analysis_Tweet_RFSA)

         a_ntweets = df_analysis[df_analysis.Analysis_Tweet_RFSA == 'Negative']
         a_ntweets = a_ntweets['Tweet_RFSA']

         a_n = round((a_ntweets.shape[0] / df_analysis.shape[0])*100, 1)
         a_n

Out[54]: 7.6

In [55]: # Υπολογισμός ποσοστού ουδέτερων tweets από τα επεξεργασμένα δεδομένα (Analysis_Tweet_RFSA)

         a_neutweets = df_analysis[df_analysis.Analysis_Tweet_RFSA == 'Neutral']
         a_neutweets = a_neutweets['Tweet_RFSA']

         a_neu = round((a_neutweets.shape[0] / df_analysis.shape[0])*100, 1)
         a_neu

Out[55]: 49.9
```

Εικόνα 50. Υπολογισμός ποσοστών θετικών, αρνητικών και ουδέτερων tweets αρχικών και επεξεργασμένων tweets

```
# Δημιουργία πίνακα αποτελεσμάτων

data = [{'Ποσοστό Θετικών Tweets': b_pos, 'Ποσοστό Ουδέτερων Tweets': b_neu,
         'Ποσοστό Αρνητικών Tweets': b_n, 'Σύνολο': b_pos+b_n+b_neu},
        {'Ποσοστό Θετικών Tweets': a_pos, 'Ποσοστό Ουδέτερων Tweets': a_neu,
         'Ποσοστό Αρνητικών Tweets': a_n, 'Σύνολο': a_pos+a_n+a_neu}]

df_table = pd.DataFrame(data, index=['Πολικότητα (πριν την επεξεργασία)',
                                     'Πολικότητα (μετά την επεξεργασία)'])
df_table
```

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	52.3	38.0	9.7	100.0
Πολικότητα (μετά την επεξεργασία)	42.5	49.9	7.6	100.0

Εικόνα 51 - Πίνακας 9. Δημιουργία πίνακα ποσοστών θετικών, ουδέτερων και αρνητικών tweets (με στελέχωση) (Audi)

Όπως παρατηρούμε στον παραπάνω πίνακα (Εικόνα 51), η ανάλυση συναισθήματος στα αρχικά δεδομένα έδωσε υψηλότερο ποσοστό θετικών (52.3%) απ' ότι ουδέτερων tweets (38%), ενώ όσον αφορά τα επεξεργασμένα δεδομένα το ποσοστό των ουδέτερων tweets (49.9%) είναι υψηλότερο σε σχέση με τα θετικά (42.5%). Η μεταστροφή αυτή στα ποσοστά παρατηρήθηκε στην περίπτωση που κατά την διάρκεια της επεξεργασίας εφαρμόσαμε την διαδικασία της στελέχωσης και πιθανότατα οφείλεται στην αναντιστοιχία των λέξεων που απαρτίζουν τα tweets μετά την στελέχωση και των εκφράσεων/λέξεων του λεξικού που χρησιμοποιεί η βιβλιοθήκη TextBlob για να εξάγει την πολικότητα και την αντικειμενικότητα.

➢ *Ανάλυση συναισθήματος **χωρίς την διαδικασία στελέχωσης***

Στη συνέχεια, παραλείποντας την διαδικασία της στελέχωσης, αλλά εφαρμόζοντας όλες τις υπόλοιπες διαδικασίες δεν παρατηρήθηκε κάποια αλλαγή μεταξύ θετικών και ουδέτερων tweets πριν και μετά την επεξεργασία, όπως προκύπτει από τους υπολογισμούς (Εικόνα 52) και τα αντίστοιχα ραβδογράμματα (Εικόνα 53), αλλά και από τον πίνακα των ποσοστών θετικών, αρνητικών και ουδέτερων tweets (Εικόνα 54).

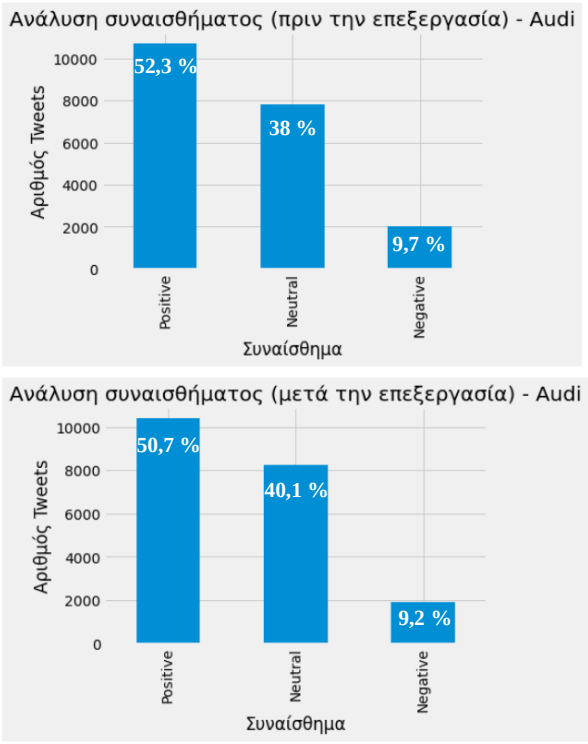
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Positive    10734
Neutral      7808
Negative     1999
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Positive    10418
Neutral      8231
Negative     1892
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 52. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (χωρίς στελέχωση) (Audi)



Εικόνα 53. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Audi (χωρίς στελέχωση)

```
# Δημιουργία πίνακα αποτελεσμάτων

data = [{'Ποσοστό Θετικών Tweets': b_pos, 'Ποσοστό Αρνητικών Tweets': b_n,
         'Ποσοστό Ουδέτερων Tweets': b_neu,'Σύνολο': b_pos+b_n+b_neu},
        {'Ποσοστό Θετικών Tweets': a_pos, 'Ποσοστό Αρνητικών Tweets': a_n,
         'Ποσοστό Ουδέτερων Tweets': a_neu,'Σύνολο': a_pos+a_n+a_neu}]

df_table = pd.DataFrame(data, index=['Πολικότητα (πριν την επεξεργασία)',
                                     'Πολικότητα (μετά την επεξεργασία)'])
df_table
```

	Ποσοστό Θετικών Tweets	Ποσοστό Αρνητικών Tweets	Ποσοστό Ουδέτερων Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	52.3	9.7	38.0	100.0
Πολικότητα (μετά την επεξεργασία)	50.7	9.2	40.1	100.0

Εικόνα 54 - Πίνακας 10. Δημιουργία πίνακα ποσοστών θετικών, ουδέτερων και αρνητικών tweets (χωρίς στελέχωση) (Audi)

❖ **Ανάλυση συναισθήματος των tweets των αυτοκινητοβιομηχανιών Chevrolet, Chrysler, KIA και Volkswagen**

Στη συνέχεια της μελέτης μας προχωρήσαμε στην εφαρμογή των παραπάνω βημάτων, της εξόρυξης και εκκαθάρισης των δεδομένων, της προεπεξεργασίας και της ανάλυσης συναισθήματος και για τις αυτοκινητοβιομηχανίες Chevrolet, Chrysler, KIA και Volkswagen. Η ανάλυση συναισθήματος των tweets πραγματοποιήθηκε αρχικά ενσωματώνοντας την διαδικασία της στελέχωσης στον αλγόριθμο και αφαιρώντας τον στη συνέχεια, ώστε να διαπιστώσουμε την επιρροή του στο τελικό αποτέλεσμα.

- **Chevrolet**

➢ *Εκκαθάριση δεδομένων*

Η βάση δεδομένων των tweets όπως Chevrolet αποτελούνταν αρχικά από 798 εγγραφές (Εικόνα 55). Μετά την αφαίρεση των διπλών εγγραφών προέκυψαν 766 εγγραφές τις οποίες έγινε ο καθαρισμός από τις αναφορές (@), τα σύμβολα hastag (#), τα προθέματα rt των retweets και τους υπερσυνδέσμους (Εικόνα 56).

```
# Εμφάνιση στήλης Tweet αρχικής βάσης δεδομένων
df
```

	Tweet
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv
2	@autoshowchevy #chevrolet free headphones please
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!
4	@autoshowchevy #chevrolet i want free headphones
...	...
793	only now! #incredible #chevrolet corvette 50th anniversary 2003 only for $31,500.00 http://t.co/iudcrawkgy
794	@autoshowchevy - i am at the auto show! #chevrolet
795	so many chevy trucks. #chevrolet #chevy #elcamino #truckyeah #truck http://t.co/hw9jqes0xd
796	@autoshowchevy i wanna take a stingray hwy1 #chevrolet
797	~ car wash ~ #carwash #washing #chevrolet #prisma #neuquen #instalavado http://t.co/21i0ycdxjk

798 rows × 1 columns

Εικόνα 55. Αρχική μορφή των tweets όπως αυτοκινητοβιομηχανίας Chevrolet

```
# Εμφάνιση βάσης δεδομένων μετά την αφαίρεση των διπλότυπων εγγραφών και την εκκαθάριση των δεδομένων
df
```

	Tweet	Tweet_clean
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	"this musclecar is a 1969 chevrolet impala 2dr ht. would you exchange yours for this?
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	own a 2013 chevrolet tahoe lt1 for an affordable price! contact us now!
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphones please
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolutely loving the show cars, eyeing the test drives!
4	@autoshowchevy #chevrolet i want free headphones	chevrolet i want free headphones
...	...	...
761	only now! #incredible #chevrolet corvette 50th anniversary 2003 only for $31,500.00 http://t.co/iudcrawkgy	only now! incredible chevrolet corvette 50th anniversary 2003 only for $31,500.00
762	@autoshowchevy - i am at the auto show! #chevrolet	- i am at the auto show! chevrolet
763	so many chevy trucks. #chevrolet #chevy #elcamino #truckyeah #truck http://t.co/hw9jqes0xd	so many chevy trucks. chevrolet chevy elcamino truckyeah truck
764	@autoshowchevy i wanna take a stingray hwy1 #chevrolet	i wanna take a stingray hwy1 chevrolet
765	~ car wash ~ #carwash #washing #chevrolet #prisma #neuquen #instalavado http://t.co/21i0ycdxjk	~ car wash ~ carwash washing chevrolet prisma neuquen instalavado

766 rows × 2 columns

Εικόνα 56. Μορφή των tweets πριν (Tweet) και μετά (Tweet_clean) την εκκαθάριση (Chevrolet)

➢ *Προεπεξεργασία δεδομένων*

Η προεπεξεργασία των δεδομένων περιλαμβάνει τις διαδικασίες της διακριτοποίησης, της αφαίρεσης των λέξεων διακοπής και των ειδικών χαρακτήρων και της στελέχωσης, τα αποτελέσματα των οποίων παρουσιάζονται στη συνέχεια.

✓ *Διακριτοποίηση*

Τα αποτελέσματα της διακριτοποίησης των επεξεργασμένων tweets εμφανίζονται στη στήλη ***Tokens*** (Εικόνα 57). Όπως παρατηρούμε, τα μέρη του λόγου και τα σημεία στίξης που απαρτίζουν τα tweets εμφανίζονται χωρισμένα με κόμμα και έχουν αποθηκευτεί σε μια λίστα.

	Tweet	Tweet_clean	Tokens
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	"this musclecar is a 1969 chevrolet impala 2dr ht. would you exchange yours for this?	['', this, musclecar, is, a, 1969, chevrolet, impala, 2dr, ht, ., would, you, exchange, yours, for, this, ?]
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	own a 2013 chevrolet tahoe lt1 for an affordable price! contact us now!	[own, a, 2013, chevrolet, tahoe, lt1, for, an, affordable, price, !, contact, us, now, !]
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphones please	[chevrolet, free, headphones, please]
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolutely loving the show cars, eyeing the test drives!	[chevrolet, absolutely, loving, the, show, cars, ,, eyeing, the, test, drives, !]
4	@autoshowchevy #chevrolet i want free headphones	chevrolet i want free headphones	[chevrolet, i, want, free, headphones]

Εικόνα 57. Αποτελέσματα διακριτοποίησης (Tokens) (Chevrolet)

✓ *Αφαίρεση λέξεων διακοπής*

Τα αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής αποθηκεύτηκαν στη στήλη ***Tweet_no_sw*** (Εικόνα 58) .

	Tweet	Tweet_clean	Tokens	Tweet_no_sw
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	"this musclecar is a 1969 chevrolet impala 2dr ht. would you exchange yours for this?	['', this, musclecar, is, a, 1969, chevrolet, impala, 2dr, ht, ., would, you, exchange, yours, for, this, ?]	['', musclecar, 1969, chevrolet, impala, 2dr, ht, ., would, exchange, ?]
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	own a 2013 chevrolet tahoe lt1 for an affordable price! contact us now!	[own, a, 2013, chevrolet, tahoe, lt1, for, an, affordable, price, !, contact, us, now, !]	[2013, chevrolet, tahoe, lt1, affordable, price, !, contact, us, !]
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphones please	[chevrolet, free, headphones, please]	[chevrolet, free, headphones, please]
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolutely loving the show cars, eyeing the test drives!	[chevrolet, absolutely, loving, the, show, cars, ,, eyeing, the, test, drives, !]	[chevrolet, absolutely, loving, show, cars, ,, eyeing, test, drives, !]
4	@autoshowchevy #chevrolet i want free headphones	chevrolet i want free headphones	[chevrolet, i, want, free, headphones]	[chevrolet, want, free, headphones]

Εικόνα 58. Αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής (Tweet_no_sw)(Chevrolet)

✓ *Αφαίρεση ειδικών χαρακτήρων*

Τα αποτελέσματα μετά την αφαίρεση των ειδικών χαρακτήρων αποθηκεύτηκαν στη στήλη ***Tweet_semifinal*** (Εικόνα 59).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw	punctuation	digits	Tweet_semifinal
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	"this musclecar is a 1969 chevrolet impala 2dr ht. would you exchange yours for this?	['', this, musclecar, is, a, 1969, chevrolet, impala, 2dr, ht, ., would, you, exchange, yours, for, this, ?]	['', musclecar, 1969, chevrolet, impala, 2dr, ht, ., would, exchange, ?]	[musclecar, 1969, chevrolet, impala, 2dr, ht, would, exchange]	[musclecar, chevrolet, impala, ht, would, exchange]	[musclecar, chevrolet, impala, ht, would, exchange]
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	own a 2013 chevrolet tahoe lt1 for an affordable price! contact us now!	[own, a, 2013, chevrolet, tahoe, lt1, for, an, affordable, price, !, contact, us, now, !]	[2013, chevrolet, tahoe, lt1, affordable, price, !, contact, us, !]	[2013, chevrolet, tahoe, lt1, affordable, price, contact, us]	[chevrolet, tahoe, lt1, affordable, price, contact, us]	[chevrolet, tahoe, lt1, affordable, price, contact, us]
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphones please	[chevrolet, free, headphones, please]	[chevrolet, free, headphones, please]	[chevrolet, free, headphones, please]	[chevrolet, free, headphones, please]	[chevrolet, free, headphones, please]
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolutely loving the show cars, eyeing the test drives!	[chevrolet, absolutely, loving, the, show, cars, ,, eyeing, the, test, drives, !]	[chevrolet, absolutely, loving, show, cars, ,, eyeing, test, drives, !]	[chevrolet, absolutely, loving, show, cars, eyeing, test, drives]	[chevrolet, absolutely, loving, show, cars, eyeing, test, drives]	[chevrolet, absolutely, loving, show, cars, eyeing, test, drives]
4	@autoshowchevy #chevrolet i want free headphones	chevrolet i want free headphones	[chevrolet, i, want, free, headphones]	[chevrolet, want, free, headphones]	[chevrolet, want, free, headphones]	[chevrolet, want, free, headphones]	[chevrolet, want, free, headphones]

Εικόνα 59. Αποτελέσματα μετά την αφαίρεση ειδικών χαρακτήρων (Tweet_semifinal) (Chevrolet)

✓ *Στελέχωση*

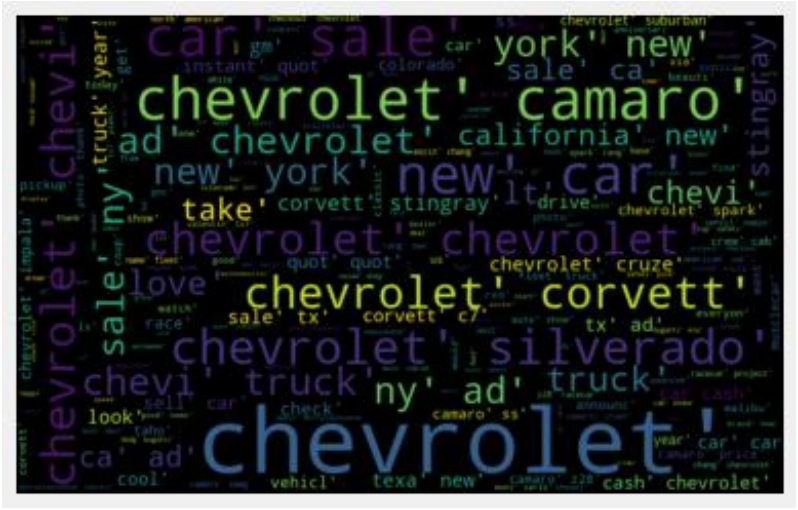
Η διαδικασία της στελέχωσης οδήγησε στην δημιουργία όπως στήλης ***Tweet_final***, τα δεδομένα της οποίας θα χρησιμοποιηθούν στη συνέχεια για την δημιουργία του σύννεφου λέξεων και της ανάλυσης συναισθήματος (Εικόνα 60).

	Tweet	Tweet_semifinal	Tweet_final
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	[musclecar, chevrolet, impala, ht, would, exchange]	[musclecar, chevrolet, impala, ht, would, exchang]
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	[chevrolet, tahoe, lt1, affordable, price, contact, us]	[chevrolet, taho, lt1, afford, price, contact, us]
2	@autoshowchevy #chevrolet free headphones please	[chevrolet, free, headphones, please]	[chevrolet, free, headphon, pleas]
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	[chevrolet, absolutely, loving, show, cars, eyeing, test, drives]	[chevrolet, absolut, love, show, car, eye, test, drive]
4	@autoshowchevy #chevrolet i want free headphones	[chevrolet, want, free, headphones]	[chevrolet, want, free, headphon]

Εικόνα 60. Αποτελέσματα μετά την στελέχωση (Tweet_final) των tweets (Chevrolet)

✓ *Σύννεφο λέξεων*

Στο σύννεφο λέξεων (Εικόνα 61) παρατηρούμε την συχνή επανάληψη και το μεγάλο μέγεθος της επωνυμίας Silverado όπως ήταν αναμενόμενο, καθώς τα tweets αφορούν την συγκριμένη αυτοκινητοβιομηχανία. Οι λέξεις camaro και corvett αναφέρονται στα σπορ μοντέλα Camaro και Corvette της Chevrolet, για τα οποία φαίνεται από τα tweets ότι αποτέλεσαν αντικείμενο έντονου σχολιασμού και συζήτησης μεταξύ των χρηστών. Οι λέξεις car και sale εμφανίζονται συχνά πιθανότατα λόγω της ανακοίνωσης εκπτώσεων εν' όψει της διοργάνωσης του SuperBowl. Τέλος οι λέξεις Silverado και truck, αφορούν το ημιφορτικό Silverado όπως Chevrolet.



Εικόνα 61. Σύννεφο λέξεων αυτοκινητοβιομηχανίας Chevrolet

✓ *Τελική μορφή δεδομένων πριν την ανάλυση συναισθήματος*

Σε αυτό το στάδιο και προκειμένου να διενεργήσουμε την ανάλυση συναισθήματος επαναφέραμε τα tweets από μορφή λίστας σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis) (Εικόνα 62)

	Tweet	Tweet_RFSA
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	musclecar chevrolet impala ht would exchang
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	chevrolet taho lt1 afford price contact us
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphon pleas
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolut love show car eye test drive
4	@autoshowchevy #chevrolet i want free headphones	chevrolet want free headphon
...	...	...
761	only now! #incredible #chevrolet corvette 50th anniversary 2003 only for $31,500.00 http://t.co/iudcrawkgy	incred chevrolet corvett anniversari
762	@autoshowchevy - i am at the auto show! #chevrolet	auto show chevrolet
763	so many chevy trucks. #chevrolet #chevy #elcamino #truckyeah #truck http://t.co/hw9jqes0xd	mani chevi truck chevrolet chevi elcamino truckyeah truck
764	@autoshowchevy i wanna take a stingray hwy1 #chevrolet	wanna take stingray hwi chevrolet
765	~ car wash ~ #carwash #washing #chevrolet #prisma #neuquen #instalavado http://t.co/21i0ycdxjk	car wash carwash wash chevrolet prisma neuquen instalavado

766 rows × 2 columns

Εικόνα 62. Μετατροπή tweets σε μορφή συμβολοσειράς και αποθήκευση σε νέα στήλη (Tweet_RFSA) (Chevrolet)

- *Ανάλυση συναισθήματος*

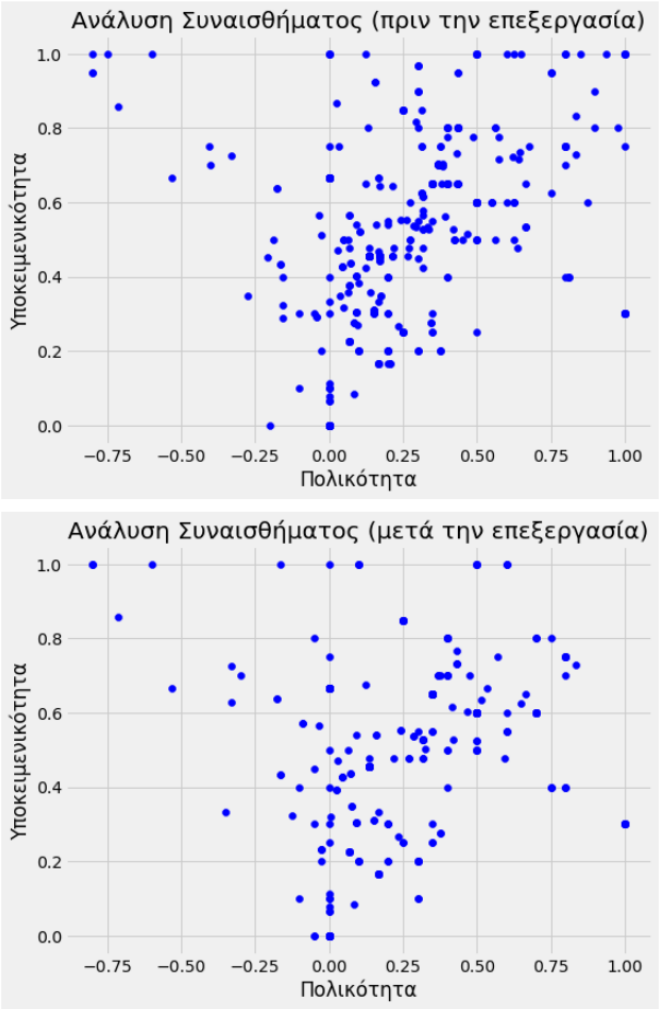
Για την ανάλυση συναισθήματος υπολογίσαμε αρχικά την πολικότητα και την αντικειμενικότητα των tweets πριν την επεξεργασία και μετά από αυτήν, έτσι ώστε να συγκρίνουμε τα αποτελέσματα στη συνέχεια. Οι στήλες Polarity_Tweet, Subjectivity_Tweet και Analysis_Tweet αφορούν τα αρχικά δεδομένα, ενώ οι στήλες Polarity_Tweet_RFSA, Subjectivity_Tweet_RFSA και Analysis_Tweet_RFSA αναφέρονται στα επεξεργασμένα δεδομένα (Εικόνα 63).

	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA	Analysis_Tweet	Analysis_Tweet_RFSA
0	rt @chevroleteurope "this #musclecar is a 1969 #chevrolet #impala 2dr ht. would you exchange yours for this? http://t.co/mef8a6r52k"	musclecar chevrolet impala ht would exchang	0.0000	0.000000	0.0	0.0	Neutral	Neutral
1	own a 2013 #chevrolet tahoe lt1 for an affordable price! contact us now! http://t.co/szqonlwlyv	chevrolet taho lt1 afford price contact us	0.9375	1.000000	0.0	0.0	Positive	Neutral
2	@autoshowchevy #chevrolet free headphones please	chevrolet free headphon pleas	0.4000	0.800000	0.4	0.8	Positive	Positive
3	@autoshowchevy #chevrolet absolutely loving the show cars, eyeing the test drives!	chevrolet absolut love show car eye test drive	0.7500	0.950000	0.5	0.6	Positive	Positive
4	@autoshowchevy #chevrolet i want free headphones	chevrolet want free headphon	0.4000	0.800000	0.4	0.8	Positive	Positive
...	...	...	...	...	...	...	...	...
761	only now! #incredible #chevrolet corvette 50th anniversary 2003 only for $31,500.00 http://t.co/iudcrawkgy	incred chevrolet corvett anniversari	0.3000	0.966667	0.0	0.0	Positive	Neutral
762	@autoshowchevy - i am at the auto show! #chevrolet	auto show chevrolet	0.0000	0.000000	0.0	0.0	Neutral	Neutral
763	so many chevy trucks. #chevrolet #chevy #elcamino #truckyeah #truck http://t.co/hw9jqes0xd	mani chevi truck chevrolet chevi elcamino truckyeah truck	0.5000	0.500000	0.0	0.0	Positive	Neutral
764	@autoshowchevy i wanna take a stingray hwy1 #chevrolet	wanna take stingray hwi chevrolet	0.0000	0.000000	0.0	0.0	Neutral	Neutral
765	~ car wash ~ #carwash #washing #chevrolet #prisma #neuquen #instalavado http://t.co/21i0ycdxjk	car wash carwash wash chevrolet prisma neuquen instalavado	0.0000	0.000000	0.0	0.0	Neutral	Neutral

766 rows × 8 columns

Εικόνα 63. Πολικότητα και υποκειμενικότητα αρχικών και επεξεργασμένων tweets (Chevrolet)

Από τα διαγράμματα διασποράς (Εικόνα 64) παρατηρούμε ότι οι περισσότερες κουκίδες εμφανίζονται στο μέσο και δεξιά μέρος, που σημαίνει ότι τα tweets είναι περισσότερο θετικά για την Chevrolet και δεν έχουν μεγάλη υποκειμενικότητα. Τα αρχικά tweets (πριν την επεξεργασία) φαίνεται να εκφράζουν εντονότερο θετικό συναίσθημα σε σχέση με εκείνα μετά την επεξεργασία. Όπως, ο αριθμός των μηνυμάτων με μεγάλη υποκειμενικότητα, που εμφανίζονται στο πάνω μέρος του πρώτου διαγράμματος έχει μειωθεί σημαντικά, γεγονός που οφείλεται όπως έχουμε αναφέρει στην αφαίρεση των διπλότυπων εγγραφών και των λέξεων διακοπής και συγκεκριμένα των προσωπικών αντωνυμιών πρώτου προσώπου I και we.



Εικόνα 64. Διάγραμμα διασποράς πολικότητας – υποκειμενικότητας των tweets για την αυτοκινητοβιομηχανία Chevrolet

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Positive    412
Neutral     325
Negative     29
Name: Analysis_Tweet, dtype: int64
```

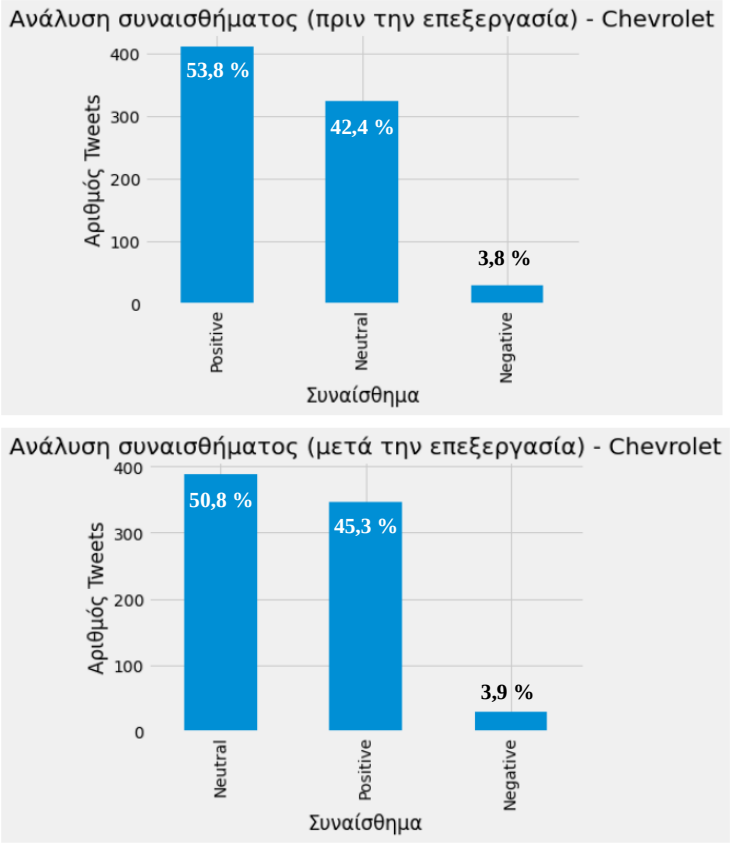
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Neutral     389
Positive    347
Negative     30
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 65. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (με στελέχωση)(Chevrolet)

Η εικόνα του διαγράμματος διασποράς επιβεβαιώνεται από των υπολογισμό του πλήθους των tweets που αντιστοιχούν σε κάθε συναίσθημα (Εικόνα 65). Παρατηρούμε ότι στην αρχική βάση δεδομένων η ανάλυση συναισθήματος έδωσε 412

θετικά, 325 ουδέτερα και 29 αρνητικά tweets. Μετά την επεξεργασία των δεδομένων η εικόνα αυτή μεταστράφηκε, έχοντας πλέον περισσότερα ουδέτερα tweets σε σχέση με τα θετικά. Τα ουδέτερα πλέον είναι 389, τα θετικά 347 και τα αρνητικά 30. Ένας σημαντικός αριθμός δηλαδή των θετικών tweets ανιχνεύτηκε πλέον ως ουδέτερος, ενώ μόνο ένα επιπλέον tweet θεωρήθηκε αρνητικό. Τα παραπάνω αποτελέσματα παρουσιάζονται και στα παρακάτω ραβδογράμματα (Εικόνα 66).



Εικόνα 66. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Chevrolet (με στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	53.8	42.4	3.8	100.0
Πολικότητα (μετά την επεξεργασία)	45.3	50.8	3.9	100.0

Πίνακας 11. Ποσοστό θετικών, ουδέτερων και αρνητικών tweets (με στελέχωση) (Chevrolet)

Στον Πίνακα 11 εμφανίζονται τα αντίστοιχα ποσοστά της πολικότητας των tweets, τα οποία είναι ενδεικτικά για την άποψη που έχουν οι χρήστες για την συγκεκριμένη αυτοκινητοβιομηχανία. Σε μεγάλο βαθμό είναι θετική, ωστόσο η μεταβολή που παρατηρήθηκε δηλώνει μια κάπως συγκεχυμένη γνώμη που θα μπορούσε να έχει δύο ερμηνείες. Αξίζει να σημειώσουμε ωστόσο ότι το ποσοστό των αρνητικών σχολίων είναι πολύ μικρό στο σύνολο του δείγματος.

➢ *Αποτελέσματα ανάλυσης συναισθήματος* ***χωρίς της διαδικασία της στελέχωσης***

Παραλείποντας την διαδικασία της στελέχωσης, δεν παρατηρήθηκε κάποια αλλαγή μεταξύ θετικών και ουδέτερων tweets πριν και μετά την επεξεργασία, όπως προκύπτει από τους υπολογισμούς (Εικόνα 67) και τα αντίστοιχα ραβδογράμματα (Εικόνα 68), αλλά και από τον πίνακα των ποσοστών θετικών, αρνητικών και ουδέτερων tweets (Πίνακας 12).

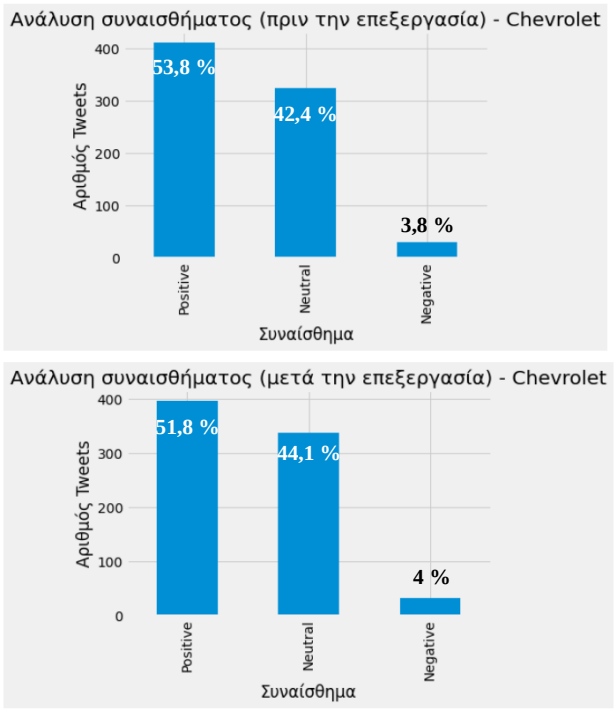
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Positive    412
Neutral     325
Negative     29
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Positive    397
Neutral     338
Negative     31
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 67. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (χωρίς στελέχωση) (Chevrolet)



Εικόνα 68. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Chevrolet (χωρίς στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	53.8	42.4	3.8	100.0
Πολικότητα (μετά την επεξεργασία)	51.8	44.1	4.0	99.9

Πίνακας 12.Δημιουργία πίνακα ποσοστών θετικών, ουδέτερων και αρνητικών tweets (χωρίς στελέχωση) (Chevrolet)

- **Chrysler**

➢ *Εκκαθάριση δεδομένων*

Η βάση δεδομένων των tweets της Chrysler αποτελούνταν αρχικά από 11613 εγγραφές (Εικόνα 69). Μετά την αφαίρεση των διπλών εγγραφών προέκυψαν 9484 εγγραφές στις οποίες έγινε ο καθαρισμός από τις αναφορές (@), τα σύμβολα hastag (#), τα προθέματα rt των retweets και τους υπερσυνδέσμους (Εικόνα 70).

```
# Εμφάνιση στήλης Tweet αρχικής βάσης δεδομένων
df
```

	Tweet
0	#chrysler equity value in 2009: 0.*today* :8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge
2	happy new year for #marchionne: #fiat buys rest of #chrysler for 3.65*b*, *just* 1.75b in cash... much less than analysts estimates
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa
...	...
11608	rt @automotive_news: #chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt httpâ€¦
11609	rt @automotive_news: #chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt httpâ€¦
11610	m.c. of laredo, tx just bought a 2012 #chrysler #200 from a dealer in laredo, tx for $18,995.00!
11611	santa fe @econupdate starts now on @ksfr. tune in! #internationalwomensday #unemployed #subsidies #google #chrysler #bernanke #inequality
11612	santa fe @econupdate starts now on @ksfr. tune in! #internationalwomensday #unemployed #subsidies #google #chrysler #bernanke #inequality

11613 rows × 1 columns

Εικόνα 69. Αρχική μορφή των tweets της αυτοκινητοβιομηχανίας Chrysler

```
# Εμφάνιση βάσης δεδομένων μετά την αφαίρεση των διπλότυπων εγγραφών και την εκκαθάριση των δεδομένων
df
```

	Tweet	Tweet_clean
0	#chrysler equity value in 2009: 0.*today* :8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equity value in 2009: 0.*today* :8.8 billion. big win for uaw members and for american people
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat at last buys chrysler: a new stafor the new year
2	happy new year for #marchionne: #fiat buys rest of #chrysler for 3.65*b*, *just* 1.75b in cash... much less than analysts estimates	happy new year for marchionne: fiat buys rest of chrysler for 3.65*b*, *just* 1.75b in cash... much less than analysts estimates
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us about your chrysler car for a chance to be featured on the forward look blog.
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strikes $4.35 billion deal to buy rest of chrysler
...	...	...
9479	#chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt http://t.co/ykvvageyml	chrysler nears china production deal. renegade and cherokee are prime candidates:
9480	rt @automotive_news: #chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt httpâ€¦	_news: chrysler nears china production deal. renegade and cherokee are prime candidates: httpâ€¦
9481	i want to go from red to black what y'all think #1999 #chrysler #300m #timetochangeitup http://t.co/n8qp7h2w73	i want to go from red to black what y'all think 1999 chrysler 300m timetochangeitup
9482	m.c. of laredo, tx just bought a 2012 #chrysler #200 from a dealer in laredo, tx for $18,995.00!	m.c. of laredo, tx just bought a 2012 chrysler 200 from a dealer in laredo, tx for $18,995.00!
9483	santa fe @econupdate starts now on @ksfr. tune in! #internationalwomensday #unemployed #subsidies #google #chrysler #bernanke #inequality	santa fe starts now on . tune in! internationalwomensday unemployed subsidies google chrysler bernanke inequality

9484 rows × 2 columns

Εικόνα 70. Μορφή των tweets πριν (Tweet) και μετά (Tweet_clean) την εκκαθάριση (Chrysler)

➢ *Προεπεξεργασία δεδομένων*

Η προεπεξεργασία των δεδομένων περιλαμβάνει τις διαδικασίες της διακριτοποίησης, της αφαίρεσης των λέξεων διακοπής και των ειδικών χαρακτήρων και της στελέχωσης, τα αποτελέσματα των οποίων παρουσιάζονται στη συνέχεια.

✓ *Διακριτοποίηση*

Τα αποτελέσματα της διακριτοποίησης των επεξεργασμένων tweets εμφανίζονται στη στήλη ***Tokens*** (Εικόνα 71). Όπως παρατηρούμε, τα μέρη του λόγου και τα

σημεία στίξης που απαρτίζουν τα tweets εμφανίζονται χωρισμένα με κόμμα και έχουν αποθηκευτεί σε μια λίστα.

	Tweet	Tweet_clean	Tokens
0	#chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people	[chrysler, equity, value, in, 2009, :, , 0, . , today, :,, 8.8, billion, ., big, win, for, uaw, members, and, for, american, people]
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat at last buys chrysler: a new stafor the new year	[fiat, at, last, buys, chrysler, :, a, new, stafor, the, new, year]
2	happy new year for #marchionne: #fiat buys rest of #chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	happy new year for marchionne: fiat buys rest of chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	[happy, new, year, for, marchionne, :, fiat, buys, rest, of, chrysler, for, , 3.65, b, , , just,, 1.75, b, in, cash, ..., much, less, than, analysts, estimates]
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us about your chrysler car for a chance to be featured on the forward look blog.	[tell, us, about, your, chrysler, car, for, a, chance, to, be, featured, on, the, forward, look, blog, .]
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strikes $4.35 billion deal to buy rest of chrysler	[fiat, strikes, $, 4.35, billion, deal, to, buy, rest, of, chrysler]

Εικόνα 71. Αποτελέσματα διακριτοποίησης (Tokens)

✓ *Αφαίρεση λέξεων διακοπής*

Τα αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής αποθηκεύτηκαν στη στήλη ***Tweet_no_sw*** (Εικόνα 72).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw
0	#chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people	[chrysler, equity, value, in, 2009, :, , 0, . , today, :,, 8.8, billion, ., big, win, for, uaw, members, and, for, american, people]	[chrysler, equity, value, 2009, :, , 0, . , today, :,, 8.8, billion, ., big, win, uaw, members, american, people]
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat at last buys chrysler: a new stafor the new year	[fiat, at, last, buys, chrysler, :, a, new, stafor, the, new, year]	[fiat, last, buys, chrysler, :, new, stafor, new, year]
2	happy new year for #marchionne: #fiat buys rest of #chrysler for $3.65b$, $just$ 1.75b in cash... much less than analysts estimates	happy new year for marchionne: fiat buys rest of chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	[happy, new, year, for, marchionne, :, fiat, buys, rest, of, chrysler, for, , 3.65, b, , , just,, 1.75, b, in, cash, ..., much, less, than, analysts, estimates]	[happy, new, year, marchionne, :, fiat, buys, rest, chrysler, , 3.65, b, , , 1.75, b, cash, ..., much, less, analysts, estimates]
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us about your chrysler car for a chance to be featured on the forward look blog.	[tell, us, about, your, chrysler, car, for, a, chance, to, be, featured, on, the, forward, look, blog, .]	[tell, us, chrysler, car, chance, featured, forward, look, blog, .]
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strikes $4.35 billion deal to buy rest of chrysler	[fiat, strikes, $, 4.35, billion, deal, to, buy, rest, of, chrysler]	[fiat, strikes, $, 4.35, billion, deal, buy, rest, chrysler]

Εικόνα 72. Αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής (Tweet_no_sw)(Chrysler)

✓ *Αφαίρεση ειδικών χαρακτήρων*

Τα αποτελέσματα μετά την αφαίρεση των ειδικών χαρακτήρων αποθηκεύτηκαν στη στήλη ***Tweet_semifinal*** (Εικόνα 73).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw	punctuation	digits	Tweet_semifinal
0	#chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equity value in 2009: $0.today$: 8.8 billion. big win for uaw members and for american people	[chrysler, equity, value, in, 2009, :, , 0, . , today, :,, 8.8, billion, ., big, win, for, uaw, members, and, for, american, people]	[chrysler, equity, value, 2009, :, , 0, . , today, :,, 8.8, billion, ., big, win, uaw, members, american, people]	[chrysler, equity, value, 2009, 0, today, 8.8, billion, big, win, uaw, members, american, people]	[chrysler, equity, value, today, billion, big, win, uaw, members, american, people]	[chrysler, equity, value, today, billion, big, win, uaw, members, american, people]
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat at last buys chrysler: a new stafor the new year	[fiat, at, last, buys, chrysler, :, a, new, stafor, the, new, year]	[fiat, last, buys, chrysler, :, new, stafor, new, year]	[fiat, last, buys, chrysler, new, stafor, new, year]	[fiat, last, buys, chrysler, new, stafor, new, year]	[fiat, last, buys, chrysler, new, stafor, new, year]
2	happy new year for #marchionne: #fiat buys rest of #chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	happy new year for marchionne: fiat buys rest of chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	[happy, new, year, for, marchionne, :, fiat, buys, rest, of, chrysler, for, , 3.65, b, , , just,, 1.75, b, in, cash, ..., much, less, than, analysts, estimates]	[happy, new, year, marchionne, :, fiat, buys, rest, chrysler, , 3.65, b, , , 1.75, b, cash, ..., much, less, analysts, estimates]	[happy, new, year, marchionne, fiat, buys, rest, chrysler, 3.65, b, 1.75, b, cash, ..., much, less, analysts, estimates]	[happy, new, year, marchionne, fiat, buys, rest, chrysler, b, b, cash, ..., much, less, analysts, estimates]	[happy, new, year, marchionne, fiat, buys, rest, chrysler, cash, ..., much, less, analysts, estimates]
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us about your chrysler car for a chance to be featured on the forward look blog.	[tell, us, about, your, chrysler, car, for, a, chance, to, be, featured, on, the, forward, look, blog, .]	[tell, us, chrysler, car, chance, featured, forward, look, blog, .]	[tell, us, chrysler, car, chance, featured, forward, look, blog]	[tell, us, chrysler, car, chance, featured, forward, look, blog]	[tell, us, chrysler, car, chance, featured, forward, look, blog]
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strikes $4.35 billion deal to buy rest of chrysler	[fiat, strikes, $, 4.35, billion, deal, to, buy, rest, of, chrysler]	[fiat, strikes, $, 4.35, billion, deal, buy, rest, chrysler]	[fiat, strikes, 4.35, billion, deal, buy, rest, chrysler]	[fiat, strikes, billion, deal, buy, rest, chrysler]	[fiat, strikes, billion, deal, buy, rest, chrysler]

Εικόνα 73. Αποτελέσματα μετά την αφαίρεση ειδικών χαρακτήρων (Tweet_semifinal) (Chrysler)

✓ *Στελέχωση*

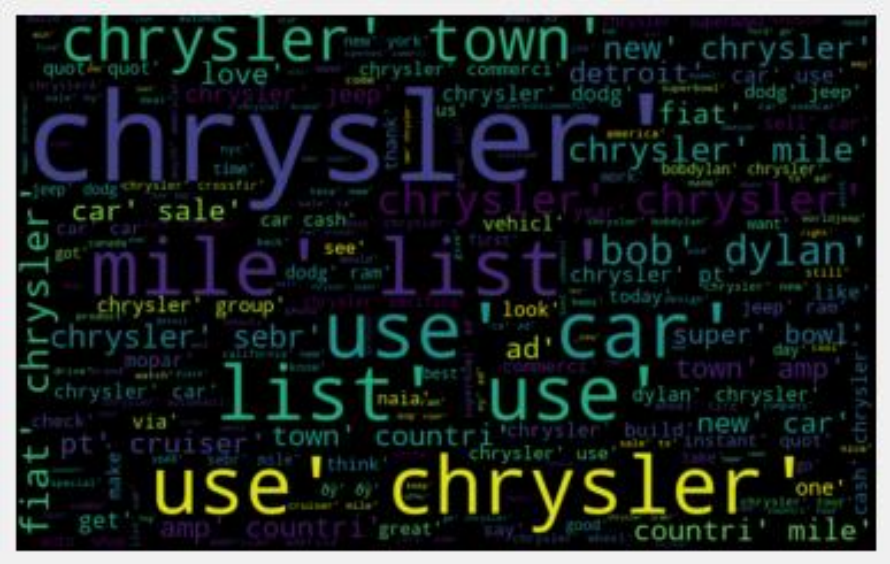
Η διαδικασία της στελέχωσης οδήγησε στην δημιουργία όπως στήλης ***Tweet_final***, τα δεδομένα της οποίας θα χρησιμοποιηθούν στη συνέχεια για την δημιουργία του σύννεφου λέξεων και της ανάλυσης συναισθήματος (Εικόνα 74).

	Tweet	Tweet_semifinal	Tweet_final
0	#chrysler equity value in 2009: $0.today$:8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	[chrysler, equity, value, today, billion, big, win, uaw, members, american, people]	[chrysler, equiti, valu, today, billion, big, win, uaw, member, american, peopl]
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	[fiat, last, buys, chrysler, new, stafor, new, year]	[fiat, last, buy, chrysler, new, stafor, new, year]
2	happy new year for #marchionne: #fiat buys rest of #chrysler for $3.65b$, $just$1.75b in cash... much less than analysts estimates	[happy, new, year, marchionne, fiat, buys, rest, chrysler, cash, ..., much, less, analysts, estimates]	[happi, new, year, marchionn, fiat, buy, rest, chrysler, cash, ..., much, less, analyst, estim]
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	[tell, us, chrysler, car, chance, featured, forward, look, blog]	[tell, us, chrysler, car, chanc, featur, forward, look, blog]
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	[fiat, strikes, billion, deal, buy, rest, chrysler]	[fiat, strike, billion, deal, buy, rest, chrysler]

Εικόνα 74. Αποτελέσματα μετά την στελέχωση (Tweet_final) των tweets (Chrysler)

✓ *Σύννεφο λέξεων*

Στο σύννεφο λέξεων που δημιουργήθηκε (Εικόνα 75) εμφανίζεται συχνά και μεγάλα γράμματα η επωνυμία της Chrysler, οι λέξεις use, list, mile και car, ενώ με μικρότερη γραμματοσειρά εμφανίζονται λέξεις όπως town, country, new, sale κ.α. Η λέξη list προέκυψε μετά την στελέχωση της λέξης listing που χρησιμοποιείται στα tweets και αναφέρεται στην ταξινόμηση των αυτοκινήτων που εισάγει η εταιρεία σε άλλες χώρες εκτός της Αμερικής. Στο σύννεφο εμφανίζεται και το όνομα του Bob Dylan ο οποίος συμμετείχε σε διαφημιστικό της συγκεκριμένης αυτοκινητοβιομηχανίας το οποίο προβλήθηκε κατά την διάρκεια του SuperBowl. Η συχνή χρήση της λέξης mile οφείλεται στην δημοσίευση αγγελιών από τους χρήστες στις οποίες αναφέρουν τα μίλια που έχει διανύσει το αυτοκίνητο που έχουν προς πώληση. Παρά τις πληροφορίες που μας δίνει το συγκεκριμένο σύννεφο σχετικά με όλα τα παραπάνω, δεν είναι δυνατόν να καθοριστεί το συναίσθημα που εκφράζεται από του χρήστες για την συγκεκριμένη αυτοκινητοβιομηχανία.



Εικόνα 75. Σύννεφο λέξεων αυτοκινητοβιομηχανίας Chrysler

✓ *Τελική μορφή δεδομένων πριν την ανάλυση συναισθήματος*

Σε αυτό το στάδιο και προκειμένου να διενεργήσουμε την ανάλυση συναισθήματος επαναφέραμε τα tweets από μορφή λίστας σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis) (Εικόνα 76)

	Tweet	Tweet_RFSA
0	#chrysler equity value in 2009: 0.*today* :8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equiti valu today billion big win uaw member american peopl
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat last buy chrysler new stafor new year
2	happy new year for #marchionne: #fiat buys rest of #chrysler for 3.65*b*, *just*1.75b in cash... much less than analysts estimates	happi new year marchionn fiat buy rest chrysler cash ... much less analyst estim
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us chrysler car chanc featur forward look blog
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strike billion deal buy rest chrysler
...	...	...
9479	#chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt http://t.co/ykvvageyml	chrysler near china product deal renegad cheroke prime candid
9480	rt @automotive_news: #chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt httpâ€¦	_new chrysler near china product deal renegad cheroke prime candid httpâ
9481	i want to go from red to black what y'all think #1999 #chrysler #300m #timetochangeitup http://t.co/n8qp7h2w73	want go red black y'all think chrysler timetochangeitup
9482	m.c. of laredo, tx just bought a 2012 #chrysler #200 from a dealer in laredo, tx for $18,995.00!	laredo tx bought chrysler dealer laredo tx
9483	santa fe @econupdate starts now on @ksfr. tune in! #internationalwomensday #unemployed #subsidies #google #chrysler #bernanke #inequality	santa fe start tune internationalwomensday unemploy subsidi googl chrysler bernank inequ

9484 rows × 2 columns

Εικόνα 76. Μετατροπή tweets σε μορφή συμβολοσειράς και αποθήκευση σε νέα στήλη (Tweet_RFSA) (Chrysler)

➢ *Ανάλυση συναισθήματος*

Ακολουθώντας την ίδια διαδικασία με πριν προέκυψαν οι στήλες Polarity_Tweet, Subjectivity_Tweet και Analysis_Tweet που αφορούν τα αρχικά δεδομένα, ενώ οι στήλες Polarity_Tweet_RFSA, Subjectivity_Tweet_RFSA και Analysis_Tweet_RFSA αναφέρονται στα επεξεργασμένα δεδομένα (Εικόνα 77).

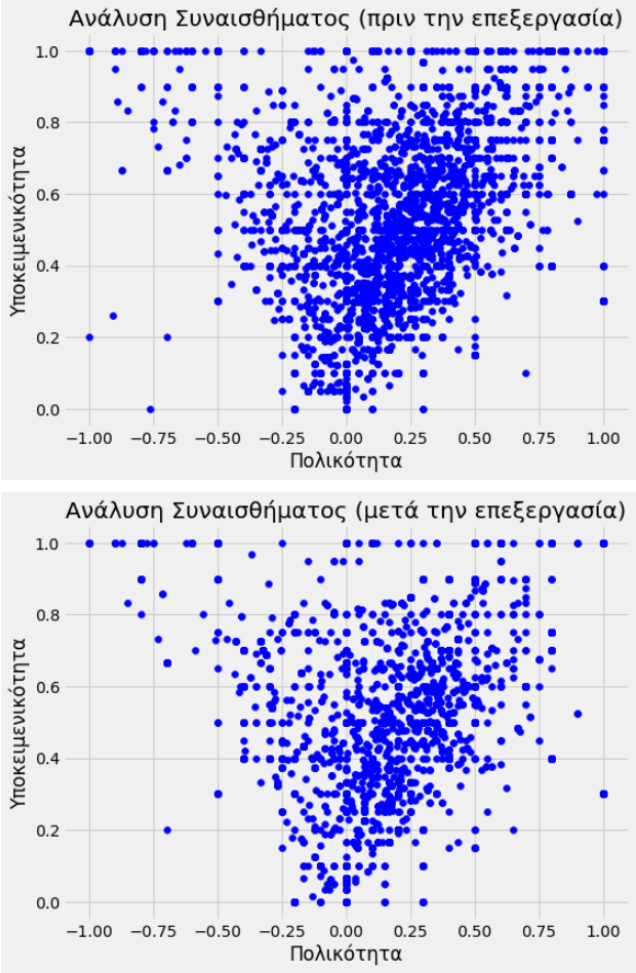
	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA	Analysis_Tweet	Analysis_Tweet_RFSA
0	#chrysler equity value in 2009: 0.*today* :8.8 billion. big win for uaw members and for american people http://t.co/nn7fxwlrt2	chrysler equiti valu today billion big win uaw member american peopl	0.266667	0.166667	0.266667	0.166667	Positive	Positive
1	#fiat at last buys #chrysler: a new start for the new year http://t.co/oordgp32ge	fiat last buy chrysler new stafor new year	0.090909	0.325253	0.090909	0.325253	Positive	Positive
2	happy new year for #marchionne: #fiat buys rest of #chrysler for 3.65*b*, *just*1.75b in cash... much less than analysts estimates	happi new year marchionn fiat buy rest chrysler cash ... much less analyst estim	0.256566	0.507071	-0.015152	0.260606	Positive	Negative
3	tell us about your #chrysler car for a chance to be featured on the forward look blog. http://t.co/f3qpygkzij http://t.co/ybztyaerxy	tell us chrysler car chanc featur forward look blog	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
4	#fiat strikes $4.35 billion deal to buy rest of #chrysler http://t.co/eshslv6iys http://t.co/ixotkjxpoa	fiat strike billion deal buy rest chrysler	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
...	...	...	...	...	...	...	...	...
9479	#chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt http://t.co/ykvvageyml	chrysler near china product deal renegad cheroke prime candid	0.000000	0.000000	0.350000	0.600000	Neutral	Positive
9480	rt @automotive_news: #chrysler nears china @jeep production deal. #renegade and #cherokee are prime candidates: http://t.co/y84weurfpt httpâ€¦	_new chrysler near china product deal renegad cheroke prime candid httpâ	0.000000	0.000000	0.278788	0.551515	Neutral	Positive
9481	i want to go from red to black what y'all think #1999 #chrysler #300m #timetochangeitup http://t.co/n8qp7h2w73	want go red black y'all think chrysler timetochangeitup	-0.083333	0.216667	-0.083333	0.216667	Negative	Negative
9482	m.c. of laredo, tx just bought a 2012 #chrysler #200 from a dealer in laredo, tx for $18,995.00!	laredo tx bought chrysler dealer laredo tx	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
9483	santa fe @econupdate starts now on @ksfr. tune in! #internationalwomensday #unemployed #subsidies #google #chrysler #bernanke #inequality	santa fe start tune internationalwomensday unemploy subsidi googl chrysler bernank inequ	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral

9484 rows × 8 columns

Εικόνα 77. Πολικότητα και υποκειμενικότητα αρχικών και επεξεργασμένων tweets (Chrysler)

Από τα διαγράμματα διασποράς (Εικόνα 78) παρατηρούμε ότι οι περισσότερες κουκίδες εμφανίζονται στο μέσο και δεξιά μέρος, που σημαίνει ότι τα tweets είναι περισσότερο ουδέτερα και θετικά για την Chrysler και δεν έχουν μεγάλη υποκειμενικότητα. Τα αρχικά tweets (πριν την επεξεργασία) φαίνεται να εκφράζουν εντονότερο θετικό συναίσθημα σε σχέση με εκείνα μετά την επεξεργασία. Επίσης κι εδώ, όπως και πριν, ο αριθμός των μηνυμάτων με μεγάλη υποκειμενικότητα, που

εμφανίζονται στο πάνω μέρος του πρώτου διαγράμματος έχει μειωθεί σημαντικά, γεγονός που οφείλεται όπως έχουμε αναφέρει στην αφαίρεση των διπλότυπων εγγραφών και των λέξεων διακοπής και συγκεκριμένα των προσωπικών αντωνυμιών πρώτου προσώπου I και we.



Εικόνα 78. Διάγραμμα διασποράς πολικότητας – υποκειμενικότητας των tweets για την Chrysler

Η εικόνα του διαγράμματος διασποράς επιβεβαιώνεται κι εδώ από των υπολογισμό του πλήθους των tweets που αντιστοιχούν σε κάθε συναίσθημα (Εικόνα 79). Παρατηρούμε ότι στην αρχική βάση δεδομένων η ανάλυση συναισθήματος έδωσε 4529 ουδέτερα, 4056 θετικά και 899 αρνητικά tweets. Η ίδια εικόνα εμφανίζεται και μετά την επεξεργασία των δεδομένων, με 5586 ουδέτερα tweets, 3275 θετικά και 623 αρνητικά. Εμφανίζεται δηλαδή μία σημαντική αύξηση των ουδέτερων, σε βάρος τόσο των θετικών όσο και των αρνητικών tweets. Τα παραπάνω αποτελέσματα παρουσιάζονται και στα παρακάτω ραβδογράμματα (Εικόνα 80).

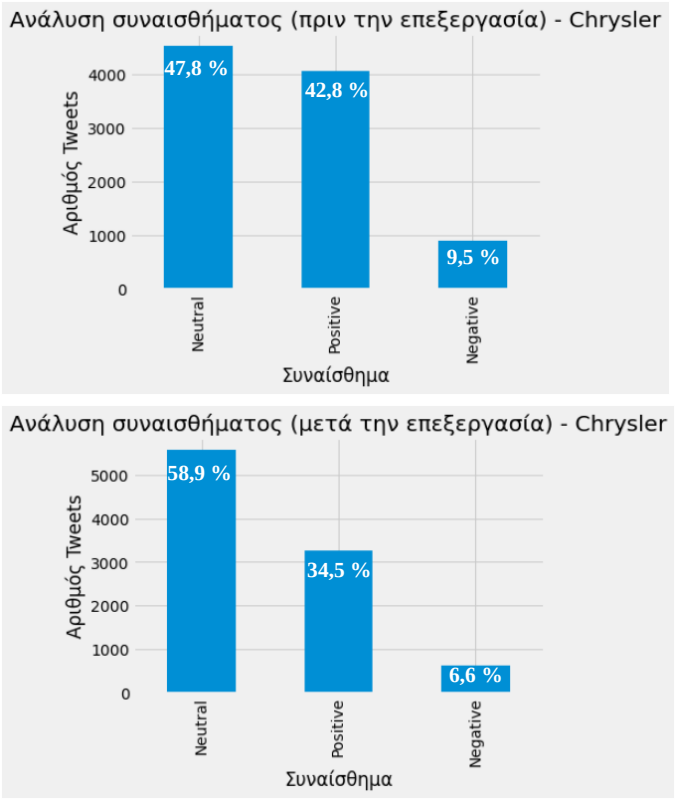
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Neutral     4529
Positive    4056
Negative     899
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Neutral     5586
Positive    3275
Negative     623
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 79. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (με στελέχωση)(Chrysler)



Εικόνα 80. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Chrysler (με στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	42.8	47.8	9.5	100.1
Πολικότητα (μετά την επεξεργασία)	34.5	58.9	6.6	100.0

Πίνακας 13. Πίνακας ποσοστών θετικών, ουδέτερων και αρνητικών tweets (με στελέχωση) (Chrysler)

Στον Πίνακα 13 εμφανίζονται τα αντίστοιχα ποσοστά της πολικότητας των tweets, τα οποία είναι ενδεικτικά για την άποψη που έχουν οι χρήστες για την συγκεκριμένη αυτοκινητοβιομηχανία. Τόσο πριν, όσο και μετά την επεξεργασία τα σχόλια

παραμένουν ουδέτερα, με τα θετικά tweets να είναι εμφανώς λιγότερα. Αξίζει να σημειώσουμε ότι στην περίπτωση της Chrysler παρατηρείτε κι ένα αρκετά υψηλό ποσοστό αρνητικών tweets σε σχέση με τα αντίστοιχα ποσοστά των υπολοίπων αυτοκινητοβιομηχανιών.

- *Αποτελέσματα ανάλυσης συναισθήματος* ***χωρίς της διαδικασία της στελέχωσης***

  Παραλείποντας την διαδικασία της στελέχωσης, δεν παρατηρήθηκε κάποια αλλαγή μεταξύ θετικών και ουδέτερων tweets πριν και μετά την επεξεργασία. Παρατηρούμε, ωστόσο ότι ο αριθμός των ουδέτερων tweets μειώθηκε στα επεξεργασμένα δεδομένα, ενώ παράλληλα είχαμε αύξηση των θετικών. Οι υπολογισμοί του αριθμού των tweets της κάθε κατηγορίας (Εικόνα 81), τα αντίστοιχα ραβδογράμματα (Εικόνα 82), αλλά και ο πίνακας των ποσοστών θετικών, αρνητικών και ουδέτερων tweets (Πίνακας 14) δίνονται παρακάτω.

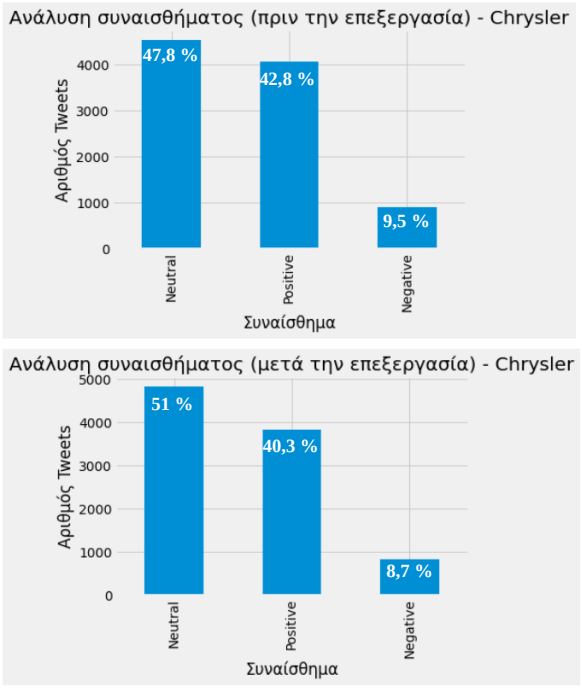
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Neutral     4529
Positive    4056
Negative     899
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Neutral     4837
Positive    3826
Negative     821
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 81. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (χωρίς στελέχωση) (Chrysler)



Εικόνα 82. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Chrysler (χωρίς στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	42.8	47.8	9.5	100.1
Πολικότητα (μετά την επεξεργασία)	40.3	51.0	8.7	100.0

Πίνακας 14. Ποσοστά θετικών, ουδέτερων και αρνητικών tweets (χωρίς στελέχωση) (Chrysler)

- **KIA**

➢ *Εκκαθάριση δεδομένων*

Η βάση δεδομένων των tweets της KIA αποτελούνταν αρχικά από 4362 εγγραφές (Εικόνα 83). Μετά την αφαίρεση των διπλών εγγραφών προέκυψαν 1143 εγγραφές στις οποίες έγινε ο καθαρισμός από τις αναφορές (@), τα σύμβολα hastag (#), τα προθέματα rt των retweets και τους υπερσυνδέσμους (Εικόνα 84).

```
# Εμφάνιση στήλης Tweet αρχικής βάσης δεδομένων
df
```

	Tweet
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m
...	...
4357	it's fine if kia sells the k900 through just a third of its dealers - s it's logical that an... http://t.co/erfpkbgzvm #kiak900
4358	it's fine if kia sells the k900 through just a third of its dealers - s it's logical that an... http://t.co/erfpkbgzvm #kiak900
4359	it's fine if kia sells the k900 through just a third of its dealers - s it's logical that an... http://t.co/erfpkbgzvm #kiak900
4360	it's fine if kia sells the k900 through just a third of its dealers - s it's logical that an... http://t.co/erfpkbgzvm #kiak900
4361	rt @kia: take the red key, and youâ€™ll never look at luxury the same again. #kiak900 http://t.co/sqctmxldi5 http://t.co/uagpeosyk1

4361 rows × 1 columns

Εικόνα 83. Αρχική μορφή των tweets της αυτοκινητοβιομηχανίας KIA

```
# Εμφάνιση βάσης δεδομένων μετά την αφαίρεση των διπλότυπων εγγραφών και την εκκαθάριση των δεδομένων
df
```

	Tweet	Tweet_clean
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	the kiak900 will challenge everything you think about kia. read why via :
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	the kiak900 will challenge everything you think about kia. read why via :
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king of the road. kiak900
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	: classic lines. upstaattitude. kiak900
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	: one reason to be on the nice list. kiak900
...	...	...
1138	watching the new kia k900 commercial makes me want to watch the matrix movies #kiak900	watching the new kia k900 commercial makes me want to watch the matrix movies kiak900
1139	dog, the #kiak900 has reclining backseats. car sex will never be the same.	dog, the kiak900 has reclining backseats. car sex will never be the same.
1140	paid attention to the #kiak900 ad for 1st time & like it...but don't like $64k base price tag.	paid attention to the kiak900 ad for 1st time & like it...but don't like $64k base price tag.
1141	the surround-view monitor: itâ€™s like watching yourself park. #kiak900 http://t.co/2f0rcjsh60	the surround-view monitor: itâ€™s like watching yourself park. kiak900
1142	â€œ@kia: morpheus, levitating cars, sparks andâ€¦opera? watch our #kiak900 big game commercial http://t.co/e1nkm3rcwi http://t.co/jgvsaozj4hâ€¦ ðŸ˜ðŸ˜	â€œ: morpheus, levitating cars, sparks andâ€¦opera? watch our kiak900 big game commercial ðŸ˜ðŸ˜

1143 rows × 2 columns

Εικόνα 84. Μορφή των tweets πριν (Tweet) και μετά (Tweet_clean) την εκκαθάριση

➢ *Προεπεξεργασία δεδομένων*

Η προεπεξεργασία των δεδομένων περιλαμβάνει τις διαδικασίες της διακριτοποίησης, της αφαίρεσης των λέξεων διακοπής και των ειδικών χαρακτήρων και της στελέχωσης, τα αποτελέσματα των οποίων παρουσιάζονται βηματικά στη συνέχεια.

✓ Διακριτοποίηση

Τα αποτελέσματα της διακριτοποίησης των επεξεργασμένων tweets εμφανίζονται στη στήλη ***Tokens*** (Εικόνα 85). Όπως παρατηρούμε, τα μέρη του λόγου και τα σημεία στίξης που απαρτίζουν τα tweets εμφανίζονται χωρισμένα με κόμμα και έχουν αποθηκευτεί σε μια λίστα.

	Tweet	Tweet_clean	Tokens
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king of the road. kiak900	[king, of, the, road, ., kiak, 900]
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	: classic lines. upstaattitude. kiak900	[:, classic, lines, ., upstaattitude, ., kiak, 900]
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	: one reason to be on the nice list. kiak900	[:, one, reason, to, be, on, the, nice, list, ., kiak, 900]

Εικόνα 85. Αποτελέσματα διακριτοποίησης (Tokens)

✓ *Αφαίρεση λέξεων διακοπής*

Τα αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής αποθηκεύτηκαν στη στήλη ***Tweet_no_sw*** (Εικόνα 86).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]	[kiak, 900, challenge, everything, think, kia, ., read, via, :]
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]	[kiak, 900, challenge, everything, think, kia, ., read, via, :]
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king of the road. kiak900	[king, of, the, road, ., kiak, 900]	[king, road, ., kiak, 900]
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	: classic lines. upstaattitude. kiak900	[:, classic, lines, ., upstaattitude, ., kiak, 900]	[:, classic, lines, ., upstaattitude, ., kiak, 900]
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	: one reason to be on the nice list. kiak900	[:, one, reason, to, be, on, the, nice, list, ., kiak, 900]	[:, one, reason, nice, list, ., kiak, 900]

Εικόνα 86. Αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής (Tweet_no_sw)(KIA)

✓ *Αφαίρεση ειδικών χαρακτήρων*

Τα αποτελέσματα μετά την αφαίρεση των ειδικών χαρακτήρων αποθηκεύτηκαν στη στήλη ***Tweet_semifinal*** (Εικόνα 87).

```
# Αποθήκευση των σημείων στίξης στην μεταβλητή punctuations
punctuations = list(string.punctuation)

# Αφαίρεση σημείων στίξης από την στήλη Tweet_no_sw
df_clean['punctuation'] = df_clean['Tweet_no_sw'].apply(lambda x: [i for i in x if i not in punctuations])


# Αφαίρεση ψηφίων (αριθμών) από την στήλη punctuation
df_clean['digits'] = df_clean['punctuation'].apply(lambda x: [i for i in x if i[0] not in list(string.digits)])


# Επιλογή των tweets με αριθμό λέξεων μεγαλύτερο του 1
df_clean['Tweet_semifinal'] = df_clean['digits'].apply(lambda x: [i for i in x if len(i) > 1])

df_clean.head()
```

	Tweet	Tweet_clean	Tokens	Tweet_no_sw	punctuation	digits	Tweet_semifinal
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]	[kiak, 900, challenge, everything, think, kia, ., read, via, :]	[kiak, 900, challenge, everything, think, kia, read, via]	[kiak, challenge, everything, think, kia, read, via]	[kiak, challenge, everything, think, kia, read, via]
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	the kiak900 will challenge everything you think about kia. read why via :	[the, kiak, 900, will, challenge, everything, you, think, about, kia, ., read, why, via, :]	[kiak, 900, challenge, everything, think, kia, ., read, via, :]	[kiak, 900, challenge, everything, think, kia, read, via]	[kiak, challenge, everything, think, kia, read, via]	[kiak, challenge, everything, think, kia, read, via]
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king of the road. kiak900	[king, of, the, road, ., kiak, 900]	[king, road, ., kiak, 900]	[king, road, kiak, 900]	[king, road, kiak]	[king, road, kiak]
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	: classic lines. upstaattitude. kiak900	[:, classic, lines, ., upstaattitude, ., kiak, 900]	[:, classic, lines, ., upstaattitude, ., kiak, 900]	[classic, lines, upstaattitude, kiak, 900]	[classic, lines, upstaattitude, kiak]	[classic, lines, upstaattitude, kiak]
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	: one reason to be on the nice list. kiak900	[:, one, reason, to, be, on, the, nice, list, ., kiak, 900]	[:, one, reason, nice, list, ., kiak, 900]	[one, reason, nice, list, kiak, 900]	[one, reason, nice, list, kiak]	[one, reason, nice, list, kiak]

Εικόνα 87. Αποτελέσματα μετά την αφαίρεση ειδικών χαρακτήρων (Tweet_semifinal) (KIA)

✓ *Στελέχωση*

Η διαδικασία της στελέχωσης οδήγησε στην δημιουργία όπως στήλης ***Tweet_final***, τα δεδομένα της οποίας θα χρησιμοποιηθούν στη συνέχεια για την δημιουργία του σύννεφου λέξεων και της ανάλυσης συναισθήματος (Εικόνα 88).

	Tweet	Tweet_semifinal	Tweet_final
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	[kiak, challenge, everything, think, kia, read, via]	[kiak, challeng, everyth, think, kia, read, via]
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	[kiak, challenge, everything, think, kia, read, via]	[kiak, challeng, everyth, think, kia, read, via]
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	[king, road, kiak]	[king, road, kiak]
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	[classic, lines, upstaattitude, kiak]	[classic, line, upstaattitud, kiak]
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	[one, reason, nice, list, kiak]	[one, reason, nice, list, kiak]

Εικόνα 88. Αποτελέσματα μετά την στελέχωση (Tweet_final) των tweets (KIA)

✓ *Σύννεφο λέξεων*

Στο σύννεφο λέξεων για την αυτοκινητοβιομηχανία KIA (Εικόνα 89) εμφανίζεται συχνά και με μεγάλη γραμματοσειρά η λέξη kiak η οποία όπως παρατηρούμε από τα αρχικά tweets στην Εικόνα 89 αναφέρεται στο πολυτελές μοντέλο της KIA K900. Το K900 είναι μια πολυτελής λιμουζίνα της KIA που παρουσιάστηκε πρώτη φορά στα πλαίσια του SuperBowl το 2014, όπως υποδηλώνουν οι λέξεις new, luxury και reveal που χρησιμοποιούν οι χρήστες στα tweets τους. Συχνά εμφανίζονται και οι λέξεις whole, world και back. Οι πρώτες δύο χρησιμοποιούνται μαζί δηλώνοντας την αίσθηση ενός νέου κόσμου στο χώρο των πολυτελών αυτοκινήτων με την παρουσίαση του συγκεκριμένου μοντέλου. Επιπλέον, η λέξη morpheu αναφέρεται στον χαρακτήρα Μορφέα της ταινίας Matrix, ο οποίος συμμετείχε σε διαφήμιση της εταιρίας KIA κατά τη διάρκεια του SuperBowl του

2014 και προκάλεσε αίσθηση λόγω της μεγάλης εμπορικής επιτυχίας που είχε η συγκεκριμένη ταινία 11 χρόνια πριν.

Εικόνα 89. Σύννεφο λέξεων αυτοκινητοβιομηχανίας KIA

✓ *Τελική μορφή δεδομένων πριν την ανάλυση συναισθήματος*

Σε αυτό το στάδιο και προκειμένου να διενεργήσουμε την ανάλυση συναισθήματος επαναφέραμε τα tweets από μορφή λίστας σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis) (Εικόνα 90).

	Tweet	Tweet_RFSA
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	kiak challeng everyth think kia read via
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	kiak challeng everyth think kia read via
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king road kiak
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	classic line upstaattitud kiak
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	one reason nice list kiak
...	...	...
1138	watching the new kia k900 commercial makes me want to watch the matrix movies #kiak900	watch new kia k900 commerci make want watch matrix movi kiak
1139	dog, the #kiak900 has reclining backseats. car sex will never be the same.	dog kiak reclin backseat car sex never
1140	paid attention to the #kiak900 ad for 1st time & like it...but don't like $64k base price tag.	paid attent kiak ad time like ... like base price tag
1141	the surround-view monitor: itâ€™s like watching yourself park. #kiak900 http://t.co/2f0rcjsh60	surround-view monitor itâ like watch park kiak
1142	â€œ@kia: morpheus, levitating cars, sparks andâ€¦opera? watch our #kiak900 big game commercial http://t.co/e1nkm3rcwi http://t.co/jgvsaozj4hâ€□ ðŸ˜›ðŸ˜›	morpheu levit car spark andâ opera watch kiak big game commerci ðÿ ðÿ

1143 rows × 2 columns

Εικόνα 90. Μετατροπή tweets σε μορφή συμβολοσειράς και αποθήκευση σε νέα στήλη (Tweet_RFSA) (KIA)

➢ *Ανάλυση συναισθήματος*

Ακολουθώντας την ίδια διαδικασία με πριν προέκυψαν οι στήλες Polarity_Tweet, Subjectivity_Tweet και Analysis_Tweet που αφορούν τα αρχικά δεδομένα, ενώ οι

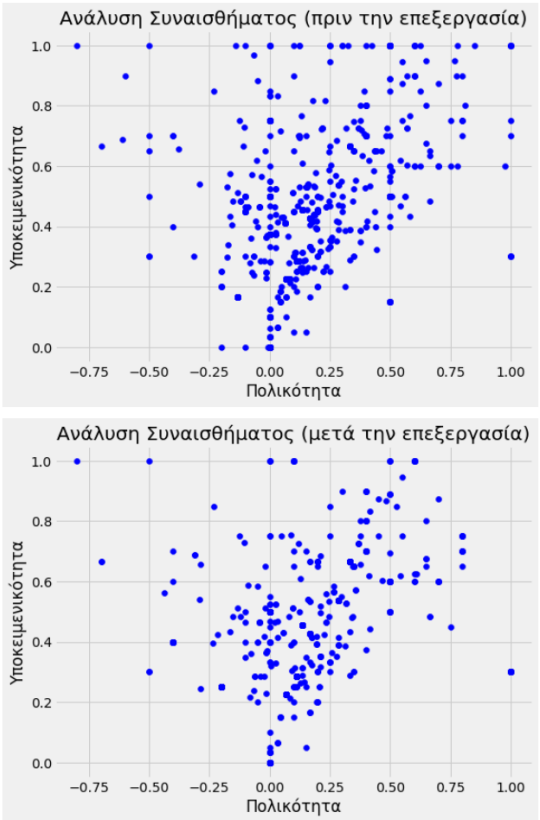
στήλες Polarity_Tweet_RFSA, Subjectivity_Tweet_RFSA και Analysis_Tweet_RFSA αναφέρονται στα επεξεργασμένα δεδομένα (Εικόνα 91).

	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA	Analysis_Tweet	Analysis_Tweet_RFSA
0	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/wcrelskaeu	kiak challeng everyth think kia read via	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
1	the #kiak900 will challenge everything you think about kia. read why via @usatoday: http://t.co/vrzxsfogbr	kiak challeng everyth think kia read via	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
2	king of the road. #kiak900 http://t.co/lu1n5ax8wz	king road kiak	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
3	rt @kia: classic lines. upstart attitude. #kiak900 http://t.co/v5ecb4sddr	classic line upstaattitud kiak	0.166667	0.166667	0.166667	0.166667	Positive	Positive
4	rt @kia: one reason to be on the nice list. #kiak900 http://t.co/rsp4zict9m	one reason nice list kiak	0.600000	1.000000	0.600000	1.000000	Positive	Positive
...	...	...	...	...	...	...	...	...
1138	watching the new kia k900 commercial makes me want to watch the matrix movies #kiak900	watch new kia k900 commerci make want watch matrix movi kiak	0.068182	0.227273	0.136364	0.454545	Positive	Positive
1139	dog. the #kiak900 has reclining backseats. car sex will never be the same.	dog kiak reclin backseat car sex never	0.000000	0.125000	0.000000	0.000000	Neutral	Neutral
1140	paid attention to the #kiak900 ad for 1st time & like it...but don't like $64k base price tag.	paid attent kiak ad time like ... like base price tag	-0.800000	1.000000	-0.800000	1.000000	Negative	Negative
1141	the surround-view monitor: itâ€™s like watching yourself park. #kiak900 http://t.co/2f0rcjsh60	surround-view monitor itâ like watch park kiak	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
1142	â€œ@kia: morpheus, levitating cars, sparks andâ€¦opera? watch our #kiak900 big game commercial http://t.co/e1nkm3rcwi http://t.co/jgvsaozj4hâ€¦ ðŸ˜ðŸ˜	morpheu levit car spark andâ opera watch kiak big game commerci ðŸ ðŸ	-0.133333	0.166667	-0.200000	0.250000	Negative	Negative

1143 rows × 8 columns

Εικόνα 91. Πολικότητα και υποκειμενικότητα αρχικών και επεξεργασμένων tweets (KIA)

Από τα διαγράμματα διασποράς (Εικόνα 92) παρατηρούμε ότι οι περισσότερες κουκίδες εμφανίζονται στο μέσο και δεξιά μέρος, που σημαίνει ότι τα tweets είναι περισσότερο θετικά για την KIA και δεν έχουν μέτρια υποκειμενικότητα. Όπως και πριν, τα αρχικά tweets φαίνεται να έχουν πιο έντονο το θετικό στοιχείο σε σχέση με εκείνα μετά την επεξεργασία. Κι εδώ, ο αριθμός των μηνυμάτων με μεγάλη υποκειμενικότητα, που εμφανίζονται στο πάνω μέρος του πρώτου διαγράμματος έχει μειωθεί σημαντικά, γεγονός που οφείλεται όπως έχουμε αναφέρει στην αφαίρεση των διπλότυπων εγγραφών και των λέξεων διακοπής και συγκεκριμένα των προσωπικών αντωνυμιών πρώτου προσώπου I και we.



Εικόνα 92. Διάγραμμα διασποράς πολικότητας – υποκειμενικότητας των tweets για την KIA

Από των υπολογισμό του πλήθους των tweets που αντιστοιχούν σε κάθε συναίσθημα (Εικόνα 93) παρατηρούμε ότι στην αρχική βάση δεδομένων η ανάλυση συναισθήματος έδωσε 612 θετικά, 403 ουδέτερα και 128 αρνητικά tweets. Η ίδια εικόνα εμφανίζεται και μετά την επεξεργασία των δεδομένων, με 538 θετικά, 487 ουδέτερα και 118 αρνητικά tweets. Εμφανίζεται δηλαδή μία σημαντική αύξηση των ουδέτερων, σε βάρος τόσο των θετικών όσο και των αρνητικών tweets. Τα παραπάνω αποτελέσματα παρουσιάζονται και στα παρακάτω ραβδογράμματα (Εικόνα 94)

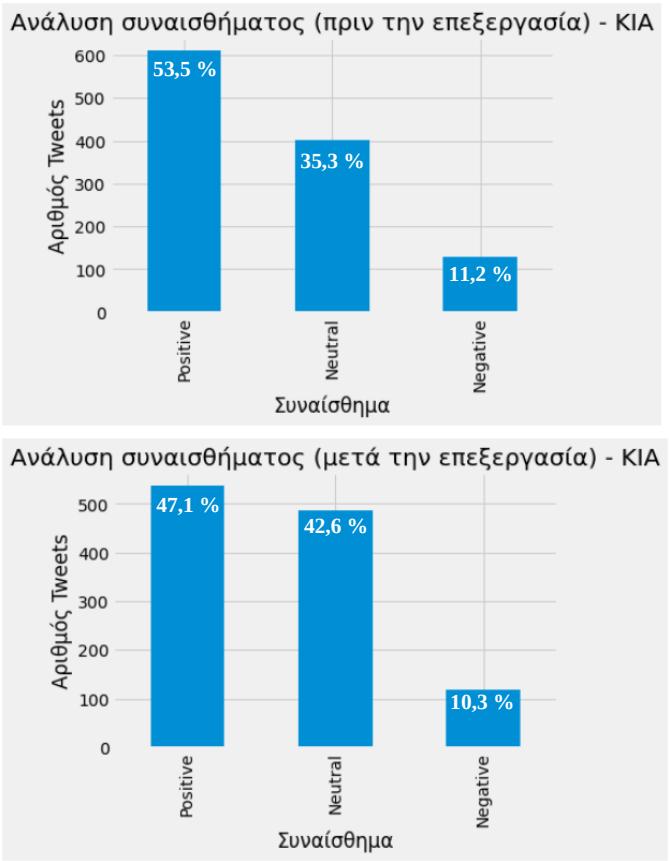
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Positive    612
Neutral     403
Negative    128
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Positive    538
Neutral     487
Negative    118
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 93. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (με στελέχωση)(ΚΙΑ)



Εικόνα 94. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία ΚΙΑ (με στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	53.5	35.3	11.2	100.0
Πολικότητα (μετά την επεξεργασία)	47.1	42.6	10.3	100.0

Πίνακας 15. Ποσοστά θετικών, ουδέτερων και αρνητικών tweets (με στελέχωση) (ΚΙΑ)

Στον Πίνακα 15 εμφανίζονται τα αντίστοιχα ποσοστά της πολικότητας των tweets, τα οποία είναι ενδεικτικά για την άποψη που έχουν οι χρήστες για την συγκεκριμένη αυτοκινητοβιομηχανία. Τόσο πριν, όσο και μετά την επεξεργασία τα σχόλια παραμένουν θετικά, με τα ουδέτερα tweets να είναι εμφανώς λιγότερα, κυρίως πριν την επεξεργασία των δεδομεών. Αξίζει να σημειώσουμε ότι και στην περίπτωση της ΚΙΑ παρατηρείτε κι ένα αρκετά υψηλό ποσοστό αρνητικών tweets αναλογικά με το σύνολο των tweets.

- *Αποτελέσματα ανάλυσης συναισθήματος* ***χωρίς της διαδικασία της στελέχωσης***

Παραλείποντας την διαδικασία της στελέχωσης, δεν παρατηρήθηκε κάποια αλλαγή μεταξύ θετικών και ουδέτερων tweets πριν και μετά την επεξεργασία, όπως προκύπτει από τους υπολογισμούς (Εικόνα 95) και τα αντίστοιχα ραβδογράμματα (Εικόνα 96), αλλά και από τον πίνακα των ποσοστών θετικών, αρνητικών και ουδέτερων tweets (Πίνακας 16).

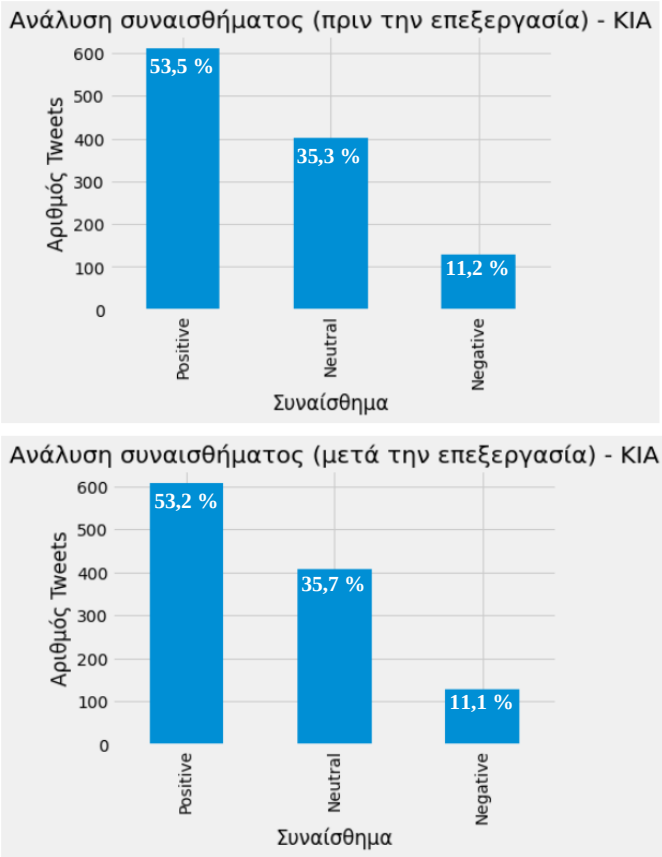
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Positive    612
Neutral     403
Negative    128
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Positive    608
Neutral     408
Negative    127
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 95. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (χωρίς στελέχωση) (ΚΙΑ)



Εικόνα 96. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία KIA (χωρίς στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	53.5	35.3	11.2	100.0
Πολικότητα (μετά την επεξεργασία)	53.2	35.7	11.1	100.0

Πίνακας 16. Ποσοστά θετικών, ουδέτερων και αρνητικών tweets (χωρίς στελέχωση) (KIA)

- **Volkswagen**

➢ *Εκκαθάριση δεδομένων*

Η βάση δεδομένων των tweets της Volkswagen αποτελούνταν αρχικά από 25153 εγγραφές (Εικόνα 97). Μετά την αφαίρεση των διπλών εγγραφών προέκυψαν 19272 εγγραφές στις οποίες έγινε ο καθαρισμός από τις αναφορές (@), τα σύμβολα hastag (#), τα προθέματα rt των retweets και τους υπερσυνδέσμους (Εικόνα 98).

```
# Εμφάνιση στήλης Tweet αρχικής βάσης δεδομένων
df
```

	Tweet
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen
...	...
25151	rt @caseyjbird: why not head to the congo from the comfort of your desk. http://t.co/1n0ymbdnj7 #instanttestdrive #volkswagen ... share theâ€¦
25152	ðÿ˜™#ultimatedubs #volkswagen #vw #golf #gti #mk4 #turbo #tdi #gt #vr6 #german #volk #rabbit #carpornâ€¦ http://t.co/11pipai5xw
25153	vw r32 #vw #volkswagen #stance #slammed #stancenation #low #like #love #hot #instacar #blue #slammedâ€¦ http://t.co/qvgpxgacfd
25154	rt @volkswagenrally: 250! the victory in ss22 @rallymexico is number 250 for @sebogier! congratulations to our #volkswagen driver #1! #gogiâ€¦
25155	rt @gruset: my new car arrived today. i am very excited :d#vw #volkswagen #up! http://t.co/z5agw2pjun

25153 rows × 1 columns

Εικόνα 97. Αρχική μορφή των tweets της αυτοκινητοβιομηχανίας Volkswagen

```
# Εμφάνιση βάσης δεδομένων μετά την αφαίρεση των διπλότυπων εγγραφών και την εκκαθάριση των δεδομένων
df
```

	Tweet	Tweet_clean
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	i am looking for porsche audi volkswagen bmw lotus ford ecoboost vehicles that want a free dyno, nearly...
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spy shots: new china-only volkswagen sedan (
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news! volkswagen jetta gl 2.0l 2005 only for $7,995.00
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl 2.0l 2005 in new york. new car for sale in ny added
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	: happy new year. another year older to our precious vws woohoooo!!! volkswagen
...	...	...
19267	high-flying #volkswagen â€" one-two in mexico | news | http://t.co/hsfava8flh http://t.co/mp90cu3grl	high-flying volkswagen â€" one-two in mexico | news |
19268	mmm:) #volkswagen #vw #vwgolf http://t.co/bez8uzlnzs	mmm:) volkswagen vw vwgolf
19269	can't beat an old vw camper ðÿ'œ #volkswagen #campervan #clean #classic #donningtonraceway http://t.co/aletmixhek	can't beat an old vw camper ðÿ'œ volkswagen campervan clean classic donningtonraceway
19270	ðÿ˜™#ultimatedubs #volkswagen #vw #golf #gti #mk4 #turbo #tdi #gt #vr6 #german #volk #rabbit #carpornâ€¦ http://t.co/11pipai5xw	ðÿ˜™ultimatedubs volkswagen vw golf gti mk4 turbo tdi gt vr6 german volk rabbit carpornâ€¦
19271	vw r32 #vw #volkswagen #stance #slammed #stancenation #low #like #love #hot #instacar #blue #slammedâ€¦ http://t.co/qvgpxgacfd	vw r32 vw volkswagen stance slammed stancenation low like love hot instacar blue slammedâ€¦

19272 rows × 2 columns

Εικόνα 98. Μορφή των tweets πριν (Tweet) και μετά (Tweet_clean) την εκκαθάριση

➢ *Προεπεξεργασία δεδομένων*

Η προεπεξεργασία των δεδομένων περιλαμβάνει τις διαδικασίες της διακριτοποίησης, της αφαίρεσης των λέξεων διακοπής και των ειδικών χαρακτήρων και της στελέχωσης, τα αποτελέσματα των οποίων παρουσιάζονται βηματικά στη συνέχεια.

✓ Διακριτοποίηση

Τα αποτελέσματα της διακριτοποίησης των επεξεργασμένων tweets εμφανίζονται στη στήλη ***Tokens*** (Εικόνα 99). Όπως παρατηρούμε, τα μέρη του λόγου και τα σημεία στίξης που απαρτίζουν τα tweets εμφανίζονται χωρισμένα με κόμμα και έχουν αποθηκευτεί σε μια λίστα.

	Tweet	Tweet_clean	Tokens
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	i am looking for porsche audi volkswagen bmw lotus ford ecoboost vehicles that want a free dyno, nearly...	[i, am, looking, for, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, that, want, a, free, dyno, ., nearly, ...]
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spy shots: new china-only volkswagen sedan (	[spy, shots, :, new, china-only, volkswagen, sedan, (]
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news! volkswagen jetta gl 2.0l 2005 only for $7,995.00	[good, news, !, volkswagen, jetta, gl, 2.0, l, 2005, only, for, $, 7,995, ., 00]
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl 2.0l 2005 in new york. new car for sale in ny added	[volkswagen, jetta, gl, 2.0, l, 2005, in, new, york, ., new, car, for, sale, in, ny, added]
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	: happy new year. another year older to our precious vws woohoooo!!! volkswagen	[:, happy, new, year, ., another, year, older, to, our, precious, vws, woohoooo, !, !, !, volkswagen]

Εικόνα 99. Αποτελέσματα διακριτοποίησης (Tokens) (Volkswagen)

✓ *Αφαίρεση λέξεων διακοπής*

Τα αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής αποθηκεύτηκαν στη στήλη ***Tweet_no_sw*** (Εικόνα 100).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	i am looking for porsche audi volkswagen bmw lotus ford ecoboost vehicles that want a free dyno, nearly...	[i, am, looking, for, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, that, want, a, free, dyno, ., nearly, ...]	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, ., nearly, ...]
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spy shots: new china-only volkswagen sedan (	[spy, shots, :, new, china-only, volkswagen, sedan, (]	[spy, shots, :, new, china-only, volkswagen, sedan, (]
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news! volkswagen jetta gl 2.0l 2005 only for $7,995.00	[good, news, !, volkswagen, jetta, gl, 2.0, l, 2005, only, for, $, 7,995, ., 00]	[good, news, !, volkswagen, jetta, gl, 2.0, l, 2005, $, 7,995, ., 00]
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl 2.0l 2005 in new york. new car for sale in ny added	[volkswagen, jetta, gl, 2.0, l, 2005, in, new, york, ., new, car, for, sale, in, ny, added]	[volkswagen, jetta, gl, 2.0, l, 2005, new, york, ., new, car, sale, ny, added]
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	: happy new year. another year older to our precious vws woohoooo!!! volkswagen	[:, happy, new, year, ., another, year, older, to, our, precious, vws, woohoooo, !, !, !, volkswagen]	[:, happy, new, year, ., another, year, older, precious, vws, woohoooo, !, !, !, volkswagen]

Εικόνα 100. Αποτελέσματα μετά την αφαίρεση των λέξεων διακοπής (Tweet_no_sw)(Volkswagen)

✓ *Αφαίρεση ειδικών χαρακτήρων*

Τα αποτελέσματα μετά την αφαίρεση των ειδικών χαρακτήρων αποθηκεύτηκαν στη στήλη ***Tweet_semifinal*** (Εικόνα 101).

	Tweet	Tweet_clean	Tokens	Tweet_no_sw	punctuation	digits	Tweet_semifinal
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	i am looking for porsche audi volkswagen bmw lotus ford ecoboost vehicles that want a free dyno, nearly...	[i, am, looking, for, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, that, want, a, free, dyno, ., nearly, ...]	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, ., nearly, ...]	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, nearly, ...]	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, nearly, ...]	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, nearly, ...]
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spy shots: new china-only volkswagen sedan (	[spy, shots, :, new, china-only, volkswagen, sedan, (]	[spy, shots, :, new, china-only, volkswagen, sedan, (]	[spy, shots, new, china-only, volkswagen, sedan]	[spy, shots, new, china-only, volkswagen, sedan]	[spy, shots, new, china-only, volkswagen, sedan]
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news! volkswagen jetta gl 2.0l 2005 only for $7,995.00	[good, news, !, volkswagen, jetta, gl, 2.0, l, 2005, only, for, $, 7,995, ., 00]	[good, news, !, volkswagen, jetta, gl, 2.0, l, 2005, $, 7,995, ., 00]	[good, news, volkswagen, jetta, gl, 2.0, l, 2005, 7,995, 00]	[good, news, volkswagen, jetta, gl, l]	[good, news, volkswagen, jetta, gl]
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl 2.0l 2005 in new york. new car for sale in ny added	[volkswagen, jetta, gl, 2.0, l, 2005, in, new, york, ., new, car, for, sale, in, ny, added]	[volkswagen, jetta, gl, 2.0, l, 2005, new, york, ., new, car, sale, ny, added]	[volkswagen, jetta, gl, 2.0, l, 2005, new, york, new, car, sale, ny, added]	[volkswagen, jetta, gl, l, new, york, new, car, sale, ny, added]	[volkswagen, jetta, gl, new, york, new, car, sale, ny, added]
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	: happy new year. another year older to our precious vws woohoooo!!! volkswagen	[:, happy, new, year, ., another, year, older, to, our, precious, vws, woohoooo, !, !, !, volkswagen]	[:, happy, new, year, ., another, year, older, precious, vws, woohoooo, !, !, !, volkswagen]	[happy, new, year, another, year, older, precious, vws, woohoooo, volkswagen]	[happy, new, year, another, year, older, precious, vws, woohoooo, volkswagen]	[happy, new, year, another, year, older, precious, vws, woohoooo, volkswagen]

Εικόνα 101. Αποτελέσματα μετά την αφαίρεση ειδικών χαρακτήρων (Tweet_semifinal) (Volkswagen)

✓ *Στελέχωση*

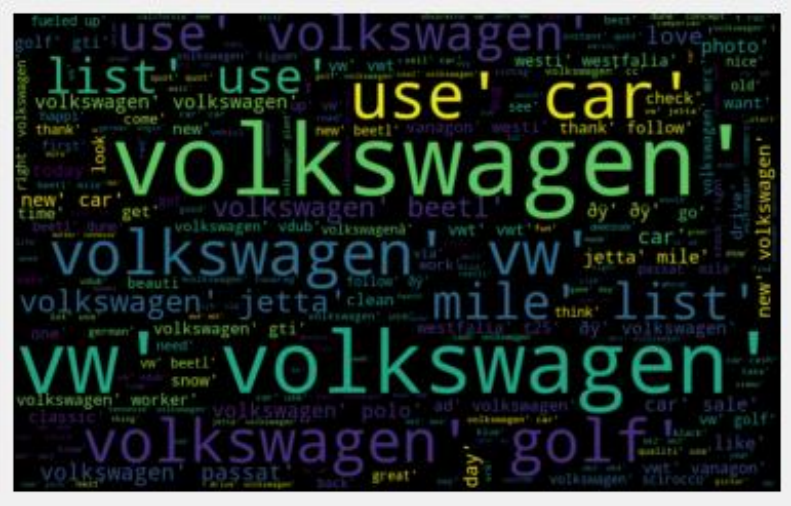
Η διαδικασία της στελέχωσης οδήγησε στην δημιουργία όπως στήλης ***Tweet_final***, τα δεδομένα της οποίας θα χρησιμοποιηθούν στη συνέχεια για την δημιουργία του σύννεφου λέξεων και της ανάλυσης συναισθήματος (Εικόνα 102).

	Tweet	Tweet_semifinal	Tweet_final
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	[looking, porsche, audi, volkswagen, bmw, lotus, ford, ecoboost, vehicles, want, free, dyno, nearly, ...]	[look, porsch, audi, volkswagen, bmw, lotu, ford, ecoboost, vehicl, want, free, dyno, nearli, ...]
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	[spy, shots, new, china-only, volkswagen, sedan]	[spi, shot, new, china-onli, volkswagen, sedan]
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	[good, news, volkswagen, jetta, gl]	[good, news, volkswagen, jetta, gl]
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	[volkswagen, jetta, gl, new, york, new, car, sale, ny, added]	[volkswagen, jetta, gl, new, york, new, car, sale, ny, ad]
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	[happy, new, year, another, year, older, precious, vws, woohoooo, volkswagen]	[happi, new, year, anoth, year, older, preciou, vw, woohoooo, volkswagen]

Εικόνα 102. Αποτελέσματα μετά την στελέχωση (Tweet_final) των tweets (Volkswagen)

✓ *Σύννεφο λέξεων*

Οι λέξεις που εμφανίζονται συχνότερα και έχουν μεγαλύτερο μέγεθος στο σύννεφο λέξεων της αυτοκινητοβιομηχανίας Volkswagen (Εικόνα 103) είναι η ίδια επωνυμία, η car, use, list, mile, golf και jetta. Τα αρχικά VW είναι μια συντομογραφία της επωνυμίας Volkswagen που χρησιμοποιείτε ευρέως. Οι λέξεις use και car συνήθως χρησιμοποιούνται μαζί και εμφανίζονται συχνά σε συζητήσεις που αφορούν αυτοκίνητα. Η λέξη golf αφορά το μοντέλο Golf που είναι από τα πιο γνωστά και με τις μεγαλύτερες πωλήσεις στον κόσμο μοντέλο της εταιρίας, ενώ αναδείχθηκε αυτοκίνητο της χρονιάς στην κατηγορία του το 2013, έναν χρόνο πριν την λήψη των συγκεκριμένων tweets. Αποτελεί κατά συνέπεια αντικείμενο σχολιασμού από τους χρήστες του Twitter. Επίσης η λέξη jetta αναφέρεται στο μοντέλο Jetta της κατηγορίας sedan, το οποίο τον Απρίλιο του 2014 ήταν το αυτοκίνητο της εταιρίας με τις περισσότερες πωλήσεις στον κόσμο. Από το συγκεκριμένο σύννεφο δεν μπορούμε να εξάγουμε συμπεράσματα για την πολικότητα των tweets, καθώς δεν εμφανίζονται συχνά και έντονα λέξεις που να δηλώνουν συναισθήματα.



Εικόνα 103. Σύννεφο λέξεων αυτοκινητοβιομηχανίας Volkswagen

✓ *Τελική μορφή δεδομένων πριν την ανάλυση συναισθήματος*

Σε αυτό το στάδιο και προκειμένου να διενεργήσουμε την ανάλυση συναισθήματος επαναφέραμε τα tweets από μορφή λίστας σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis) (Εικόνα 104)

	Tweet	Tweet_RFSA
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	look porsch audi volkswagen bmw lotu ford ecoboost vehicl want free dyno nearli ...
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spi shot new china-onli volkswagen sedan
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news volkswagen jetta gl
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl new york new car sale ny ad
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	happi new year anoth year older preciou vw woohoooo volkswagen
...	...	...
19267	high-flying #volkswagen â€" one-two in mexico | news | http://t.co/hsfava8flh http://t.co/mp90cu3grl	high-fli volkswagen one-two mexico news
19268	mmm:) #volkswagen #vw #vwgolf http://t.co/bez8uzlnzs	mmm :) volkswagen vw vwgolf
19269	can't beat an old vw camper ðÿ˜œ #volkswagen #campervan #clean #classic #donningtonraceway http://t.co/aletmixhek	can't beat old vw camper ðÿ volkswagen campervan clean classic donningtonraceway
19270	ðÿ˜™#ultimatedubs #volkswagen #vw #golf #gti #mk4 #turbo #tdi #gt #vr6 #german #volk #rabbit #carpornâ€¦ http://t.co/11pipai5xw	ðÿ ultimatedub volkswagen vw golf gti mk4 turbo tdi gt vr6 german volk rabbit carpornâ
19271	vw r32 #vw #volkswagen #stance #slammed #stancenation #low #like #love #hot #instacar #blue #slammedâ€¦ http://t.co/qvgpxgacfd	vw r32 vw volkswagen stanc slam stancen low like love hot instacar blue slammedâ

19272 rows × 2 columns

Εικόνα 104. Μετατροπή tweets σε μορφή συμβολοσειράς και αποθήκευση σε νέα στήλη (Tweet_RFSA) (Volkswagen)

- *Ανάλυση συναισθήματος*

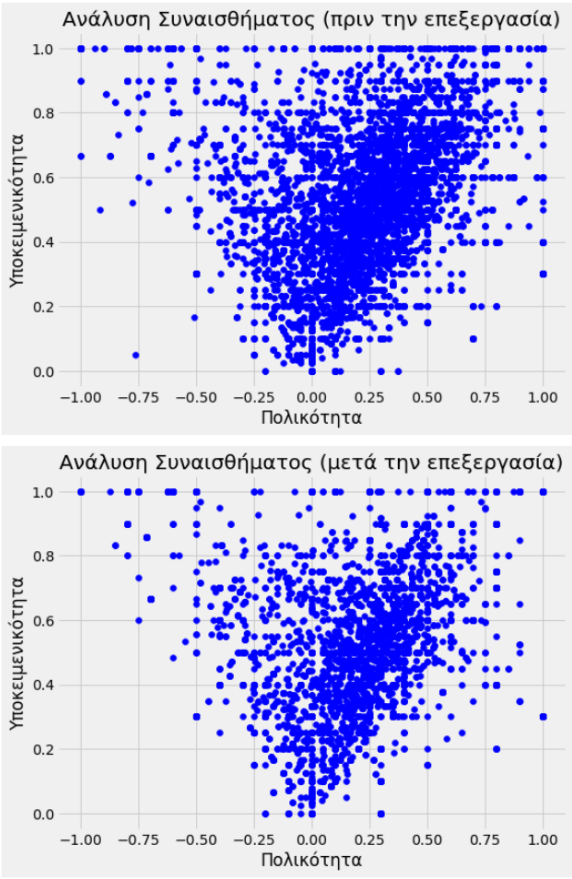
Ακολουθώντας την ίδια διαδικασία με πριν προέκυψαν οι στήλες Polarity_Tweet, Subjectivity_Tweet και Analysis_Tweet που αφορούν τα αρχικά δεδομένα, ενώ οι στήλες Polarity_Tweet_RFSA, Subjectivity_Tweet_RFSA και Analysis_Tweet_RFSA αναφέρονται στα επεξεργασμένα δεδομένα (Εικόνα 105).

	Tweet	Tweet_RFSA	Polarity_Tweet	Subjectivity_Tweet	Polarity_Tweet_RFSA	Subjectivity_Tweet_RFSA	Analysis_Tweet	Analysis_Tweet_RFSA
0	i am looking for #porsche #audi #volkswagen #bmw #lotus #ford #ecoboost vehicles that want a free dyno, nearly... http://t.co/qejkoigbug	look porsch audi volkswagen bmw lotu ford ecoboost vehicl want free dyno nearli ...	0.250000	0.600000	0.400000	0.800000	Positive	Positive
1	spy shots: new china-only #volkswagen sedan (http://t.co/an0w8zstce) http://t.co/syrrs1in3b	spi shot new china-onli volkswagen sedan	0.136364	0.454545	0.136364	0.454545	Positive	Positive
2	good #news! #volkswagen jetta gl 2.0l 2005 only for $7,995.00 http://t.co/os3dfkdfwn	good news volkswagen jetta gl	0.437500	0.800000	0.700000	0.600000	Positive	Positive
3	#volkswagen jetta gl 2.0l 2005 in #new york. new car for sale in #ny added http://t.co/erimcvgwrp	volkswagen jetta gl new york new car sale ny ad	0.136364	0.454545	0.136364	0.454545	Positive	Positive
4	rt @vwcs032: happy new year. another year older to our precious vws woohoooo!!! #volkswagen	happi new year anoth year older preciou vw woohoooo volkswagen	0.519898	0.696970	0.151515	0.393939	Positive	Positive
...	...	...	...	...	...	...	...	...
19267	high-flying #volkswagen â€" one-two in mexico | news | http://t.co/hsfava8flh http://t.co/mp90cu3grl	high-fli volkswagen one-two mexico news	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
19268	mmm:) #volkswagen #vw #vwgolf http://t.co/bez8uzlnzs	mmm :) volkswagen vw vwgolf	0.500000	1.000000	0.500000	1.000000	Positive	Positive
19269	can't beat an old vw camper ðÿ˜œ #volkswagen #campervan #clean #classic #donningtonraceway http://t.co/aletmixhek	can't beat old vw camper ðÿ volkswagen campervan clean classic donningtonraceway	0.211111	0.355556	0.211111	0.355556	Positive	Positive
19270	ðÿ˜™#ultimatedubs #volkswagen #vw #golf #gti #mk4 #turbo #tdi #gt #vr6 #german #volk #rabbit #carpornâ€¦ http://t.co/11pipai5xw	ðÿ ultimatedub volkswagen vw golf gti mk4 turbo tdi gt vr6 german volk rabbit carpornâ	0.000000	0.000000	0.000000	0.000000	Neutral	Neutral
19271	vw r32 #vw #volkswagen #stance #slammed #stancenation #low #like #love #hot #instacar #blue #slammedâ€¦ http://t.co/qvgpxgacfd	vw r32 vw volkswagen stanc slam stancen low like love hot instacar blue slammedâ	0.187500	0.462500	0.187500	0.462500	Positive	Positive

19272 rows × 8 columns

Εικόνα 105 . Πολικότητα και υποκειμενικότητα αρχικών και επεξεργασμένων tweets (Volkswagen)

Από τα διαγράμματα διασποράς (Εικόνα 106) παρατηρούμε και πάλι ότι οι περισσότερες κουκίδες εμφανίζονται στο μέσο και δεξιά μέρος, που σημαίνει ότι τα tweets είναι περισσότερο θετικά για την Volkswagen και δεν εμφανίζουν μεγάλη υποκειμενικότητα. Στα αρχικά tweets εκφράζεται εντονότερα το θετικό συναίσθημα σε σχέση με εκείνα μετά την επεξεργασία. Επίσης, ο αριθμός των μηνυμάτων με μεγάλη υποκειμενικότητα, που εμφανίζονται στο πάνω μέρος του πρώτου διαγράμματος έχει μειωθεί σημαντικά, γεγονός που οφείλεται όπως έχουμε αναφέρει στην αφαίρεση των διπλότυπων εγγραφών και των λέξεων διακοπής και συγκεκριμένα των προσωπικών αντωνυμιών πρώτου προσώπου I και we.



Εικόνα 106 Διάγραμμα διασποράς πολικότητας – υποκειμενικότητας των tweets για την Volkswagen

Από των υπολογισμό του πλήθους των tweets που αντιστοιχούν σε κάθε συναίσθημα (Εικόνα 107) παρατηρούμε ότι στην αρχική βάση δεδομένων η ανάλυση συναισθήματος έδωσε 9525 ουδέτερα, 8355 θετικά και 1392 αρνητικά tweets. Η ίδια εικόνα εμφανίζεται και μετά την επεξεργασία των δεδομένων, με 11525 ουδέτερα, 6633 θετικά και 1114 αρνητικά tweets. Εμφανίζεται δηλαδή μία σημαντική αύξηση των ουδέτερων, σε βάρος τόσο των θετικών όσο και των αρνητικών tweets. Τα παραπάνω αποτελέσματα παρουσιάζονται και στα παρακάτω ραβδογράμματα (Εικόνα 108).

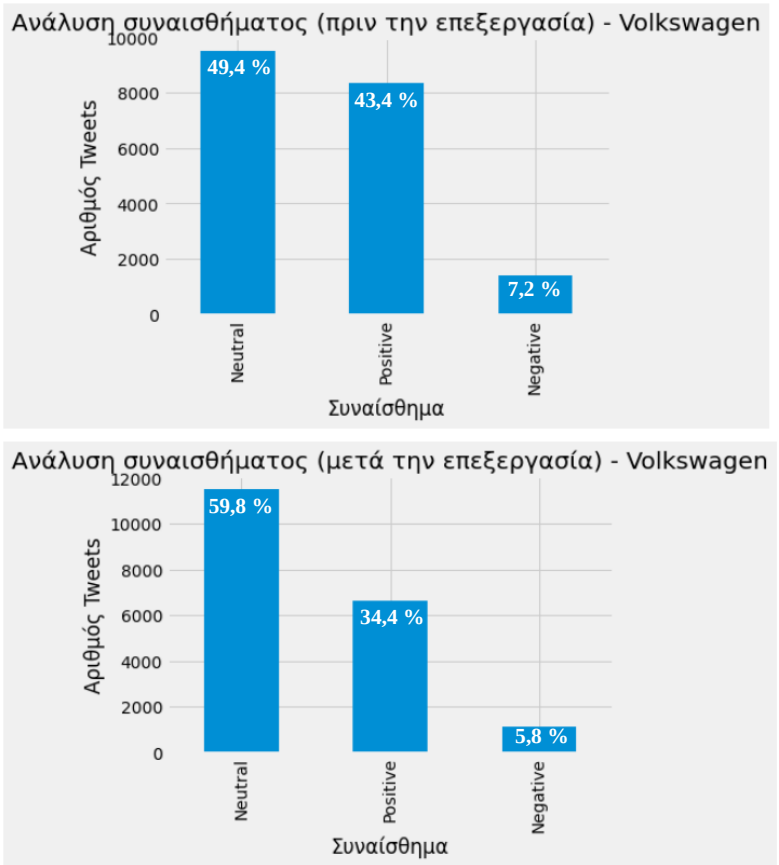
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Neutral     9525
Positive    8355
Negative    1392
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Neutral     11525
Positive     6633
Negative     1114
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 107. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (με στελέχωση) (Volkswagen)



Εικόνα 108. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Volkswagen (με στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	43.4	49.4	7.2	100.0
Πολικότητα (μετά την επεξεργασία)	34.4	59.8	5.8	100.0

Πίνακας 17. Ποσοστά θετικών, ουδέτερων και αρνητικών tweets (μεσ στελέχωση) (Volkswagen)

Στον Πίνακα 17 εμφανίζονται τα αντίστοιχα ποσοστά της πολικότητας των tweets, αναδεικνύοντας την ουδέτερη άποψη που έχουν οι χρήστες για την συγκεκριμένη αυτοκινητοβιομηχανία. Τόσο πριν, όσο και μετά την επεξεργασία τα σχόλια παραμένουν ουδέτερα σε μεγάλο ποσοστό. Είναι χαρακτηριστικό μάλιστα το γεγονός ότι μεγαλύτερη μετατόπιση προς την ουδετερότητα παρατηρείτε από τα θετικά tweets σε σχέση με τα αρνητικά. Ωστόσο το ποσοστό των αρνητικών tweets δεν είναι ιδιαίτερα υψηλό αναλογικά με το σύνολο των tweets.

- *Αποτελέσματα ανάλυσης συναισθήματος* ***χωρίς της διαδικασία της στελέχωσης***

Παραλείποντας την διαδικασία της στελέχωσης, δεν παρατηρήθηκε κάποια αλλαγή μεταξύ ουδέτερων, θετικών και αρνητικών tweets πριν και μετά την επεξεργασία, όπως προκύπτει από τους υπολογισμούς (Εικόνα 109) και τα αντίστοιχα ραβδογράμματα (Εικόνα 110), αλλά και από τον πίνακα των αντίστοιχων ποσοστών (Πίνακας 18).

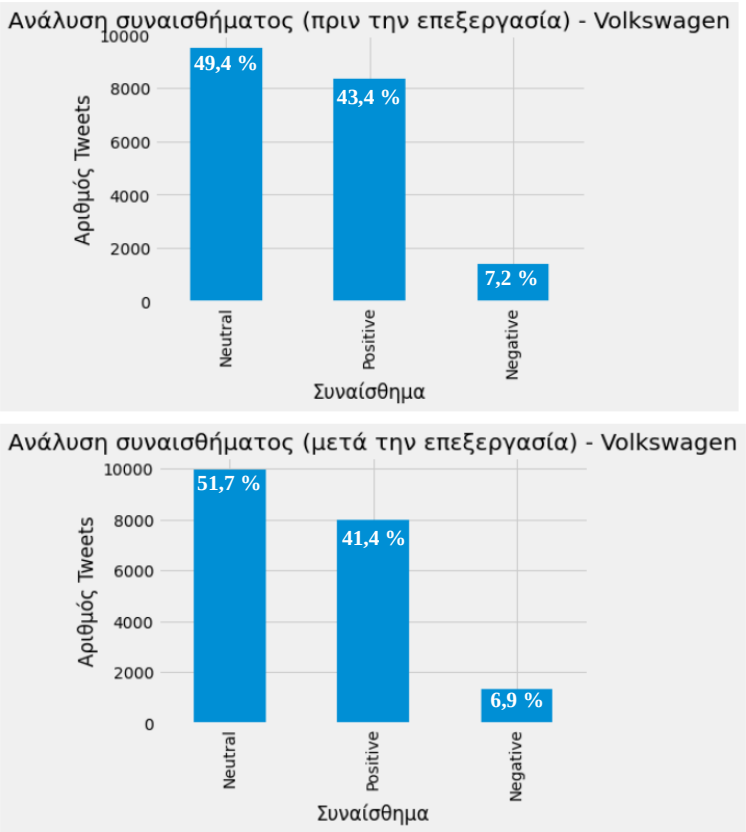
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

df_analysis['Analysis_Tweet'].value_counts()
```

```
Neutral     9525
Positive    8355
Negative    1392
Name: Analysis_Tweet, dtype: int64
```

```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

df_analysis['Analysis_Tweet_RFSA'].value_counts()
```

```
Neutral     9973
Positive    7975
Negative    1324
Name: Analysis_Tweet_RFSA, dtype: int64
```

Εικόνα 109. Υπολογισμός αριθμού θετικών, αρνητικών και ουδέτερων tweets, πριν και μετά την επεξεργασία των δεδομένων (χωρίς στελέχωση) (Volkswagen)



Εικόνα 110. Ραβδογράμματα αριθμού θετικών, αρνητικών και ουδέτερων tweets πριν και μετά την επεξεργασία για την αυτοκινητοβιομηχανία Volkswagen (χωρίς στελέχωση)

	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	43.4	49.4	7.2	100.0
Πολικότητα (μετά την επεξεργασία)	41.4	51.7	6.9	100.0

Πίνακας 18. Ποσοστά θετικών, ουδέτερων και αρνητικών tweets (χωρίς στελέχωση) (Volkswagen)

6.3. Εξόρυξη προφίλ χρηστών

Για την εξόρυξη του προφίλ των χρηστών του Twitter χρησιμοποιήσαμε δεδομένα των tweets που συλλέχθηκαν από την εφαρμογή LIWC2007 (Εικόνα 111). Πιο συγκεκριμένα, η εφαρμογή LIWC2007 ανέλυσε το σύνολο των tweets που αφορούσαν τις αυτοκινητοβιομηχανίες Audi, Chevrolet, Chrysler, KIA και Volkswagen και εξήγαγε για καθένα ένα από αυτά μια βαθμολογία/ποσοστό για κάθε μία από τις μεταβλητές του ενσωματωμένου λεξικού που διαθέτει.

	Timestamp	Tweet	TextLenght	Funct	Pronoun	Ppron	I	We	You	SheHe	They	Ipron	Article	Verbs	AuxVb	Past	Present	Future
0	fri feb 07 00:00:21 +0000 2014	rt @audi: 4 amazing race tracks. 21 hours. 1 great car	18	11,11	0	0	0	0	0	0	0	0	5,56	0	0	0	0	0
1	fri feb 07 00:00:30 +0000 2014	#throwbackthursday #audi #july2013 @princetonauc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	fri feb 07 00:00:37 +0000 2014	rt @audi: zero to 60 in under 4.6 seconds. #audinaias	14	28,57	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	fri feb 07 00:00:39 +0000 2014	rt @audi: we certainly do. #audinaias mt @naiasdeti	12	41,67	16,67	16,67	0	8,33	8,33	0	0	0	0	25	25	0	25	0
4	fri feb 07 00:00:46 +0000 2014	rt @audi: stadler: for almost 20 years, the audi #a8 h	17	41,18	5,88	5,88	0	5,88	0	0	0	0	11,76	11,76	11,76	5,88	5,88	0
5	fri feb 07 00:01:44 +0000 2014	rt @dionvmracing: since @audi means "listen"when	19	31,58	15,79	10,53	5,26	0	0	0	5,26	5,26	0	10,53	0	0	10,53	0
6	fri feb 07 00:09:57 +0000 2014	@chevrolet @corvette c7 stingray vs @porsche 911,	8	12,5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	fri feb 07 00:22:39 +0000 2014	check out the price diff between germany & ire	18	27,78	0	0	0	0	0	0	0	0	11,11	0	0	0	0	0
8	fri feb 07 00:22:58 +0000 2014	@tashmanhardware @gorillaglue oh how i miss my	7	57,14	28,57	28,57	28,57	0	0	0	0	0	0	14,29	0	0	14,29	0
9	fri feb 07 00:30:06 +0000 2014	#tthursday: red @audi tt having fun in the snow. phc	11	36,36	0	0	0	0	0	0	0	0	9,09	9,09	9,09	0	0	0
10	fri feb 07 00:41:49 +0000 2014	rt @audiforlife: #tthursday: red @audi tt having fun	12	33,33	0	0	0	0	0	0	0	0	8,33	8,33	8,33	0	0	0
11	fri feb 07 00:44:15 +0000 2014	another #boastpost with my @audi #allroad and it's	16	50	12,5	6,25	6,25	0	0	0	0	6,25	0	0	0	0	0	0
12	fri feb 07 01:01:01 +0000 2014	audi announces student finalists for the future of m	11	27,27	0	0	0	0	0	0	0	0	9,09	0	0	0	0	0
13	fri feb 07 01:13:23 +0000 2014	@jmdc88 @audi @audiforlife @autopainttech @on	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	fri feb 07 01:17:53 +0000 2014	rt @alisterrobbie: @jmdc88 @audi @audiforlife @a	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	fri feb 07 01:22:46 +0000 2014	exciting news! @pmk_bnc wins digiday video award	13	30,77	0	0	0	0	0	0	0	0	15,38	0	0	0	0	0
16	fri feb 07 01:25:22 +0000 2014	rt @pmk_bnc: exciting news! @pmk_bnc wins digid	14	28,57	0	0	0	0	0	0	0	0	14,29	0	0	0	0	0
17	fri feb 07 01:27:36 +0000 2014	corvette take on the juggernauts http://t.co/8umlfv	5	40	0	0	0	0	0	0	0	0	20	20	0	0	20	0
18	fri feb 07 01:34:40 +0000 2014	--> rt @audiforlife: #tthursday: red @audi tt havi	13	30,77	0	0	0	0	0	0	0	0	7,69	7,69	7,69	0	0	0
19	fri feb 07 01:36:42 +0000 2014	@audi please bring the rs6 avant to the us. a manual	25	56	12	8	4	4	0	0	0	4	16	20	12	0	12	4
20	fri feb 07 01:40:24 +0000 2014	rt @audiforlife: #tthursday: red @audi tt having fun	12	33,33	0	0	0	0	0	0	0	0	8,33	8,33	8,33	0	0	0
21	fri feb 07 01:40:40 +0000 2014	rt @zach_roerig: thanks for the tickets @audi !! http	5	40	0	0	0	0	0	0	0	0	20	20	0	0	20	0
22	fri feb 07 01:41:29 +0000 2014	rt @audi_nashville: @audi sent over #doghoodies m	17	41,18	11,76	11,76	0	5,88	5,88	0	0	0	0	17,65	0	5,88	11,76	0
23	fri feb 07 01:42:06 +0000 2014	â€œ@audi: we certainly do. #audinaias mt @naiasd	12	41,67	16,67	16,67	0	8,33	8,33	0	0	0	0	25	25	0	25	0
24	fri feb 07 01:42:41 +0000 2014	@alisterrobbie you have a days jump on the #audipi	7	71,43	14,29	14,29	0	0	14,29	0	0	0	28,57	14,29	14,29	0	14,29	0
25	fri feb 07 01:43:36 +0000 2014	@autopainttech it's the joy of living in the future! @	8	62,5	12,5	0	0	0	0	0	0	12,5	25	0	0	0	0	0
26	fri feb 07 01:44:42 +0000 2014	@alisterrobbie @autopainttech @audi @audiforlife	6	33,33	0	0	0	0	0	0	0	0	16,67	0	0	0	0	0
27	fri feb 07 01:47:10 +0000 2014	@alisterrobbie since you started #audipics...@jmdc	3	66,67	33,33	33,33	0	0	33,33	0	0	0	0	33,33	0	33,33	0	0
28	fri feb 07 01:48:33 +0000 2014	@alisterrobbie my friend was down under 2 days ag	10	50	10	10	10	0	0	0	0	0	0	10	10	10	0	0
29	fri feb 07 01:49:23 +0000 2014	rt @autopainttech: @alisterrobbie since you started	5	40	20	20	0	0	20	0	0	0	0	20	0	20	0	0
30	fri feb 07 01:49:49 +0000 2014	@jmdc88 @alisterrobbie @audi @audiforlife @bath	7	57,14	28,57	14,29	0	0	0	0	14,29	14,29	14,29	14,29	0	0	14,29	0
31	fri feb 07 01:55:39 +0000 2014	best in the world. we got 4 rings! rt @lesty4567: i lov	14	50	14,29	14,29	7,14	7,14	0	0	0	0	7,14	14,29	0	7,14	7,14	0
32	fri feb 07 01:59:19 +0000 2014	rt @autopainttech: @jmdc88 @alisterrobbie @audi	6	50	33,33	16,67	0	0	0	0	16,67	16,67	0	16,67	0	0	16,67	0
33	fri feb 07 01:59:59 +0000 2014	rt @ichrisharrison: best in the world. we got 4 rings!	15	46,67	13,33	13,33	6,67	6,67	0	0	0	0	6,67	13,33	0	6,67	6,67	0
34	fri feb 07 02:00:33 +0000 2014	@dionvmracing @audi learn something new every d	9	44,44	22,22	11,11	0	0	0	0	11,11	11,11	0	11,11	11,11	0	11,11	0
35	fri feb 07 02:04:00 +0000 2014	@rockville_audi thanks so much! did you see the @	11	81,82	18,18	18,18	0	9,09	9,09	0	0	0	9,09	36,36	18,18	18,18	18,18	0
36	fri feb 07 02:08:58 +0000 2014	rt @audi: the audi sport #quattro #laserlight concep	10	30	0	0	0	0	0	0	0	0	10	0	0	0	0	0
37	fri feb 07 02:13:27 +0000 2014	rent an @audi for $9?? @jimellisaudiatl rt @garyleff	10	30	0	0	0	0	0	0	0	0	10	0	0	0	0	0
38	fri feb 07 02:33:19 +0000 2014	road trip successful.. new audiâœ…ðŸ˜ @audi http:/	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	fri feb 07 02:33:42 +0000 2014	@audi i'm rewatching irobot on amc- i was pleasantl	22	40,91	9,09	4,55	4,55	0	0	0	0	4,55	13,64	9,09	9,09	9,09	0	0

Εικόνα 111. Μορφή της βάσης δεδομένων προς ανάλυση

Για τον σκοπό της ανάλυσής μας χρησιμοποιήσαμε το λογισμικό SPSS, το οποίο ενσωματώνει ένα μεγάλο αριθμό αυτοματοποιημένων εργαλείων ανάλυσης. Για την εξόρυξη του προφίλ των χρηστών του Twitter χρησιμοποιήσαμε το εργαλείο κατηγοριοποίησης TwoStep Cluster, με στόχο την δημιουργία ομάδων χρηστών με κοινά χαρακτηριστικά τα οποία θα μας οδηγήσουν στη συνέχεια στην κατηγοριοποίησή τους με βάση τα 7 επίπεδα της «Σκάλας των κοινωνικών τεχνολογικών συμπεριφορών» (σύμφωνα με την διαδικτυακή έρευνα της Forrester Research, 2010) και το πρότυπο OCEAN. Η ερμηνεία των αποτελεσμάτων της ανάλυσης βασίστηκε σε σημαντικό βαθμό στα συμπεράσματα της μελέτης των Tausczik και Pennebaker (2010) για την ψυχολογική σημασία των λέξεων [81] και της εργασίας των Schwarts et al. (Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach) [137] που

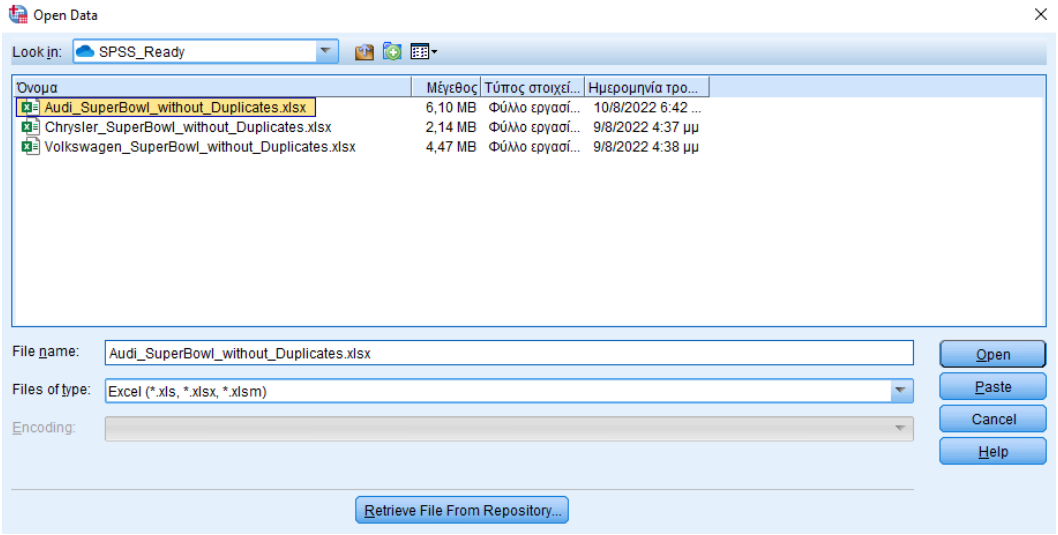
συσχετίζει της κατηγορίες λέξεων του λεξικού LIWC με το φύλο, την ηλικία και το μοντέλο προσωπικότητας OCEAN.

Τα κοινά χαρακτηριστικά των χρηστών αναφέρονται στις μεταβλητές που θα συνθέσουν τις συστάδες που θα δημιουργηθούν μετά την ανάλυση. Καθώς οι ποσοτικές μεταβλητές σχετίζονται με τις λέξεις που χρησιμοποιούνται για την σύνταξη των tweets, η χρήση συγκεκριμένων λέξεων από τους χρήστες φανερώνουν συγκεκριμένα μοτίβα αυτών. Οι μεταβλητές βάσει των οποίων θα δημιουργηθούν οι συστάδες θα καθορίσουν και τις αντίστοιχες κατηγορίες χρηστών.

Στη συνέχεια περιγράφονται τα βήματα της ανάλυσης που ακολουθήσαμε και αφορούν τα tweets των χρηστών της αυτοκινητοβιομηχανίας Audi, τα οποία εφαρμόσαμε ακολούθως και για τις υπόλοιπες εταιρίες.

6.3.1. Προετοιμασία δεδομένων

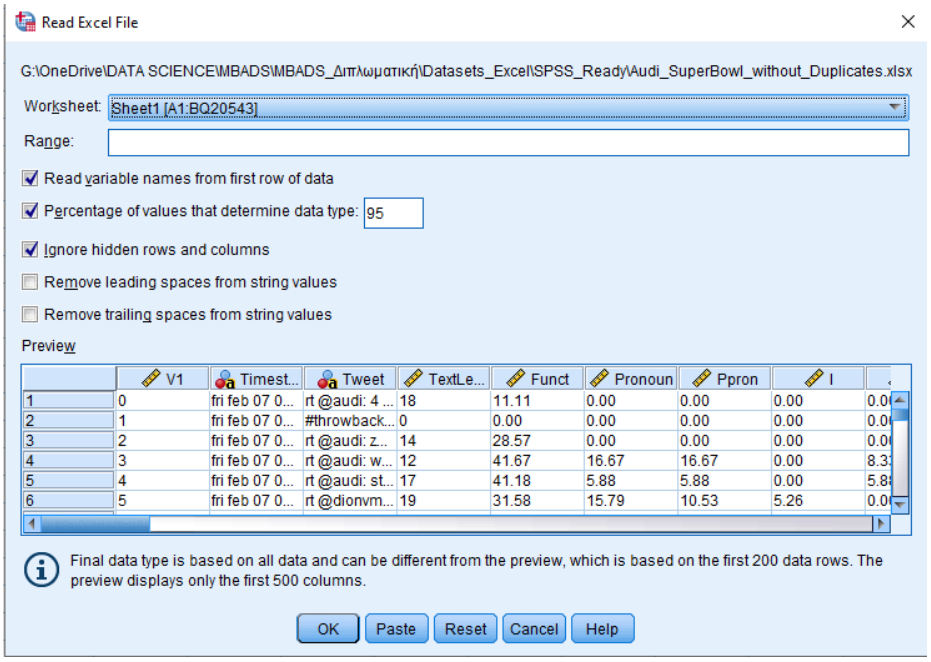
✓ Εισαγάγαμε στο SPSS τα δεδομένα των tweets για την αυτοκινητοβιομηχανία Audi (Εικόνα 112), αφού προηγουμένως επεξεργαστήκαμε την βάση δεδομένων, αφαιρώντας τις πολλαπλές εγγραφές των ίδιων tweets, όπως και στην περίπτωση της ανάλυσης συναισθήματος. Το αρχείο των δεδομένων ήταν σε μορφή excel (.xlsx) και το σύνολο των εγγραφών που προέκυψαν ήταν 20541.



Εικόνα 112. Άνοιγμα αρχείου δεδομένων της αυτοκινητοβιομηχανίας Audi

Στο παράθυρο που εμφανίστηκε δίνονται μια σειρά από επιλογές για την εισαγωγή των δεδομένων. Επιλέξαμε «Read variable names from first row of data», ώστε να εμφανιστούν τα ονόματα των μεταβλητών στην πρώτη σειρά του πίνακα, ενώ η

επιλογή «Ignore hidden rows and columns» αποκλείει τις εγγραφές που δεν εμφανίζονται στο αρχικό μας αρχείο (Εικόνα 113).



Εικόνα 113. Εισαγωγή δεδομένων στο πρόγραμμα στατιστικής ανάλυσης SPSS

Η βάση δεδομένων αποτελούνταν αρχικά από 69 μεταβλητές, εκ των οποίων οι μεταβλητές V1, Timestamp, Tweet και UserID αφαιρέθηκαν, καθώς περιείχαν την αρίθμηση των εγγραφών, την ημερομηνία και ώρα λήψης του κάθε tweet, το κείμενο του tweet και τον χρήστη που το δημοσίευσε αντίστοιχα, και δεν αποτελούν χρήσιμες μεταβλητές για την ανάλυσή μας (Εικόνα 114).

	TextLenght	Funct	Pronoun	Ppron	I	We	You	SheHe	They	Ipron	Article	Verbs
1	18	11,11	,00	,00	,00	,00	,00	,00	,00	,00	5,56	,00
2	0	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
3	14	28,57	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
4	12	41,67	16,67	16,67	,00	8,33	8,33	,00	,00	,00	,00	25,00
5	17	41,18	5,88	5,88	,00	5,88	,00	,00	,00	,00	11,76	11,76
6	19	31,58	15,79	10,53	5,26	,00	,00	,00	5,26	5,26	,00	10,53
7	8	12,50	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
8	18	27,78	,00	,00	,00	,00	,00	,00	,00	,00	11,11	,00
9	7	57,14	28,57	28,57	28,57	,00	,00	,00	,00	,00	,00	14,29
10	11	36,36	,00	,00	,00	,00	,00	,00	,00	,00	9,09	9,09
11	12	33,33	,00	,00	,00	,00	,00	,00	,00	,00	8,33	8,33
12	16	50,00	12,50	6,25	6,25	,00	,00	,00	,00	6,25	,00	,00
13	11	27,27	,00	,00	,00	,00	,00	,00	,00	,00	9,09	,00
14	1	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
15	2	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00	,00
16	13	30,77	,00	,00	,00	,00	,00	,00	,00	,00	15,38	,00
17	14	28,57	,00	,00	,00	,00	,00	,00	,00	,00	14,29	,00
18	5	40,00	,00	,00	,00	,00	,00	,00	,00	,00	20,00	20,00
19	13	30,77	,00	,00	,00	,00	,00	,00	,00	,00	7,69	7,69
20	25	56,00	12,00	8,00	4,00	4,00	,00	,00	,00	4,00	16,00	20,00
21	12	33,33	,00	,00	,00	,00	,00	,00	,00	,00	8,33	8,33
22	5	40,00	,00	,00	,00	,00	,00	,00	,00	,00	20,00	20,00
23	17	41,18	11,76	11,76	,00	5,88	5,88	,00	,00	,00	,00	17,65

Εικόνα 114. Μορφή βάσης δεδομένων προς ανάλυση

Επιπλέον ορίσαμε δύο δεκαδικά ψηφία για τις αριθμητικές μεταβλητές του λεξικού του LIWC2007. Από τις υπόλοιπες μεταβλητές οι 64 αφορούν τις κατηγορίες λέξεων του

λεξικού LIWC2007, οι οποίες είναι αριθμητικές και είναι αυτές που θα χρησιμοποιηθούν στη συνέχεια για την ανάλυση (Εικόνα 115).

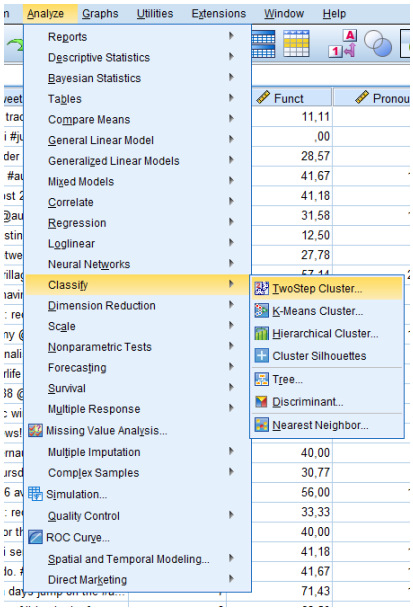
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	TextLenght	Numeric	11	2		None	None	12	Right	Scale	Input
2	Funct	Numeric	18	2		None	None	12	Right	Scale	Input
3	Pronoun	Numeric	20	2		None	None	12	Right	Scale	Input
4	Ppron	Numeric	18	2		None	None	12	Right	Scale	Input
5	I	Numeric	18	2		None	None	12	Right	Scale	Input
6	We	Numeric	18	2		None	None	12	Right	Scale	Input
7	You	Numeric	18	2		None	None	12	Right	Scale	Input
8	SheHe	Numeric	18	2		None	None	12	Right	Scale	Input
9	They	Numeric	18	2		None	None	12	Right	Scale	Input
10	Ipron	Numeric	18	2		None	None	12	Right	Scale	Input
11	Article	Numeric	20	2		None	None	12	Right	Scale	Input
12	Verbs	Numeric	18	2		None	None	12	Right	Scale	Input
13	AuxVb	Numeric	18	2		None	None	12	Right	Scale	Input
14	Past	Numeric	18	2		None	None	12	Right	Scale	Input
15	Present	Numeric	20	2		None	None	12	Right	Scale	Input
16	Future	Numeric	19	2		None	None	12	Right	Scale	Input
17	Adverbs	Numeric	20	2		None	None	12	Right	Scale	Input
18	Prep	Numeric	18	2		None	None	12	Right	Scale	Input
19	Conj	Numeric	18	2		None	None	12	Right	Scale	Input
20	Negate	Numeric	19	2		None	None	12	Right	Scale	Input
21	Quant	Numeric	18	2		None	None	12	Right	Scale	Input
22	Numbers	Numeric	20	2		None	None	12	Right	Scale	Input
23	Swear	Numeric	18	2		None	None	12	Right	Scale	Input
24	Social	Numeric	19	2		None	None	12	Right	Scale	Input
25	Family	Numeric	19	2		None	None	12	Right	Scale	Input
26	Friends	Numeric	20	2		None	None	12	Right	Scale	Input
27	Humans	Numeric	19	2		None	None	12	Right	Scale	Input

Εικόνα 115. Πίνακας μεταβλητών βάσης δεδομένων (Audi)

6.3.2. Συσταδοποίηση Δύο Βημάτων (TwoStep Cluster)

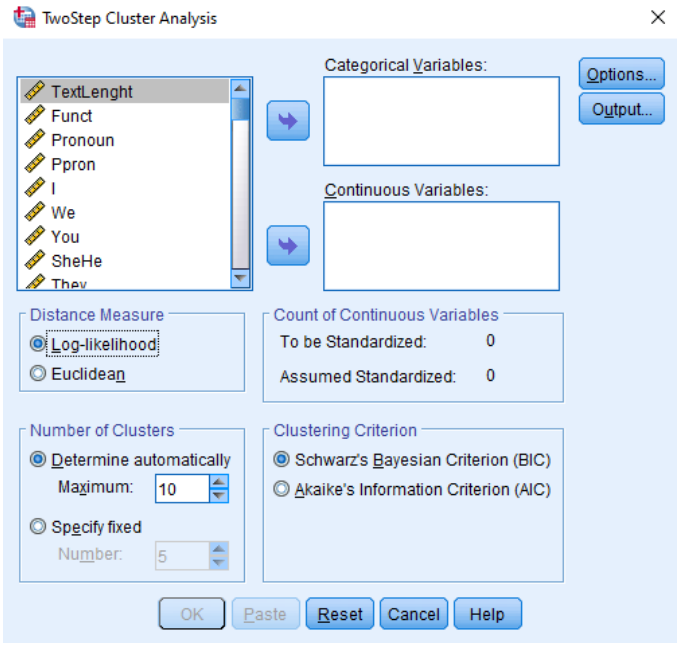
6.3.2.1.Audi

Από το μενού του προγράμματος SPSS επιλέξαμε *Analyze* → *Classify* → *TwoStep Cluster*… (Εικόνα 116)



Εικόνα 116

Στο παράθυρο που εμφανίστηκε μας δίνονται μια σειρά από επιλογές για την διεξαγωγή της ανάλυσης οι οποίες περιγράφονται στη συνέχεια (Εικόνα 117).

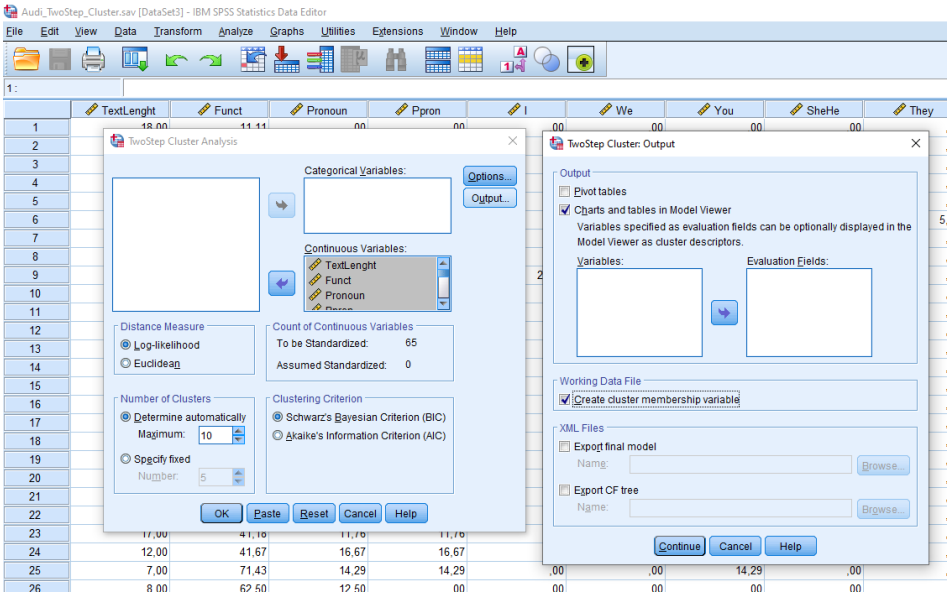


Εικόνα 117. Παράθυρο επιλογών ανάλυσης TwoStep Cluster

- **Distance Measure:** Καθορίζει τον τρόπο υπολογισμού των αποστάσεων μεταξύ των συστάδων. Υπάρχουν δύο επιλογές, η απόσταση λογαριθμικής πιθανότητας (Log-likelihood) και η Ευκλείδεια απόσταση (Euclidean). Η απόσταση λογαριθμικής πιθανότητας μπορεί να χρησιμοποιηθεί τόσο για συνεχείς όσο και για κατηγορικές μεταβλητές. Η απόσταση μεταξύ δύο συστάδων συσχετίζεται με τη μείωση της συνάρτησης του φυσικού λογαρίθμου πιθανότητας, καθώς ομαδοποιούνται σε μια συστάδα. Για τον υπολογισμό της απόστασης λογαριθμικής πιθανότητας, θεωρείται ότι οι συνεχείς μεταβλητές έχουν κανονικές κατανομές και οι κατηγορικές μεταβλητές έχουν πολυωνυμικές κατανομές και επίσης οι μεταβλητές είναι ανεξάρτητες μεταξύ τους. Η ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί μόνο εάν όλες οι μεταβλητές είναι συνεχείς. Η ευκλείδεια απόσταση μεταξύ δύο σημείων ορίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των διαφορών μεταξύ των συντεταγμένων των σημείων . Για τις συστάδες, η απόσταση μεταξύ δύο συστάδων ορίζεται ως η ευκλείδεια απόσταση μεταξύ των κέντρων τους. Το κέντρο μιας συστάδας (που ονομάζεται centroid) είναι το διάνυσμα των μέσων της συστάδας της κάθε μεταβλητής.[172]
- **Count of Continuous Variables:** Το πλαίσιο αυτό παρέχει μια σύνοψη των προδιαγραφών συνεχούς τυποποίησης μεταβλητών που έγιναν στο παράθυρο διαλόγου ***Options…*** .

- **Number of Clusters:** Δίνεται η δυνατότητα αυτόματου καθορισμού του «καλύτερου» μέγιστου αριθμού συστάδων ή χειροκίνητου καθορισμού συγκεκριμένου αριθμού συστάδων.
- **Clustering Criterion:** Αυτή η επιλογή καθορίζει τον τρόπο με τον οποίο ο αλγόριθμος αυτόματης συσταδοποίησης καθορίζει τον αριθμό των συστάδων. Μπορεί να προσδιοριστεί είτε το Κριτήριο Schwarz's Bayesian (Schwarz's Bayesian Criterion - BIC), είτε το Κριτήριο Πληροφοριών του Akaike (Akaike Information Criterion - AIC) ως κριτήριο συσταδοποίησης. Ο δείκτης αυτός χρησιμοποιείται για την εύρεση μιας αρχικής εκτίμησης για τον αριθμό των συστάδων.[173]

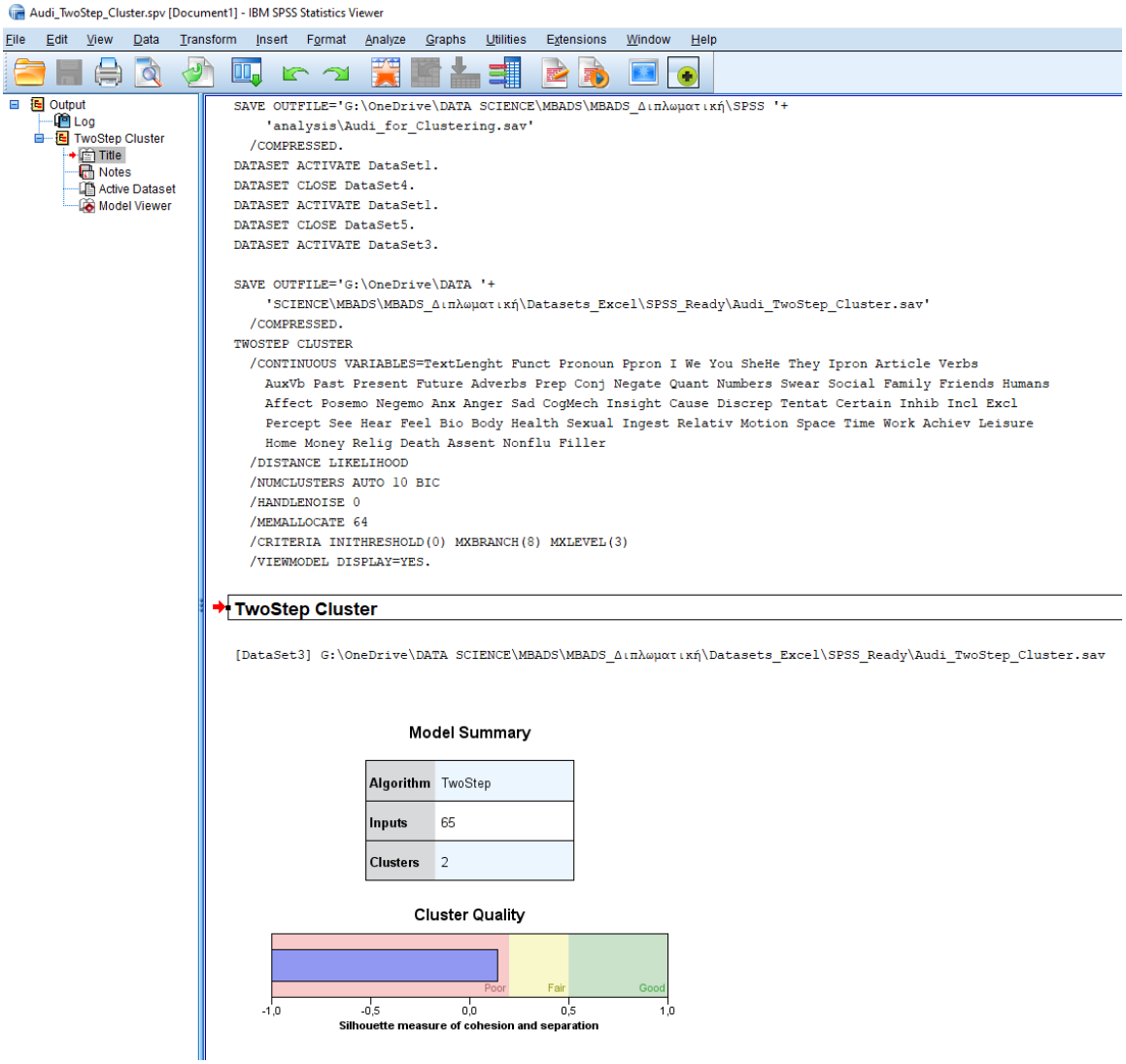
Η βάση δεδομένων μας αποτελείται από ποσοτικές μεταβλητές και ως εκ τούτου τις μεταφέραμε στο πλαίσιο ***Continuous Variables***. Επιλέξαμε για την μέτρηση απόστασης την Log-likelihood και καθορίσαμε τον μέγιστο αριθμό των συστάδων σε 10. Επιπλέον ως κριτήριο συσταδοποίησης επιλέξαμε το κριτήριο Schwarz's Bayesian (Εικόνα 118). Στη συνέχεια από το παράθυρο διαλόγου ***Output…*** «τσεκάραμε» την επιλογή Charts and tables in Model Viewer, ώστε οι μεταβλητές που καθορίζονται ως πεδία αξιολόγησης να μπορούν (προαιρετικά) να εμφανίζονται στο πρόγραμμα προβολής μοντέλων (Model Viewer) ως λέξεις που περιγράφουν τις συστάδες, καθώς και την επιλογή ***Create cluster membership variable,*** ώστε μετά την ανάλυση να δημιουργηθεί στη βάση δεδομένων μία μεταβλητή με τον αριθμό της συστάδας που ανήκει η κάθε εγγραφή.



Εικόνα 118. Καθορισμός παραμέτρων ανάλυσης

Τα αποτελέσματα της ανάλυσης εμφανίζονται στο παράθυρο Viewer του SPSS. Οι πληροφορίες που δίνονται είναι μια περίληψη του μοντέλου (Model Summary), όπου

εμφανίζονται ο αλγόριθμος που εφαρμόστηκε, το πλήθος των μεταβλητών εισόδου στον αλγόριθμο, καθώς και το πλήθος των συστάδων που δημιουργήθηκαν, ενώ στη συνέχεια εμφανίζεται ένα ραβδόγραμμα, το οποίο μας δίνει άμεση πληροφόρηση για την ποιότητα του μοντέλου (Cluster Quality) της συσταδοποίησης, σύμφωνα με μέτρο της συνοχής και του διαχωρισμού μεταξύ των συστάδων (Εικόνα 119).



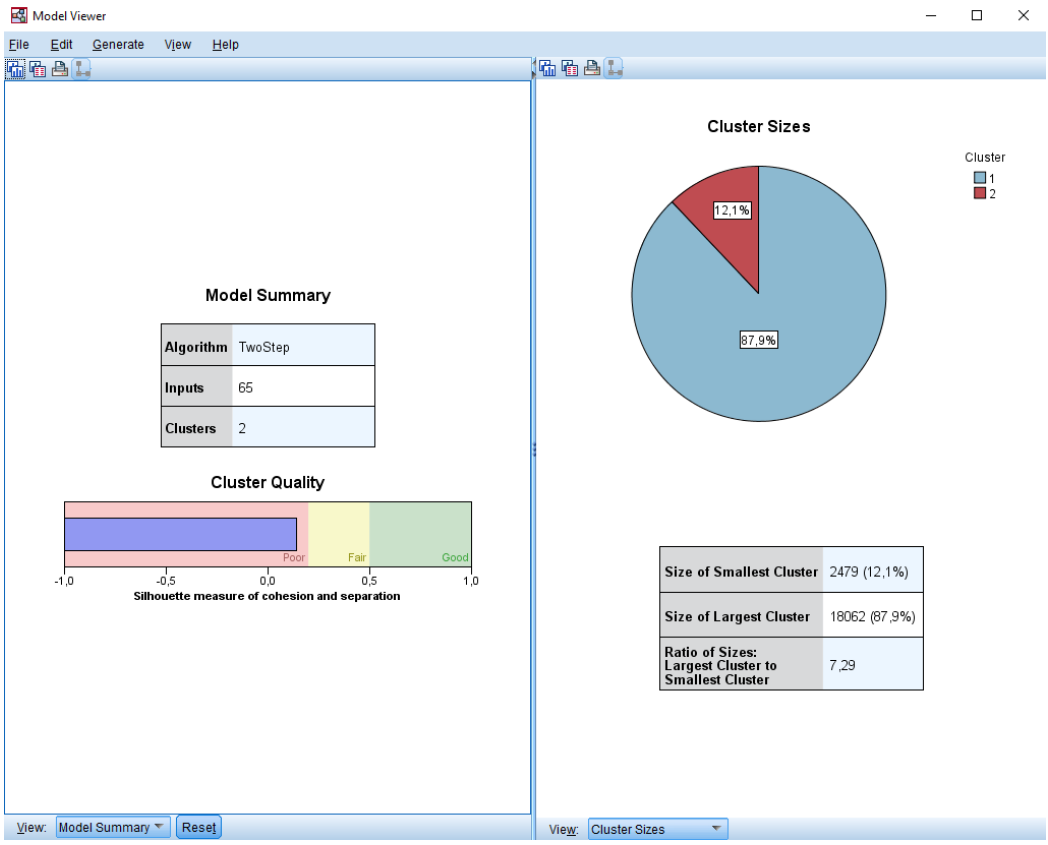
Εικόνα 119. Περίληψη μοντέλου και ποιότητα συσταδοποίησης (Audi)

Όπως παρατηρούμε, το μοντέλο μας αποτελείται από τις 65 μεταβλητές του λεξικού LIWC2007 και μετά την εφαρμογή του αλγορίθμου δημιουργήθηκαν 2 συστάδες. Ωστόσο, η ποιότητα των συστάδων είναι «φτωχή» (poor) όπως υποδηλώνει η μπάρα του διαγράμματος Cluster Quality. Τα αποτελέσματα της φτωχής, μέτριας και καλής ποιότητας βασίζονται στο έργο των Kaufman και Rousseeuw (1990) σχετικά με την ερμηνεία των δομών των συστάδων. Ένα καλό αποτέλεσμα ισοδυναμεί με δεδομένα που αντικατοπτρίζουν την αξιολόγηση των Kaufman και Rousseeuw είτε ως βάσιμα είτε ως

ισχυρές ενδείξεις της δομής μια συστάδας, ένα μέτριο αποτέλεσμα αντικατοπτρίζει στην αξιολόγησή τους, αδύναμα αποδεικτικά στοιχεία και ένα κακό αποτέλεσμα αντικατοπτρίζει στην αξιολόγησή τους, μη ύπαρξη σημαντικών αποδεικτικών στοιχείων.[174]

Η μέτρηση σιλουέτας (silhouette measure) είναι ο μέσος όρος, σε όλες τις εγγραφές, (B−A)/max(A,B), όπου A είναι η απόσταση της εγγραφής από το κέντρο της συστάδας και B είναι η απόσταση της εγγραφής από το πλησιέστερο κέντρο συστάδας στο οποίο δεν ανήκει. Ένας συντελεστής σιλουέτας 1 σημαίνει ότι όλες οι περιπτώσεις βρίσκονται απευθείας στα κέντρα συστάδων τους. Μια τιμή −1 θα σήμαινε ότι όλες οι περιπτώσεις βρίσκονται στα κέντρα συστάδων κάποιας άλλης συστάδας. Μια τιμή 0 σημαίνει, κατά μέσο όρο, ότι οι περιπτώσεις βρίσκονται σε ίση απόσταση μεταξύ του δικού τους κέντρου συστάδας και της πλησιέστερης άλλης συστάδας. [175]

Το ραβδόγραμμα είναι ένα αλληλεπιδραστικό αντικείμενο προβολής μοντέλου (Model Viewer object), το οποίο ενεργοποιείται κάνοντας διπλό κλικ πάνω του (Εικόνα 120).

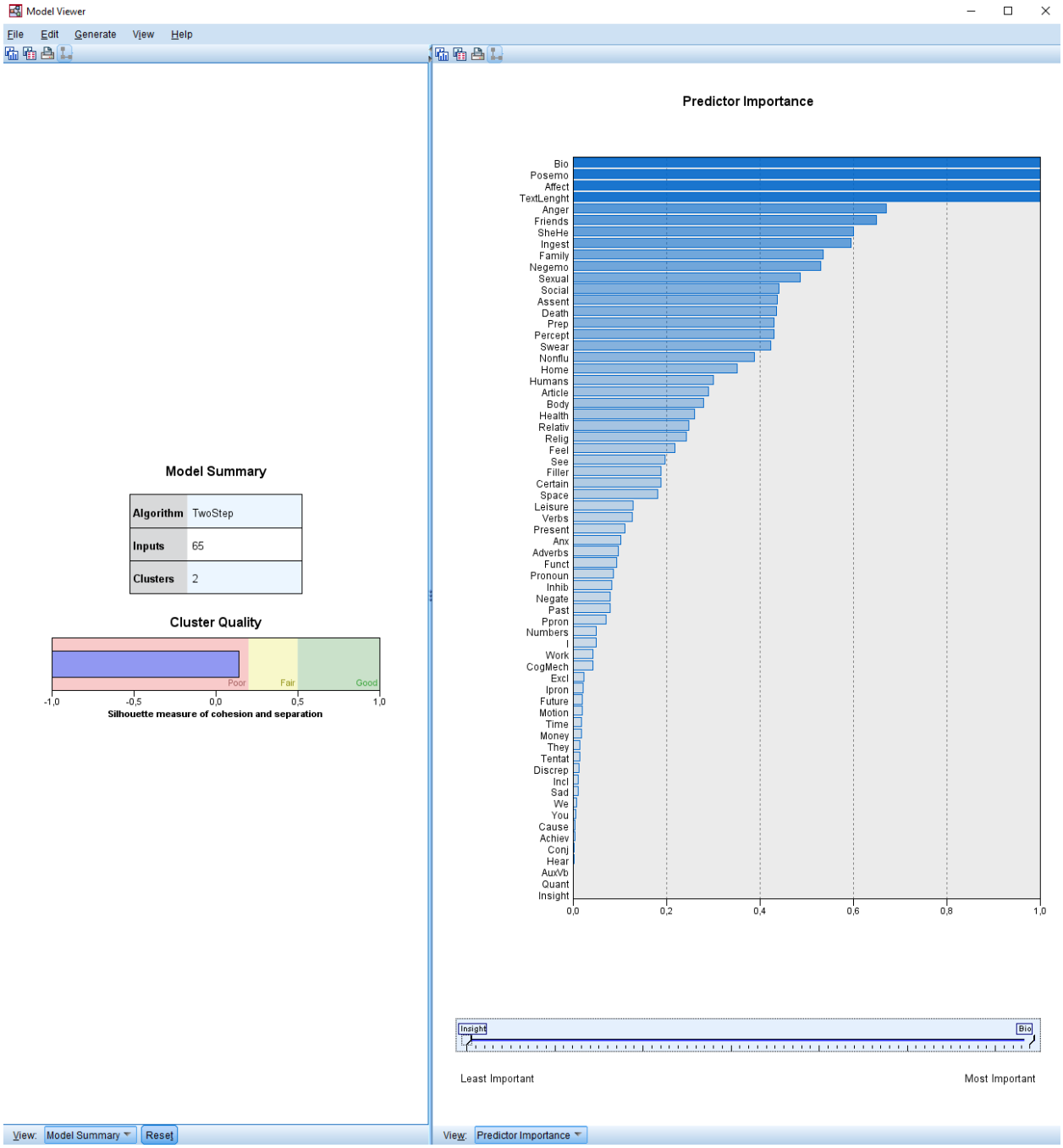


Εικόνα 120. Παράθυρο προβολής μοντέλων

Το παράθυρο Model Viewer, χωρίζεται σε δύο υποπαράθυρα (Εικόνα 121). Στο δεξί από προεπιλογή εμφανίζεται ένα διάγραμμα πίτας με πληροφορίες για το πλήθος και το

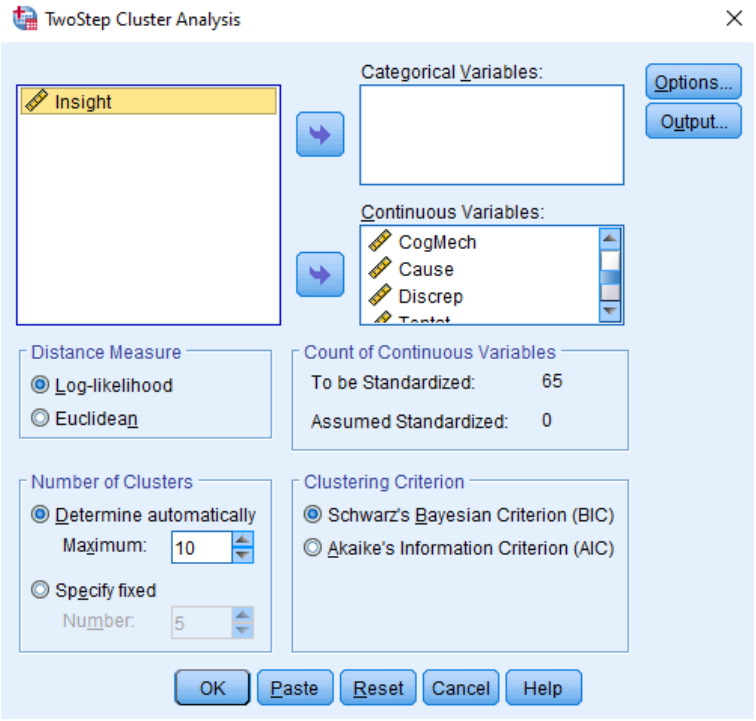
μέγεθος των συστάδων. Παρατηρούμε ότι η μεγαλύτερη συστάδα αποτελεί το 87,9% του δείγματος, ενώ η μικρότερη το 12,1%.

Επιλέγοντας στο κάτω μέρος του παραθύρου View → Predictor Importance εμφανίζεται ένα ραβδόγραμμα της σημαντικότητας της κάθε μεταβλητής. Εκείνες που βρίσκονται χαμηλότερα είναι αυτές που επηρεάζουν λιγότερο (ή και καθόλου) την δημιουργία των συστάδων. Προκειμένου να βελτιώσουμε το μοντέλο, αφαιρούμε σταδιακά σε κάθε επόμενη εφαρμογή του αλγορίθμου, την μεταβλητή που βρίσκεται χαμηλότερα στο διάγραμμα.



Εικόνα 121. Παράθυρο προβολής μοντέλων – Εμφάνιση ραβδογράμματος σημαντικότητας μεταβλητών

Όπως παρατηρούμε, μετά την πρώτη εφαρμογή του αλγορίθμου, η μεταβλητή ***Insight*** είναι η λιγότερο σημαντική μεταβλητή και είναι αυτή που θα αφαιρεθεί πρώτη στη συνέχεια (Εικόνα 122).



Εικόνα 122

Επαναλαμβάνοντας την παραπάνω διαδικασία, στόχος μας είναι να δημιουργηθούν συστάδες με καλό (Good) Cluster Quality. Μετά την διαδοχική αφαίρεση 57 μεταβλητών συνολικά μετά από κάθε εφαρμογή του αλγόριθμου προέκυψαν 2 συστάδες με καλή ποιότητα συστάδων (Εικόνα 123).

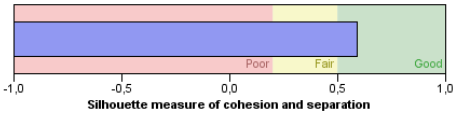
```
TWOSTEP CLUSTER
  /CONTINUOUS VARIABLES=Social Humans Negemo Anger Bio Health Sexual Home
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 10 BIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES
  /SAVE VARIABLE=TSC_3133.
```

TwoStep Cluster

Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	2

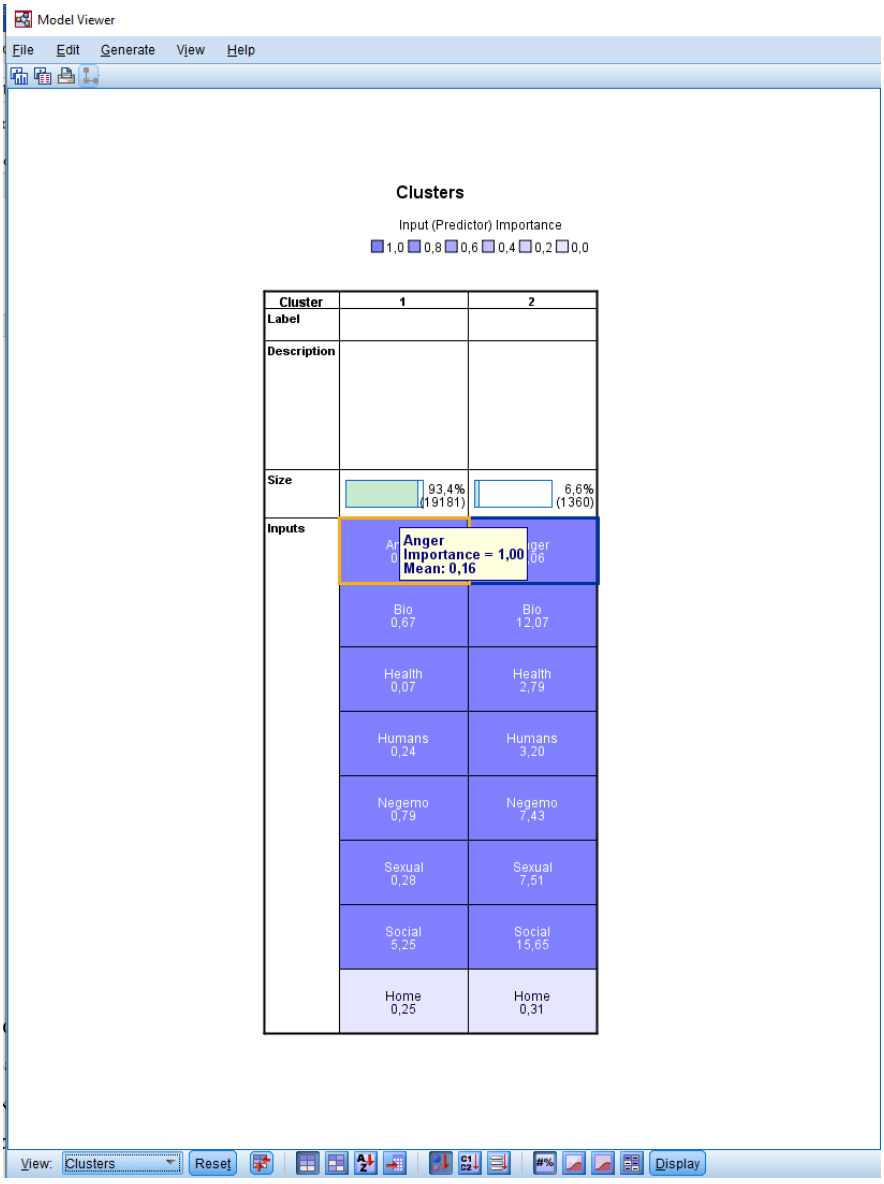
Cluster Quality



Εικόνα 123

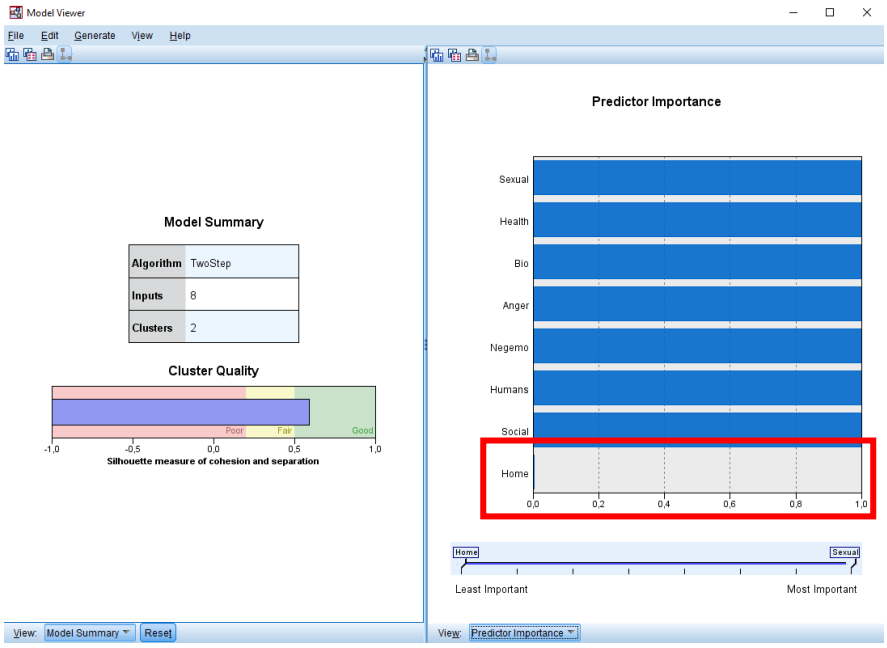
Οι 8 μεταβλητές που κρίθηκαν σημαντικές από τον αλγόριθμο για τη δημιουργία των συστάδων είναι οι: **Sexual, Health, Bio, Anger, Negemo, Humans, Social και Home**.

Στο αριστερό τμήμα του Model Viewer επιλέγοντας στο κάτω μέρος View → Clusters εμφανίζονται οι μεταβλητές που χρησιμοποιήθηκαν για τη δημιουργία των συστάδων με τον αντίστοιχο συντελεστή σημαντικότητας για την κάθε συστάδα και την αντίστοιχη μέση τιμή (Εικόνα 124).



Εικόνα 124. Πλέγμα μεταβλητών που συμμετέχουν στην δημιουργία των συστάδων (Audi)

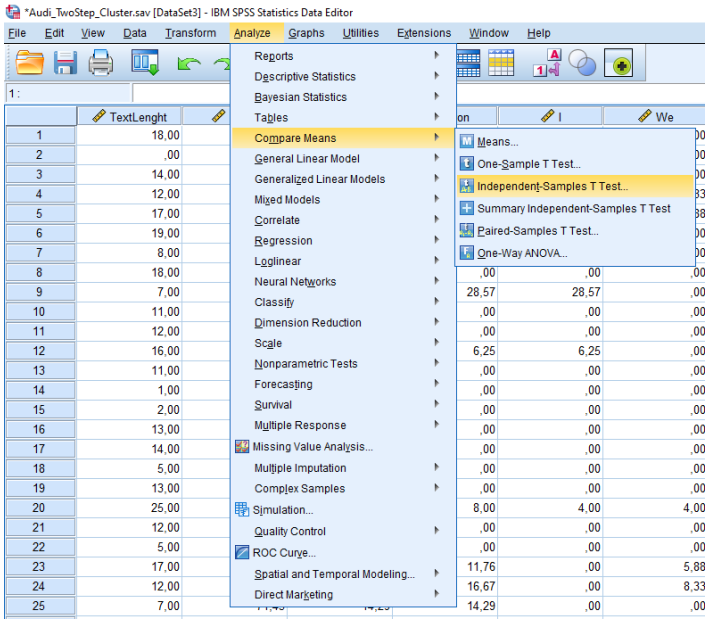
Εξετάζοντας εκ νέου την σημαντικότητα των μεταβλητών διαπιστώσαμε ότι παρά το υψηλό Cluster Quality η μεταβλητή Home φάνηκε αρχικά να μην συμμετέχει σημαντικά στην δημιουργία των συστάδων (Εικόνα 125).



Εικόνα 125

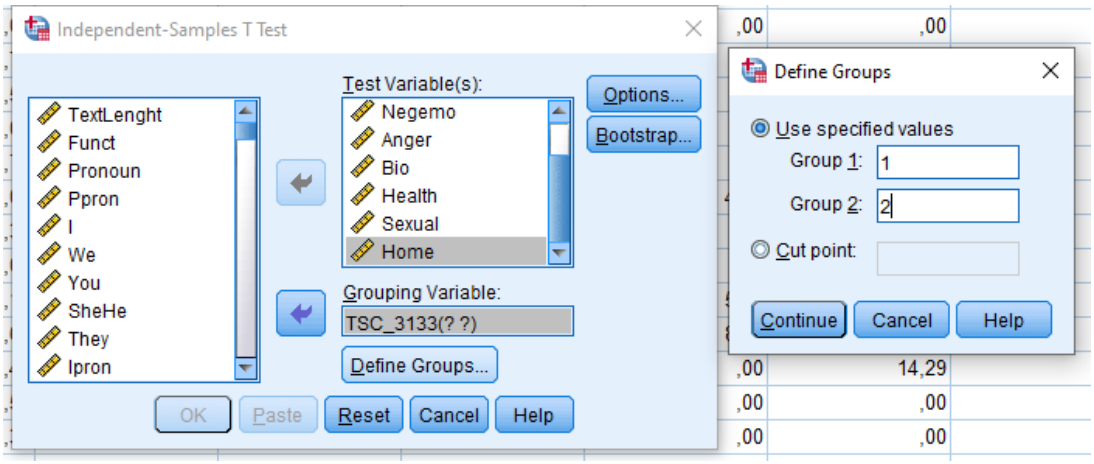
Στη συνέχεια, και προκειμένου να εξετάσουμε την διαφοροποίηση των ποσοτικών μεταβλητών ως προς τις δύο συστάδες διεξαγάγαμε έναν μη συσχετισμένο έλεγχο t-test. Ο έλεγχος αυτός χρησιμοποιείται σε περιπτώσεις προβλημάτων, στα οποία θέλουμε να συγκρίνουμε τις μέσες τιμές για την ίδια συνεχή μεταβλητή, δύο δειγμάτων που δεν είναι συζευγμένα [176]. Η ανάλυση αυτή δείχνει αν υπάρχει στατιστικά σημαντική διαφορά για κάποια από τις επιλεγμένες μεταβλητές από τον αλγόριθμο, έτσι ώστε προβούμε στην αφαίρεσή της.

Από το μενού του προγράμματος SPSS επιλέξαμε *Analyze* → *Compare Means* → *Indepedence-Samples T Test…* (Εικόνα 126)



Εικόνα 126. Διεξαγωγή μη συσχετισμένου ελέγχου t-test

Στο παράθυρο που εμφανίστηκε τοποθετήσαμε στο Test Variables(s) τις ποσοτικές μεταβλητές που χρησιμοποιήσαμε στο τελικό μας μοντέλο και στο Grouping Variable(s) την μεταβλητή των συστάδων που δημιούργησε ο αλγόριθμός (TSC_3133). Από το παράθυρο διαλόγου ***Define Groups*** καθορίσαμε τις ομάδες, δηλαδή τις συστάδες, στις οποίες θα πραγματοποιηθεί ο έλεγχος t-test (Εικόνα 127).



Εικόνα 127. Καθορισμός παραμέτρων ανάλυσης μη συσχετισμένου ελέγχου t-test

Τα αποτελέσματα του ελέγχου δίνονται στην Εικόνα. Στον πρώτο από τους δύο πίνακες (Group Statistics), εμφανίζονται τα περιγραφικά στατιστικά (πλήθος, μέση τιμή, τυπική απόκλιση και τυπικό σφάλμα) και των δύο υπό εξέταση συστάδων (Εικόνα 128). Στον δεύτερο πίνακα (Independent Samples Test) εμφανίζονται δύο διαστήματα εμπιστοσύνης και δύο τιμές p-value, Sig και Sig. (2-tailed) για την κάθε μεταβλητή (Εικόνα 129).

T-Test

Group Statistics

	TwoStep Cluster Number	N	Mean	Std. Deviation	Std. Error Mean
Social	1	19181	5,2505	7,81858	,05645
	2	1360	15,6486	16,76704	,45466
Humans	1	19181	,2385	1,26899	,00916
	2	1360	3,1981	7,69936	,20878
Negemo	1	19181	,7944	2,65082	,01914
	2	1360	7,4252	13,35745	,36220
Anger	1	19181	,1570	1,01821	,00735
	2	1360	4,0596	9,61799	,26080
Bio	1	19181	,6699	2,36312	,01706
	2	1360	12,0677	14,70711	,39880
Health	1	19181	,0694	,62304	,00450
	2	1360	2,7936	6,95242	,18852
Sexual	1	19181	,2794	1,43390	,01035
	2	1360	7,5096	12,74041	,34547
Home	1	19181	,2492	1,66734	,01204
	2	1360	,3096	3,19372	,08660

Εικόνα 128. Περιγραφικά στατιστικά μη συσχετισμένου ελέγχου t-test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Social	Equal variances assumed	1870,312	,000	-42,593	20539	,000	-10,39811	,24413	-10,87662	-9,91960
	Equal variances not assumed			-22,696	1401,204	,000	-10,39811	,45815	-11,29684	-9,49937
Humans	Equal variances assumed	7743,902	,000	-45,278	20539	,000	-2,95962	,06537	-3,08775	-2,83150
	Equal variances not assumed			-14,162	1364,240	,000	-2,95962	,20898	-3,36958	-2,54967
Negemo	Equal variances assumed	7956,043	,000	-55,136	20539	,000	-6,63078	,12026	-6,86651	-6,39506
	Equal variances not assumed			-18,281	1366,600	,000	-6,63078	,36271	-7,34231	-5,91926
Anger	Equal variances assumed	9976,703	,000	-52,234	20539	,000	-3,90258	,07471	-4,04903	-3,75614
	Equal variances not assumed			-14,958	1361,161	,000	-3,90258	,26091	-4,41441	-3,39076
Bio	Equal variances assumed	10141,770	,000	-91,918	20539	,000	-11,39780	,12400	-11,64085	-11,15475
	Equal variances not assumed			-28,554	1363,980	,000	-11,39780	,39917	-12,18085	-10,61475
Health	Equal variances assumed	10175,559	,000	-51,448	20539	,000	-2,72425	,05295	-2,82804	-2,62047
	Equal variances not assumed			-14,446	1360,548	,000	-2,72425	,18858	-3,09419	-2,35432
Sexual	Equal variances assumed	16492,692	,000	-72,415	20539	,000	-7,23023	,09984	-7,42593	-7,03452
	Equal variances not assumed			-20,919	1361,442	,000	-7,23023	,34563	-7,90825	-6,55221
Home	Equal variances assumed	6,614	,010	-1,189	20539	,234	-,06034	,05075	-,15982	,03913
	Equal variances not assumed			-,690	1411,996	,490	-,06034	,08743	-,23186	,11117

Εικόνα 129. Πίνακας αποτελεσμάτων ελέγχου ανεξαρτησίας του δείγματος (Audi)

Εδώ θα πρέπει να επισημάνουμε ότι για να πραγματοποιηθεί ένα μη συσχετισμένος έλεγχος t-test, χρειάζεται να προχωρήσουμε σε κάποιες παραδοχές. Μία από αυτές είναι η ομοιογένεια των τιμών των δύο συστάδων. Για να ελέγξουμε την ισότητα ή μη των διακυμάνσεων των δυο συστάδων χρησιμοποιούμε τον έλεγχο Levene,, ο οποίο πραγματοποιείται αυτόματα με την υλοποίηση μη συσχετισμένου ελέγχου t-test. Η μηδενική και εναλλακτική υπόθεση του ελέγχου Levene έχουν ως εξής:

- Μηδενική Υπόθεση H_0 : Οι διακυμάνσεις των δύο ομάδων είναι ίσες.
- Εναλλακτική Υπόθεση H_1: Οι διακυμάνσεις των δύο ομάδων δεν είναι ίσες.

Στον δεύτερο πίνακα παρατηρούμε ότι ο έλεγχος Levene δίνει για όλες τις μεταβλητές p-value = 0 < 0,05 και κατά συνέπεια η μηδενική υπόθεση για την ισότητα των διακυμάνσεων απορρίπτεται και άρα, η επιλογή διαστήματος εμπιστοσύνης και p-value για τον μη συσχετισμένο έλεγχο t-test θα γίνει για την κάθε μεταβλητή, από τη δεύτερη γραμμή (Equal variances not assumed). Στην περίπτωση της μεταβλητής Home παρατηρούμε ότι εφόσον για τον μη συσχετισμένο έλεγχο t-test, προέκυψε ότι p-value = 0,490, η μηδενική υπόθεση γίνεται αποδεκτή σε επίπεδο σημαντικότητας 0,05 και κατά συνέπεια, μπορούμε να συμπεράνουμε πως δεν υπάρχει στατιστικά σημαντική διαφορά

σε επίπεδο σημαντικότητας 0,05 και άρα μπορούμε να την αφαιρέσουμε από το μοντέλο μας, τρέχοντας εκ νέου τον αλγόριθμο. Ωστόσο, η τιμή του στατιστικού t είναι -0.690 και καθώς βρίσκεται εκτός των ορίων που καθορίζουν το διάστημα εμπιστοσύνης 95% (-0,23186 , 0,11117), η μηδενική υπόθεση δεν γίνεται αποδεκτή σε επίπεδο σημαντικότητας 0,05.

Αφαιρώντας την μεταβλητή Home και τρέχοντας εκ νέου τον αλγόριθμο η ποιότητα του μοντέλου μειώθηκε σημαντικά (Εικόνα 130).

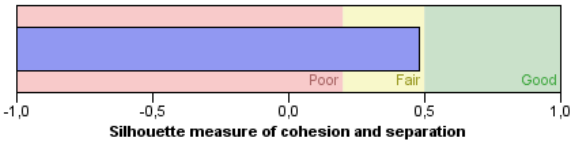
```
TWOSTEP CLUSTER
  /CONTINUOUS VARIABLES=Social Humans Negemo Anger Bio Health Sexual
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 10 BIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES
  /SAVE VARIABLE=TSC_7083.
```

TwoStep Cluster

Model Summary

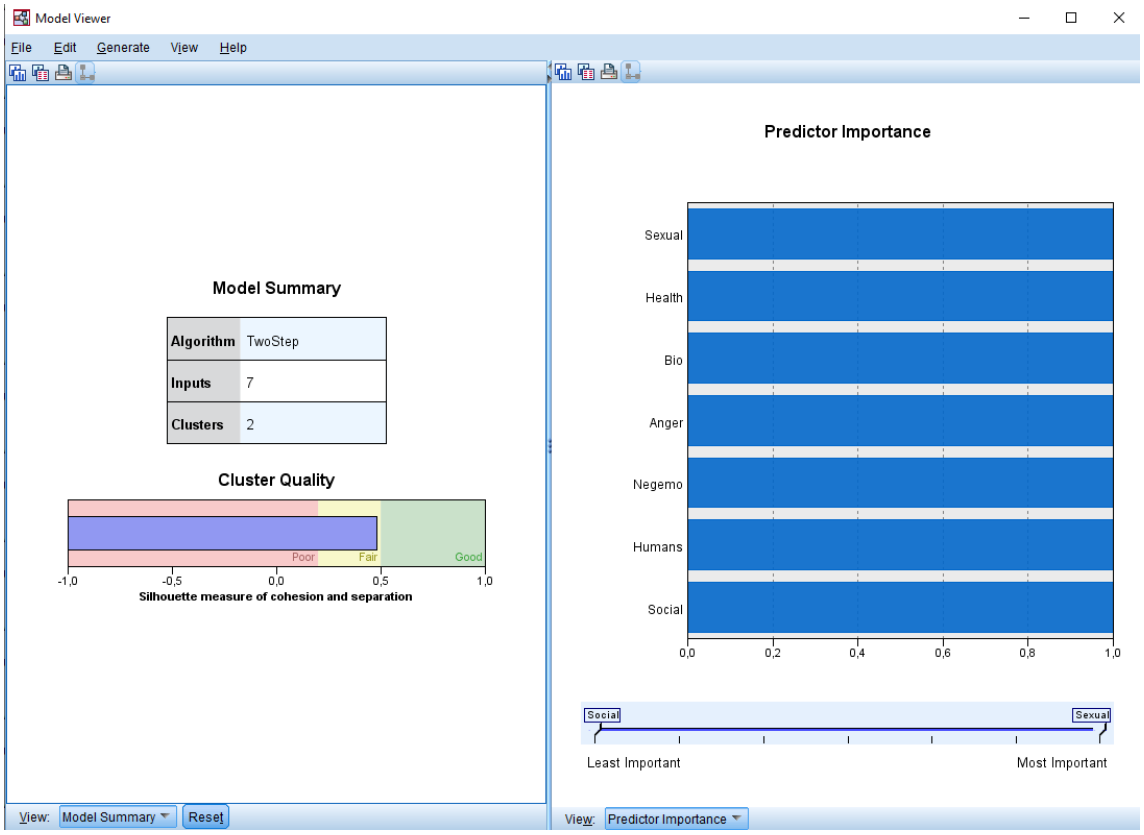
Algorithm	TwoStep
Inputs	7
Clusters	2

Cluster Quality



Εικόνα 130

Εξετάζοντας και πάλι την σημαντικότητα των μεταβλητών προκειμένου να αφαιρέσουμε την λιγότερο σημαντική και να βελτιώσουμε με αυτό τον τρόπο την ποιότητα της συσταδοποίησης, διαπιστώσαμε ότι όλες οι εναπομείναντες μεταβλητές ήταν εξίσου σημαντικές (Εικόνα 131).



Εικόνα 131

Ως εκ τούτου για το επόμενο βήμα της μελέτης μας θα χρησιμοποιήσουμε τα αποτελέσματα της συσταδοποίησης, σύμφωνα με την οποία προέκυψαν 2 συστάδες βάσει των 8 μεταβλητών: **Sexual, Health, Bio, Anger, Negemo, Humans, Social και Home.**

6.3.2.1.1. Ερμηνεία αποτελεσμάτων

Οι 8 μεταβλητές που συμμετείχαν στην δημιουργία των συστάδων μετά την εφαρμογή του αλγόριθμου είναι οι **Sexual, Health, Bio, Anger, Negemo, Humans, Social** και **Home.** Κάθε μία από αυτές τις μεταβλητές αντιστοιχεί σε μία κατηγορία του λεξικού LIWC.

Σύμφωνα με τον πίνακα των κατηγοριών αυτών η μεταβλητή **Social** ανήκει στην κατηγορία των ψυχολογικών διαδικασιών (Psychological Processes) και περιλαμβάνει ένα μεγάλο αριθμό λέξεων που υποδηλώνουν κοινωνικές διεργασίες, συμπεριλαμβανομένων όλων των προσωπικών αντωνυμιών που δεν είναι πρώτου προσώπου, καθώς και τα ρήματα που υποδηλώνουν ανθρώπινη αλληλεπίδραση, όπως για παράδειγμα οι λέξεις talking (ομιλία) και sharing (κοινή χρήση). Στην κατηγορία των ψυχολογικών διαδικασιών ανήκουν επίσης και οι μεταβλητές Humans, Anger, Negemo,

Bio, Health και Sexual. Η μεταβλητή Humans σχετίζεται με λέξεις που αφορούν την ανθρώπινη φύση, όπως adult (ενήλικας), boy (αγόρι), girl (κορίτσι). Η μεταβλητή Anger περιλαμβάνει λέξεις που δηλώνουν θυμό, όπως hate (μίσος), kill (σκοτωμός) και annoyed (ενοχλημένος, εκνευρισμένος), ενώ η μεταβλητή Negemo αναφέρεται σε λέξεις που υποδηλώνουν αρνητικά συναισθήματα, όπως hurt (πληγωμένος), ugly (άσχημος), nasty (μοχθηρός, άθλιος). Η μεταβλητή Bio αφορά λέξεις που περιγράφουν βιολογικές διεργασίες, όπως eat (τρώω), blood (αίμα) και pain (πόνος). Η μεταβλητή Health περιλαμβάνει λέξεις που αφορούν την υγεία, όπως clinic (κλινική), flu (γρίπη) και pill (χάπι), ενώ η μεταβλητή Sexual αφορά λέξεις σχετικά με την σεξουαλικότητα, όπως love (αγάπη) και incest (αιμομιξία). Τέλος, η μεταβλητή Home ανήκει στην ευρύτερη κατηγορία των Προσωπικών Ανησυχιών (Personal Concerns) και σχετίζεται με λέξεις όπως family (οικογένεια), kitchen (κουζίνα) και apartment (διαμέρισμα).

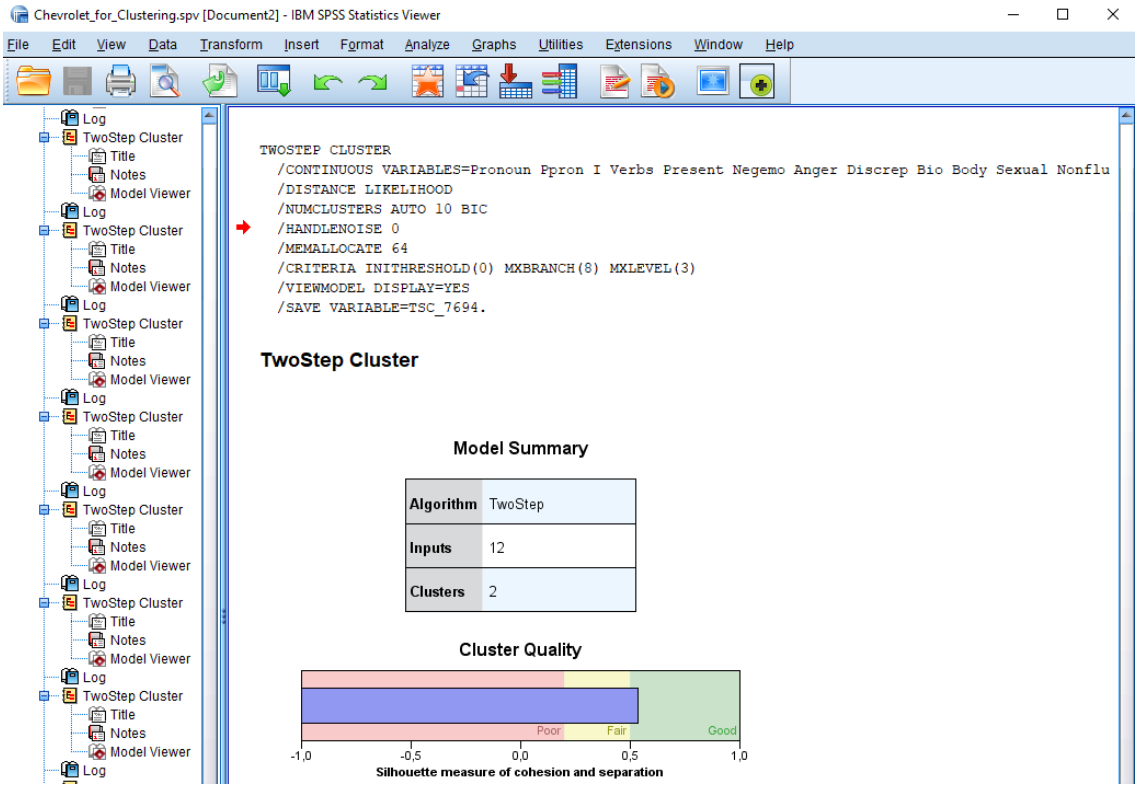
Προσπαθώντας να σκιαγραφήσουμε το προφίλ των χρηστών του Twitter μέσω των μηνυμάτων τους για την συγκεκριμένη αυτοκινητοβιομηχανία, θα μπορούσαμε να τους κατατάξουμε αρχικά στην κατηγορία των συμμετεχόντων (joiners) ή των συνομιλητών σύμφωνα με την κατηγοριοποίηση της Forrester Research, καθώς πρόκειται για χρήστες που συμμετέχουν ενεργά στα κοινωνικά μέσα (μεταβλητή social), εκφράζονται έντονα χρησιμοποιώντας λέξεις που δηλώνουν θυμό (anger) και αρνητικότητα (negemo).

Όσον αφορά την κατηγοριοποίηση σύμφωνα με το μοντέλο OCEAN, οι χρήστες θα μπορούσαν να ανήκουν στην κατηγορία τόσο των νευρωτικών προσωπικοτήτων, λόγω της εκδήλωσης αρνητικών συναισθημάτων, όσο και των εξωστρεφών καθώς εμφανίζονται δραστήριοι συμμετέχοντας σε συνομιλίες στο διαδίκτυο με ευκολία.

Στη συνέχεια της μελέτης μας εφαρμόσαμε τον αλγόριθμο συσταδοποίησης δύο βημάτων και στις υπόλοιπες βάσεις δεδομένων που αφορούσαν τα tweets των χρηστών για τις αυτοκινητοβιομηχανίες Chrysler, Chevrolet, KIA και Volkswagen, επιχειρώντας να ερμηνεύσουμε στη συνέχεια τα αποτελέσματα της ανάλυσης.

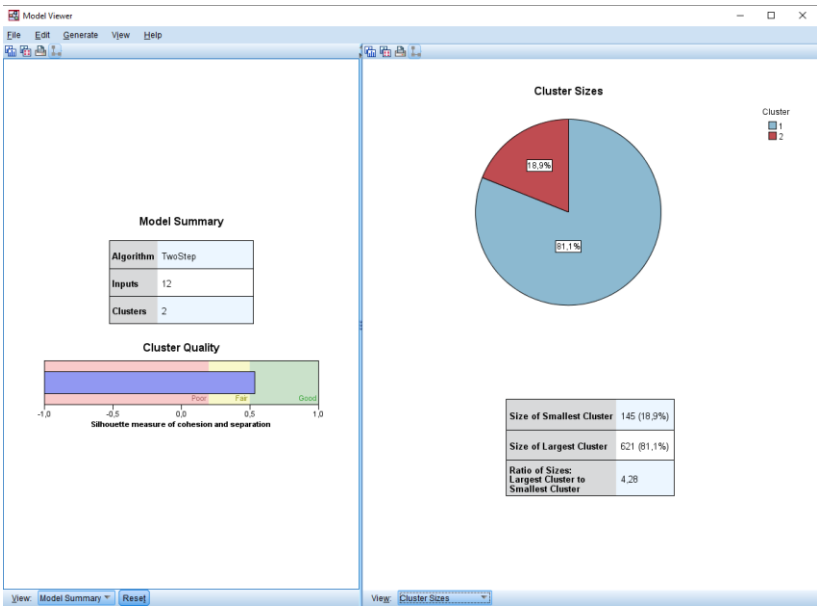
6.3.2.2.*Chevrolet*

Η βάση δεδομένων της εταιρείας Chevrolet αποτελούνταν από 766 εγγραφές και 65 μεταβλητές εισόδου του λεξικού LIWC2007. Η εφαρμογή του αλγόριθμου οδήγησε στη δημιουργία ενός μοντέλου 2 συστάδων καλής ποιότητας με συνολικά 12 μεταβλητές εισόδου (Εικόνα), τις: **Pronoun, Ppron, I, Verbs, Present, Negemo, Anger, Discrep, Bio, Body, Sexual** και **Nonflu**. Η πρώτη συστάδα αποτελεί το **18,9**% του δείγματος, ενώ η **δεύτερη** το **81,1**% (Εικόνα 132).

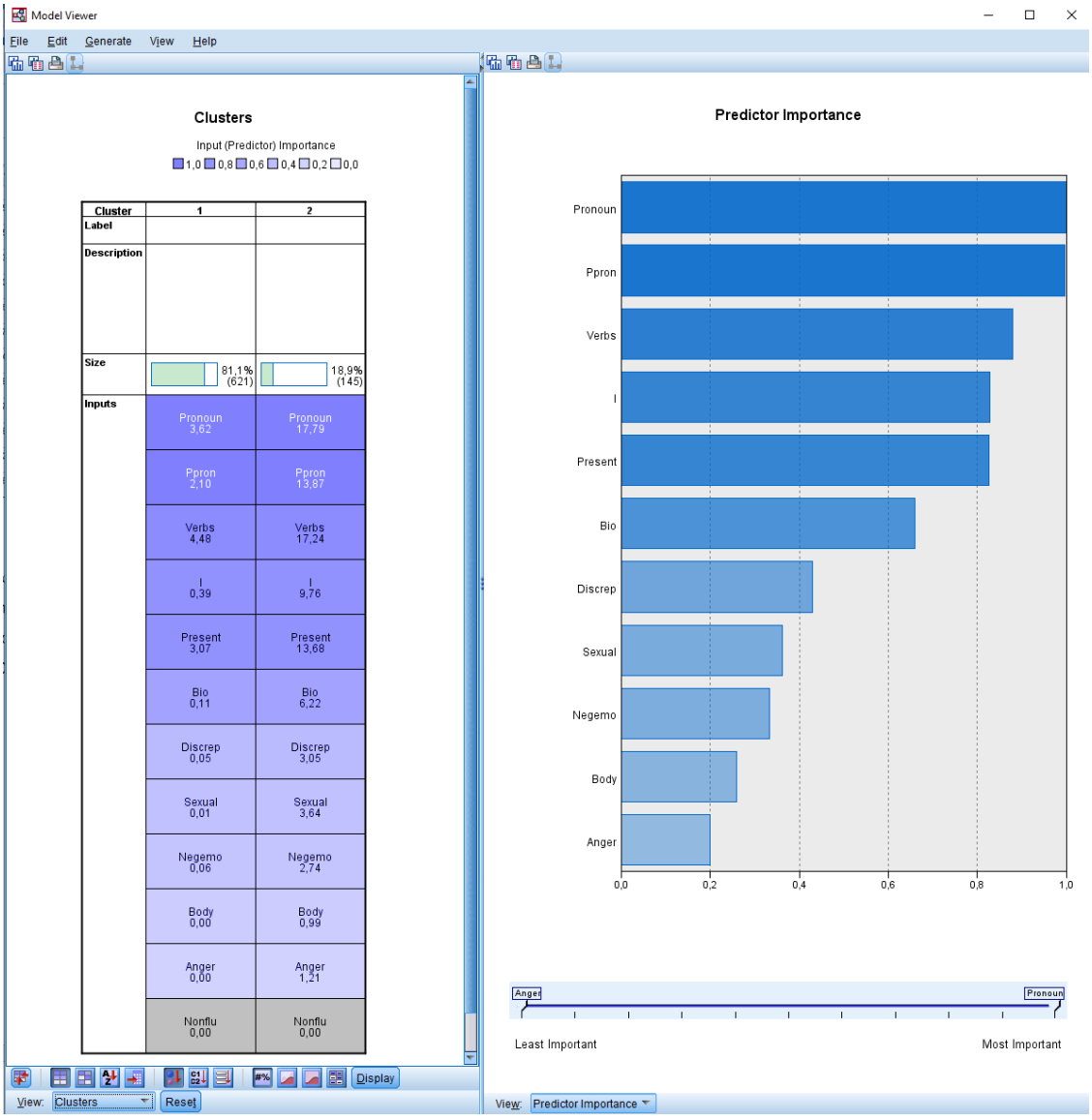


Εικόνα 132. Περίληψη μοντέλου και ποιότητα συσταδοποίησης (Chevrolet)

Στο ραβδόγραμμα των σημαντικότερων μεταβλητών για τη δημιουργία των συστάδων (Εικόνα), παρατηρούμε ότι οι μεταβλητές Anger, Body, Negemo, Sexual και Discrep, δεν συμμετέχουν σημαντικά στην δημιουργία του μοντέλου, ενώ η μεταβλητή Nonflu δεν εμφανίζεται καθόλου, καθώς δεν υπάρχει εγγραφή που να έχει μετρήσιμη τιμή γι' αυτή την μεταβλητή, όπως φαίνεται και από τα περιγραφικά στατιστικά του μοντέλου (Εικόνα 134).



Εικόνα 133. Παράθυρο προβολής μοντέλων



Εικόνα 134

Από τον μη συσχετισμένο έλεγχο t-test για τις 12 μεταβλητές του τελικού μοντέλου (Εικόνα 136) παρατηρούμε ότι το Levene Test για όλες τις μεταβλητές δίνει p-value = 0 < 0,05 και έτσι κι εδώ η μηδενική υπόθεση για την ισότητα των διακυμάνσεων απορρίπτεται και συνεπώς, η επιλογή διαστήματος εμπιστοσύνης και p-value για τον μη συσχετισμένο έλεγχο t-test θα γίνει από τη δεύτερη γραμμή (Equal variances not assumed) για την κάθε μεταβλητή. Σε κάθε περίπτωση, παρατηρούμε ότι p-value = 0 <0,05 και άρα όλες οι ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες, γεγονός που δηλώνει ότι δεν χρειάζεται να αφαιρέσουμε κάποια από τις μεταβλητές **Anger, Body, Negemo, Sexual** και **Discrep**. Η μεταβλητή **Nonflu** και πάλι δεν εμφανίζεται στον πίνακα ελέγχου ανεξαρτησίας και κατά συνέπεια μπορεί να αφαιρεθεί.

```
T-TEST GROUPS=TSC_7694(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=Pronoun Ppron I Verbs Present Negemo Anger Discrep Bio Body Sexual Nonflu
  /CRITERIA=CI(.95).
```

Group Statistics

	TwoStep Cluster Number	N	Mean	Std. Deviation	Std. Error Mean
Pronoun	1	621	3,6167	6,17191	,24767
	2	145	17,7877	16,87113	1,40107
Ppron	1	621	2,0984	4,31207	,17304
	2	145	13,8738	15,18820	1,26131
I	1	621	,3944	1,93701	,07773
	2	145	9,7610	15,11909	1,25557
Verbs	1	621	4,4776	6,93263	,27820
	2	145	17,2422	14,71945	1,22238
Present	1	621	3,0684	5,20123	,20872
	2	145	13,6810	14,10116	1,17104
Negemo	1	621	,0648	,73833	,02963
	2	145	2,7448	7,30260	,60645
Anger	1	621	,0000	,00000	,00000
	2	145	1,2077	4,45961	,37035
Discrep	1	621	,0508	,57670	,02314
	2	145	3,0506	7,16835	,59530
Bio	1	621	,1076	,85206	,03419
	2	145	6,2217	11,48899	,95411
Body	1	621	,0000	,00000	,00000
	2	145	,9912	3,16827	,26311
Sexual	1	621	,0115	,28652	,01150
	2	145	3,6354	9,61541	,79852
Nonflu	1	621	,0000	,00000[a]	,00000
	2	145	,0000	,00000[a]	,00000

a. t cannot be computed because the standard deviations of both groups are 0.

Εικόνα 135. Περιγραφικά στατιστικά μη συσχετισμένου ελέγχου t-test

T-Test

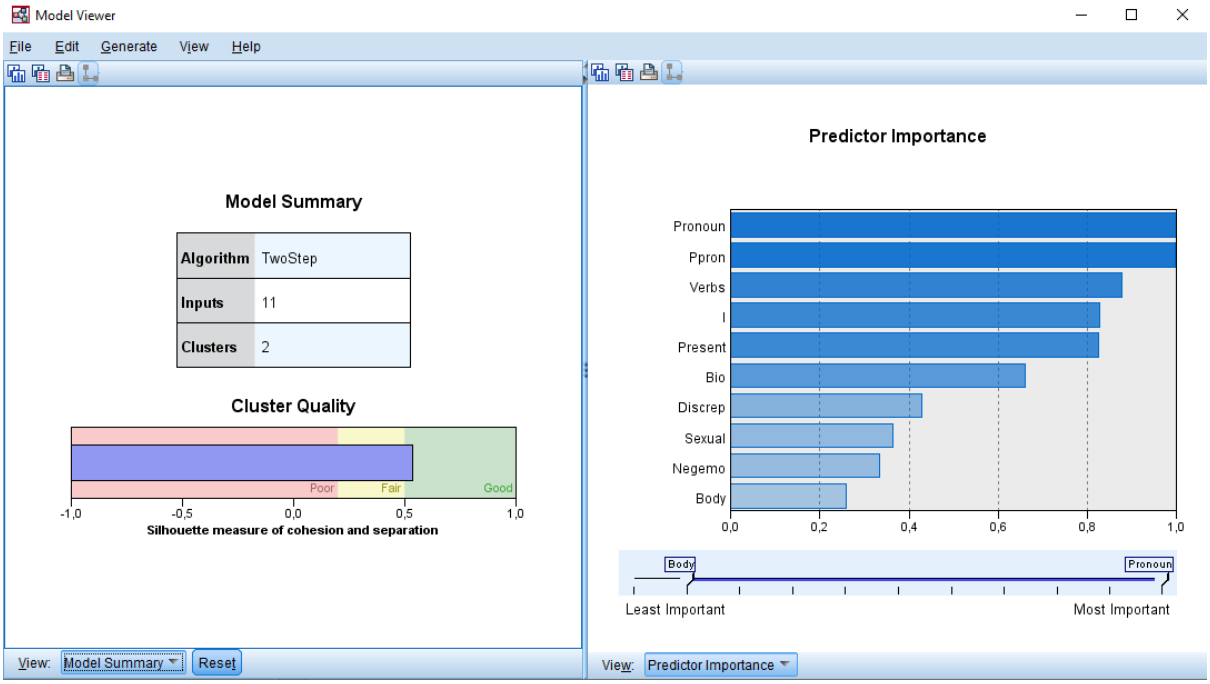
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pronoun	Equal variances assumed	265,678	,000	-16,708	764	,000	-14,17099	,84815	-15,83596	-12,50602
	Equal variances not assumed			-9,960	153,105	,000	-14,17099	1,42279	-16,98183	-11,36015
Ppron	Equal variances assumed	361,216	,000	-16,682	764	,000	-11,77544	,70586	-13,16109	-10,38979
	Equal variances not assumed			-9,249	149,459	,000	-11,77544	1,27313	-14,29109	-9,25979
I	Equal variances assumed	682,610	,000	-14,952	764	,000	-9,36659	,62643	-10,59632	-8,13686
	Equal variances not assumed			-7,446	145,105	,000	-9,36659	1,25798	-11,85291	-6,88026
Verbs	Equal variances assumed	185,134	,000	-15,489	764	,000	-12,76462	,82413	-14,38244	-11,14680
	Equal variances not assumed			-10,182	159,204	,000	-12,76462	1,25364	-15,24053	-10,28871
Present	Equal variances assumed	300,595	,000	-14,925	764	,000	-10,61263	,71104	-12,00846	-9,21680
	Equal variances not assumed			-8,922	153,258	,000	-10,61263	1,18949	-12,96255	-8,26271
Negemo	Equal variances assumed	358,660	,000	-8,970	764	,000	-2,68000	,29878	-3,26652	-2,09347
	Equal variances not assumed			-4,414	144,688	,000	-2,68000	,60717	-3,88007	-1,47993
Anger	Equal variances assumed	200,778	,000	-6,763	764	,000	-1,20772	,17857	-1,55828	-,85717
	Equal variances not assumed			-3,261	144,000	,001	-1,20772	,37035	-1,93975	-,47570
Discrep	Equal variances assumed	464,610	,000	-10,308	764	,000	-2,99985	,29101	-3,57112	-2,42858
	Equal variances not assumed			-5,035	144,435	,000	-2,99985	,59575	-4,17736	-1,82234
Bio	Equal variances assumed	638,624	,000	-13,136	764	,000	-6,11407	,46546	-7,02781	-5,20034
	Equal variances not assumed			-6,404	144,370	,000	-6,11407	,95472	-8,00111	-4,22703
Body	Equal variances assumed	293,830	,000	-7,813	764	,000	-,99117	,12686	-1,24022	-,74213
	Equal variances not assumed			-3,767	144,000	,000	-,99117	,26311	-1,51123	-,47111
Sexual	Equal variances assumed	397,684	,000	-9,394	764	,000	-3,62388	,38576	-4,38115	-2,86661
	Equal variances not assumed			-4,538	144,060	,000	-3,62388	,79860	-5,20237	-2,04540

Εικόνα 136. Πίνακας αποτελεσμάτων ελέγχου ανεξαρτησίας του δείγματος (Chevrolet)

Αφαιρώντας την μεταβλητή Nonflu και τρέχοντας εκ νέου τον αλγόριθμο η ποιότητα του μοντέλου εξακολουθεί να είναι καλή ενώ δεν παρατηρείται κάποια αλλαγή και στην σειρά σημαντικότητας των μεταβλητών.

Ως εκ τούτου για το επόμενο βήμα της μελέτης μας θα χρησιμοποιήσουμε τα αποτελέσματα της συσταδοποίησης, σύμφωνα με την οποία προέκυψαν 2 συστάδες βάσει των 11 μεταβλητών: **Pronoun, Ppron, I, Verbs, Present, Negemo, Anger, Discrep, Bio, Body** και **Sexual** (Εικόνα 137).



Εικόνα 137. Απεικόνιση μοντέλου μετά την αφαίρεση της μεταβλητής Nonflu

6.3.2.2.1. Ερμηνεία αποτελεσμάτων

Οι 11 μεταβλητές που συμμετείχαν στην δημιουργία των συστάδων μετά την εφαρμογή του αλγόριθμου είναι οι **Pronoun, Ppron, I, Verbs, Present, Negemo, Anger, Discrep, Bio, Body** και **Sexual**.

Οι μεταβλητές Pronoun, Ppron και I ανήκουν στην γενική κατηγορία των γλωσσικών διεργασιών (Linguistic Processes) και περιλαμβάνουν τις προσωπικές αντωνυμίες όλων των προσώπων. Στην ίδια γενική κατηγορία ανήκουν και οι μεταβλητές Verbs και Present. Η μεταβλητή Verb αναφέρεται στα κοινά (συνηθισμένα) ρήματα που χρησιμοποιούνται στις καθημερινές εκφράσεις, όπως walk (περπατώ), went (πήγα), see (βλέπω), ενώ η μεταβλητή Present αφορά στον ενεστωτικό χρόνο των ρημάτων, δηλαδή στο παρόν.

Οι υπόλοιπες μεταβλητές ανήκουν στην κατηγορία των ψυχολογικών διαδικασιών (Psychological Processes). Όπως αναφέραμε και παραπάνω οι μεταβλητές Negemo και Anger περιλαμβάνουν λέξεις που δηλώνουν έντονα αρνητικά συναισθήματα, ενώ οι μεταβλητές Bio, Body και Sexual αναφέρονται σε βιολογικές διεργασίες και περιλαμβάνουν λέξεις που στα πλαίσια των μηνυμάτων των κοινωνικών μέσων χρησιμοποιούνται συνήθως σε προσβλητικές εκφράσεις. Η νέα μεταβλητή που εμφανίζεται εδώ είναι η Discrep (από την λέξη discrepancy = ασυνέπεια, ασυμφωνία, απόκλιση) η οποία περιλαμβάνει τα βοηθητικά ρήματα would, could, should (καλούνται

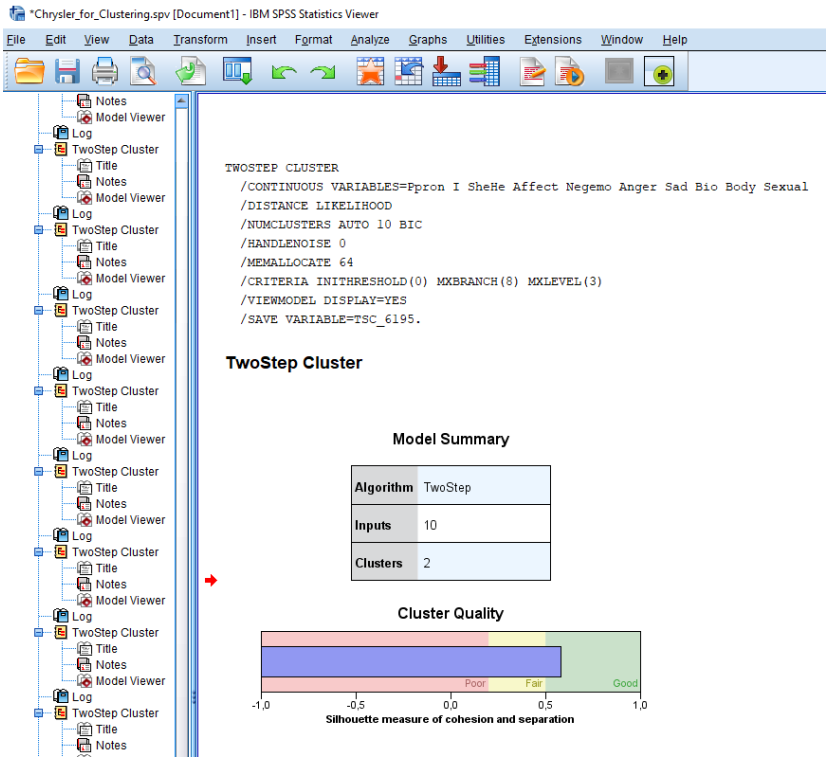
modal verbs). Τα ρήματα αυτά χρησιμοποιούνται συνήθως όταν θέλουμε να κάνουμε μια προσφορά ή να προσκαλέσουμε κάποιον, να διατυπώσουμε με ευγενικό τρόπο ένα αίτημα, να μιλήσουμε για προτιμήσεις ή επιθυμίες και να εκφράζουμε μια ικανότητα.

Μια εκτίμηση για το προφίλ των χρηστών που συμμετέχουν στον σχολιασμό της αυτοκινητοβιομηχανίας Chevrolet, είναι ότι θα μπορούσαν να στην κατηγορία των συνομιλητών (conversationalists) σύμφωνα με την κατηγοριοποίηση της Forrester Research, καθώς πρόκειται για χρήστες που ζουν στο παρόν, χρησιμοποιούν σε μεγάλο βαθμό τις προσωπικές αντωνυμίες εκφράζοντας και υπερασπίζοντας την άποψή τους, ενώ τα αρνητικά συναισθήματα που φαίνεται να έχουν δεν είναι τόσο έντονα και θα μπορούσαν να θεωρηθούν φυσικό επακόλουθο της διαδικασίας ανταλλαγής απόψεων κατά την διάρκεια μιας συνομιλίας. Επιπλέον η ύπαρξη των βοηθητικών ρημάτων would, could και should στα μηνύματά τους δηλώνουν έναν βαθμό ευγένειας μεταξύ των συνομιλητών.

Όσον αφορά την κατηγοριοποίηση σύμφωνα με το μοντέλο OCEAN, οι χρήστες θα μπορούσαν να ανήκουν στην κατηγορία των νευρωτικών. Η χρήση προσωπικών αντωνυμιών σε συνδυασμό με λέξεις που δηλώνουν αρνητισμό, νευρικότητα και ασυνέπεια αποτελούν χαρακτηριστικά των ανθρώπων αυτής της κατηγορίας.

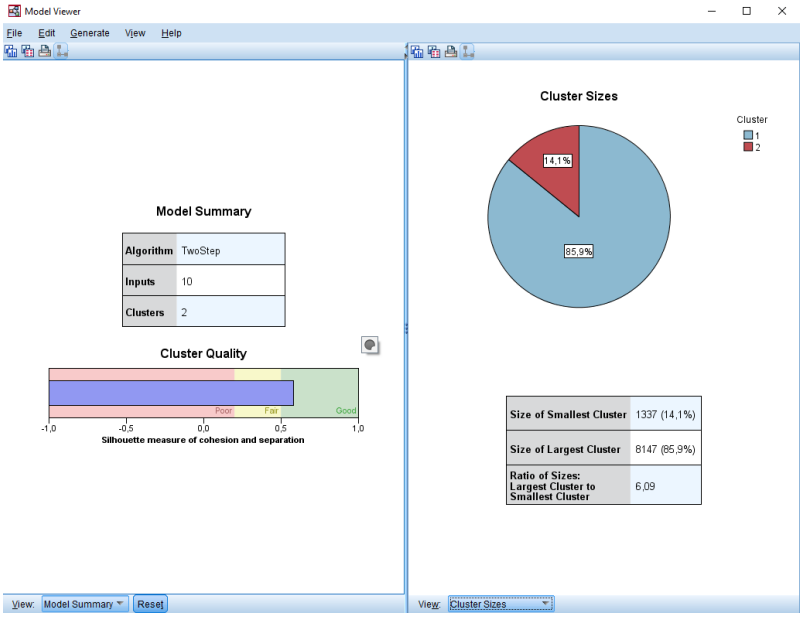
6.3.2.3. Chrysler

Η βάση δεδομένων των tweets της εταιρείας Chrysler, αποτελούνταν από 9484 εγγραφές, ενώ κι εδώ επιλέχθηκαν για την ανάλυση οι 65 μεταβλητές του λεξικού LIWC2007. Ακολουθώντας την ίδια διαδικασία με πριν, η εφαρμογή του αλγόριθμου οδήγησε στη δημιουργία ενός μοντέλου 2 συστάδων καλής ποιότητας με συνολικά 10 μεταβλητές εισόδου (Εικόνα 138), τις: **Bio, Anger, Negemo, Affect, Sad, Body, SheHe, Sexual, I, Ppron**.



Εικόνα 138

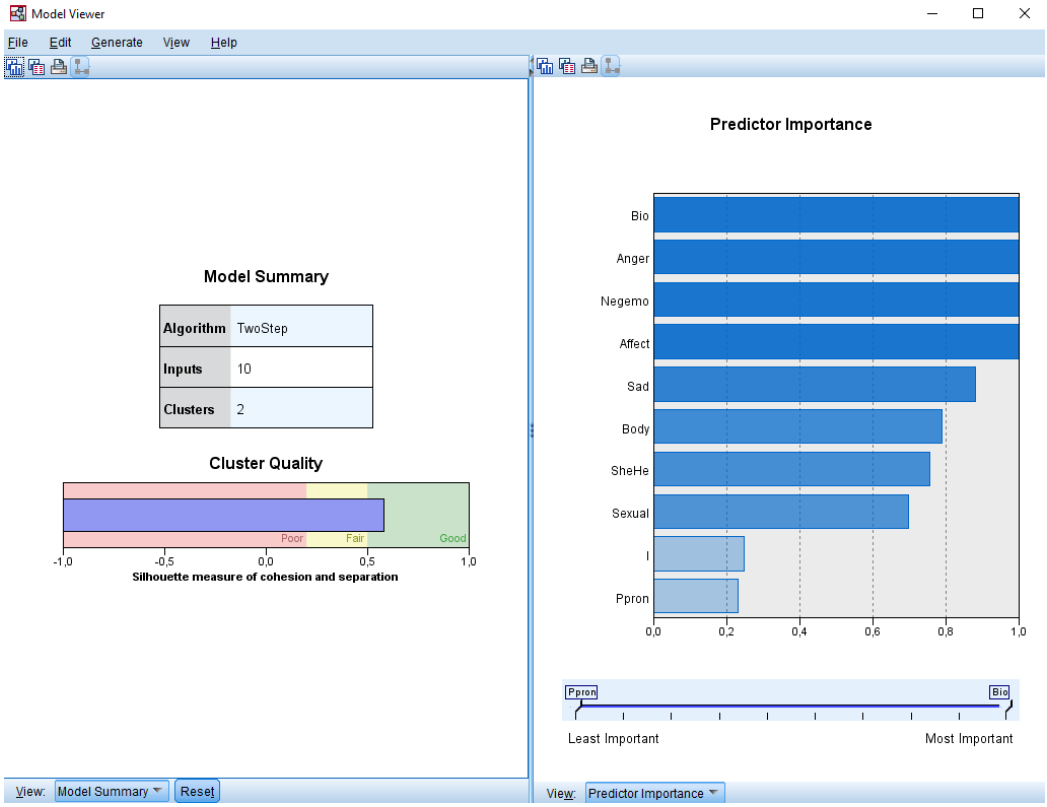
Ενεργοποιώντας το Model Viewer του διαγράμματος, παρατηρούμε την δημιουργία δύο συστάδων με την πρώτη να αποτελεί το 85,9% του δείγματος, ενώ η δεύτερη το 14,1% (Εικόνα 139).



Εικόνα 139. Παράθυρο προβολής μοντέλων

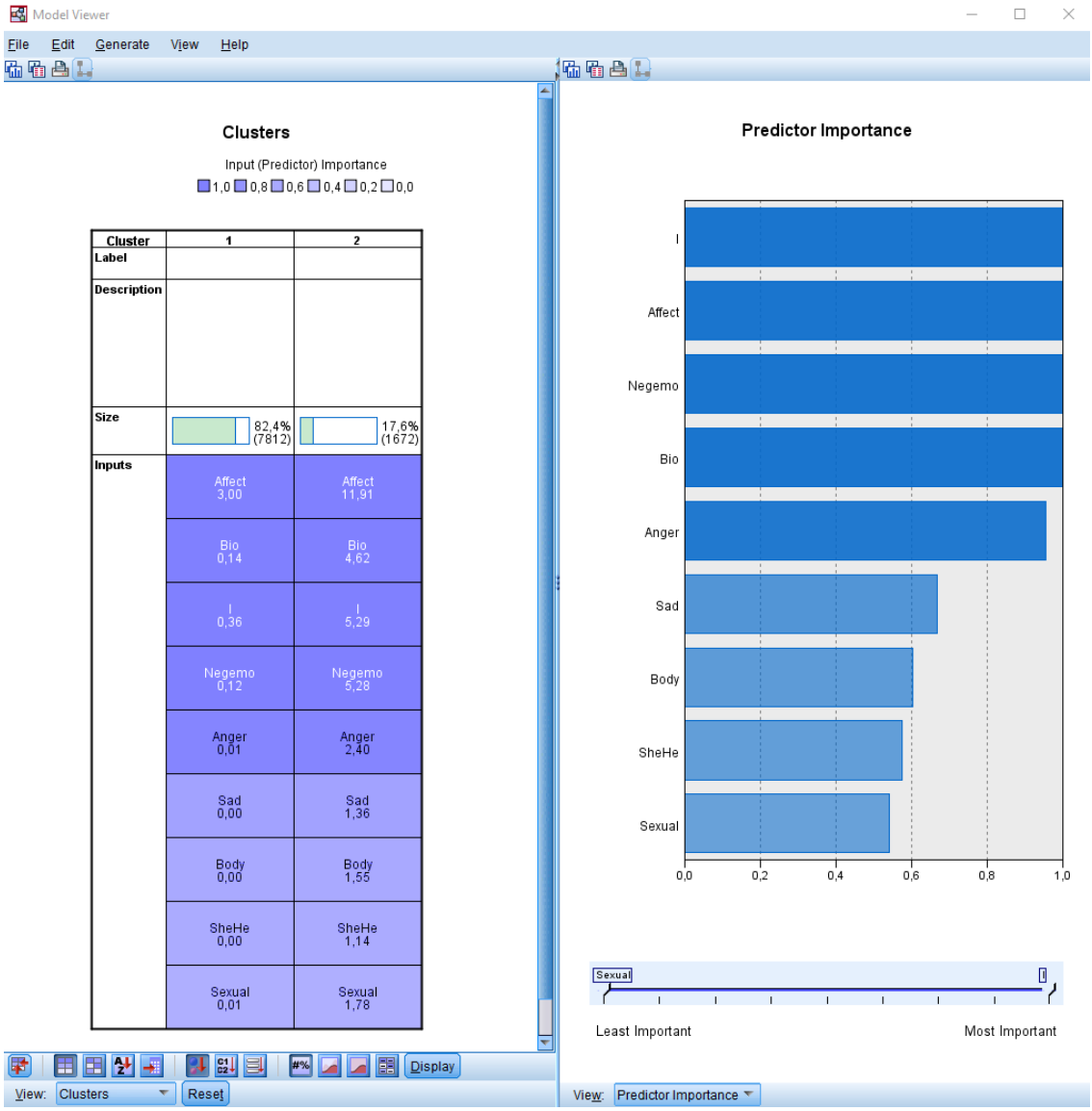
Στο ραβδόγραμμα των σημαντικότερων μεταβλητών για τη δημιουργία των συστάδων (Εικόνα 140), παρατηρούμε ότι οι μεταβλητές **I** και **Ppron**, οι οποίες ανήκουν στην κατηγορία των προσωπικών αντωνυμιών (Personal pronouns) δεν είναι πολύ

σημαντικές, και ως εκ τούτου μπορούμε να τις αφαιρέσουμε από την ανάλυση. Ωστόσο, κατά την ερμηνεία των αποτελεσμάτων, θα μπορούσαν πιθανότατα να δώσουν χρήσιμες πληροφορίες για το προφίλ των χρηστών.



Εικόνα 140

Προκειμένου να διαπιστώσουμε την ύπαρξη αυτής της πιθανότητας προχωρήσαμε στην αφαίρεση της μεταβλητής **Ppron** (Personal pronouns) και εφαρμόσαμε εκ νέου τον αλγόριθμο. Το αποτέλεσμα τώρα έδωσε και πάλι 2 συστάδες με καλή ποιότητα, ενώ στην σημαντικότητα τον μεταβλητών παρατηρούμε πλέον ότι η μεταβλητή **I** είναι από τις πιο σημαντικές (Εικόνα 141). Αυτό μας οδήγησε στο επόμενο βήμα, της διεξαγωγής ενός μη συσχετιστικού ελέγχου t-test.



Εικόνα 141

Από τον μη συσχετισμένο έλεγχο t-test για τις 10 μεταβλητές του τελικού μοντέλου (πριν την αφαίρεση της μεταβλητής Ppron (Εικόνα 142) παρατηρούμε ότι το Levene Test για όλες τις μεταβλητές δίνει p-value = 0 < 0,05 και έτσι η μηδενική υπόθεση για την ισότητα των διακυμάνσεων απορρίπτεται και συνεπώς, η επιλογή διαστήματος εμπιστοσύνης και p-value για τον μη συσχετισμένο έλεγχο t-test θα γίνει από τη δεύτερη γραμμή (Equal variances not assumed) για την κάθε μεταβλητή. Σε κάθε περίπτωση, παρατηρούμε ότι p-value = 0 <0,05 και άρα όλες οι ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες, γεγονός που δηλώνει ότι δεν χρειάζεται να αφαιρέσουμε κάποια από τις μεταβλητές **I** και **Ppron**.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Ppron	Equal variances assumed	487,402	,000	-18,077	9482	,000	-3,66653	,20283	-4,06412	-3,26894
	Equal variances not assumed			-12,962	1507,934	,000	-3,66653	,28288	-4,22141	-3,11166
I	Equal variances assumed	1109,576	,000	-18,735	9482	,000	-2,46521	,13158	-2,72314	-2,20729
	Equal variances not assumed			-10,796	1414,854	,000	-2,46521	,22835	-2,91315	-2,01728
SheHe	Equal variances assumed	5419,440	,000	-33,562	9482	,000	-1,42066	,04233	-1,50364	-1,33769
	Equal variances not assumed			-13,605	1336,141	,000	-1,42066	,10442	-1,62551	-1,21582
Affect	Equal variances assumed	668,550	,000	-41,663	9482	,000	-9,06952	,21769	-9,49624	-8,64280
	Equal variances not assumed			-26,070	1444,357	,000	-9,06952	,34789	-9,75195	-8,38709
Negemo	Equal variances assumed	7915,266	,000	-62,960	9482	,000	-6,12581	,09730	-6,31653	-5,93508
	Equal variances not assumed			-26,783	1343,662	,000	-6,12581	,22872	-6,57450	-5,67711
Anger	Equal variances assumed	8922,003	,000	-43,709	9482	,000	-2,95103	,06752	-3,08337	-2,81868
	Equal variances not assumed			-17,814	1336,936	,000	-2,95103	,16566	-3,27601	-2,62604
Sad	Equal variances assumed	6319,353	,000	-36,382	9482	,000	-1,69338	,04654	-1,78462	-1,60215
	Equal variances not assumed			-14,739	1336,049	,000	-1,69338	,11489	-1,91876	-1,46800
Bio	Equal variances assumed	8391,482	,000	-56,920	9482	,000	-5,53286	,09720	-5,72340	-5,34232
	Equal variances not assumed			-24,123	1343,049	,000	-5,53286	,22936	-5,98281	-5,08291
Body	Equal variances assumed	5156,780	,000	-34,316	9482	,000	-1,92574	,05612	-2,03575	-1,81574
	Equal variances not assumed			-13,935	1336,396	,000	-1,92574	,13820	-2,19685	-1,65464
Sexual	Equal variances assumed	4334,590	,000	-32,152	9482	,000	-2,19261	,06820	-2,32629	-2,05893
	Equal variances not assumed			-13,066	1336,511	,000	-2,19261	,16781	-2,52181	-1,86341

Εικόνα 142. Πίνακας αποτελεσμάτων ελέγχου ανεξαρτησίας του δείγματος (Chrysler)

6.3.2.3.1. Ερμηνεία αποτελεσμάτων

Οι 10 μεταβλητές που συμμετείχαν στην δημιουργία των συστάδων μετά την εφαρμογή του αλγόριθμου είναι οι **I, Affect, Negemo, Bio, Anger, Sad, Body, SheHe, Sexual, Ppron**.

Οι μεταβλητές I, SheHe, Ppron ανήκουν στην γενική κατηγορία των γλωσσικών διεργασιών (Linguistic Processes). Η μεταβλητή Sad αναφέρεται σε λέξεις σχετικές με την λύπη (sadness), όπως crying (κλάμα), grief (θλίψη) και sad (λυπημένος). Η μεταβλητή Affect αποτελείται από λέξεις που σχετίζονται με συναισθηματικές διεργασίες (affective processes), όπως happy (χαρούμενος), cry (κλαίω) και abandon (εγκαταλείπω).

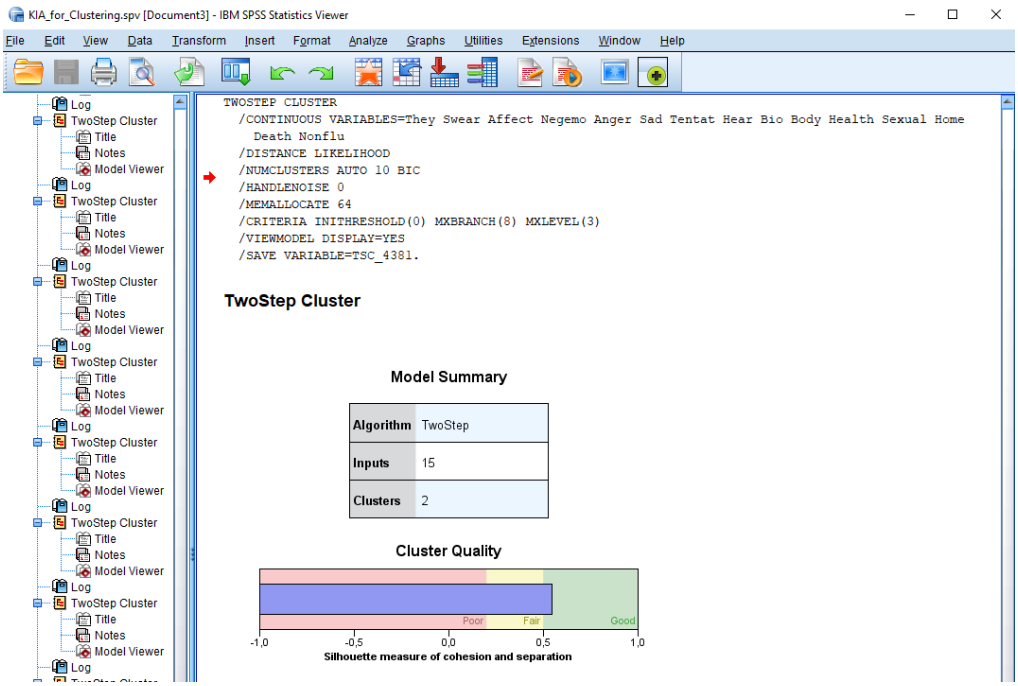
Το προφίλ των χρηστών που συμμετέχουν στον σχολιασμό της αυτοκινητοβιομηχανίας Chrysler παρουσιάζει στοιχεία της κατηγορίας των κριτικών

(critics) της κατηγοριοποίησης της Forrester Research. Οι κριτικοί κατά την αντίδρασή τους σε περιεχόμενο άλλων χρηστών κάνουν συχνή χρήση προσωπικών αντωνυμιών και λέξεων που εκφράζουν συναισθήματα, είτε αυτά είναι θετικά, είτε αρνητικά (negemo, sad, affect).

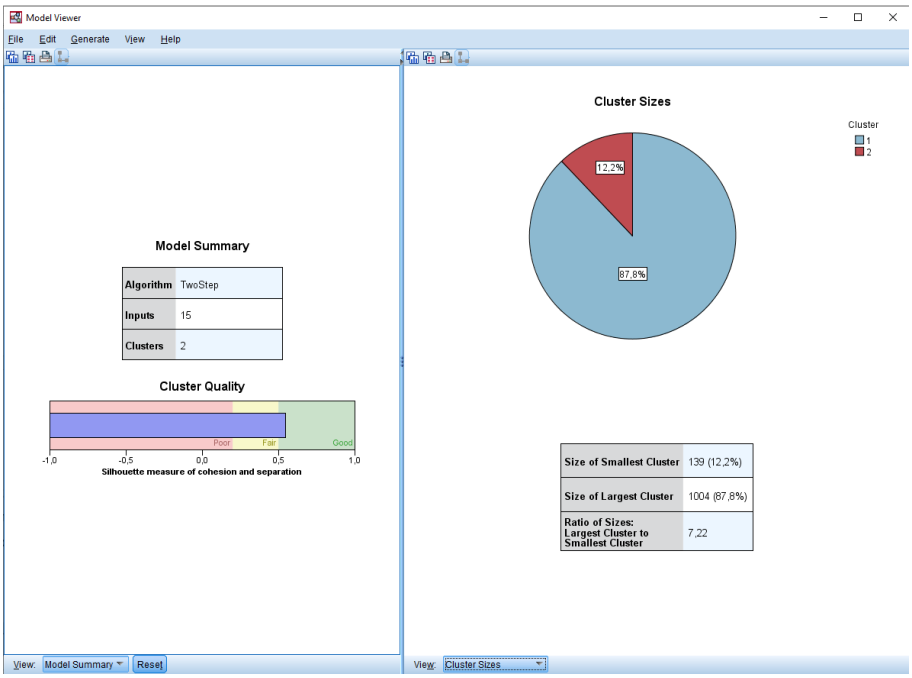
Σύμφωνα με το μοντέλο προσωπικότητας OCEAN, οι χρήστες θα μπορούσαν να ανήκουν στην κατηγορία είτε των εξωστρεφών, είτε των νευρωτικών προσωπικοτήτων. Χαρακτηριστικά όπως η χρήση λέξεων που εκφράζουν συναισθήματα, σχετίζονται με βιολογικές διεργασίες και έχουν σεξουαλικό περιεχόμενο εμφανίζουν σημαντική συσχέτιση με την προσωπικότητας των εξωστρεφών ανθρώπων. Ωστόσο, η χρήση προσωπικών αντωνυμιών και κυρίως του πρώτου προσώπου, σε συνδυασμό με την έκφραση αρνητικών συναισθημάτων όπως θυμό και θλίψη σκιαγραφεί μια νευρωτική προσωπικότητα.

6.3.2.4.*KIA*

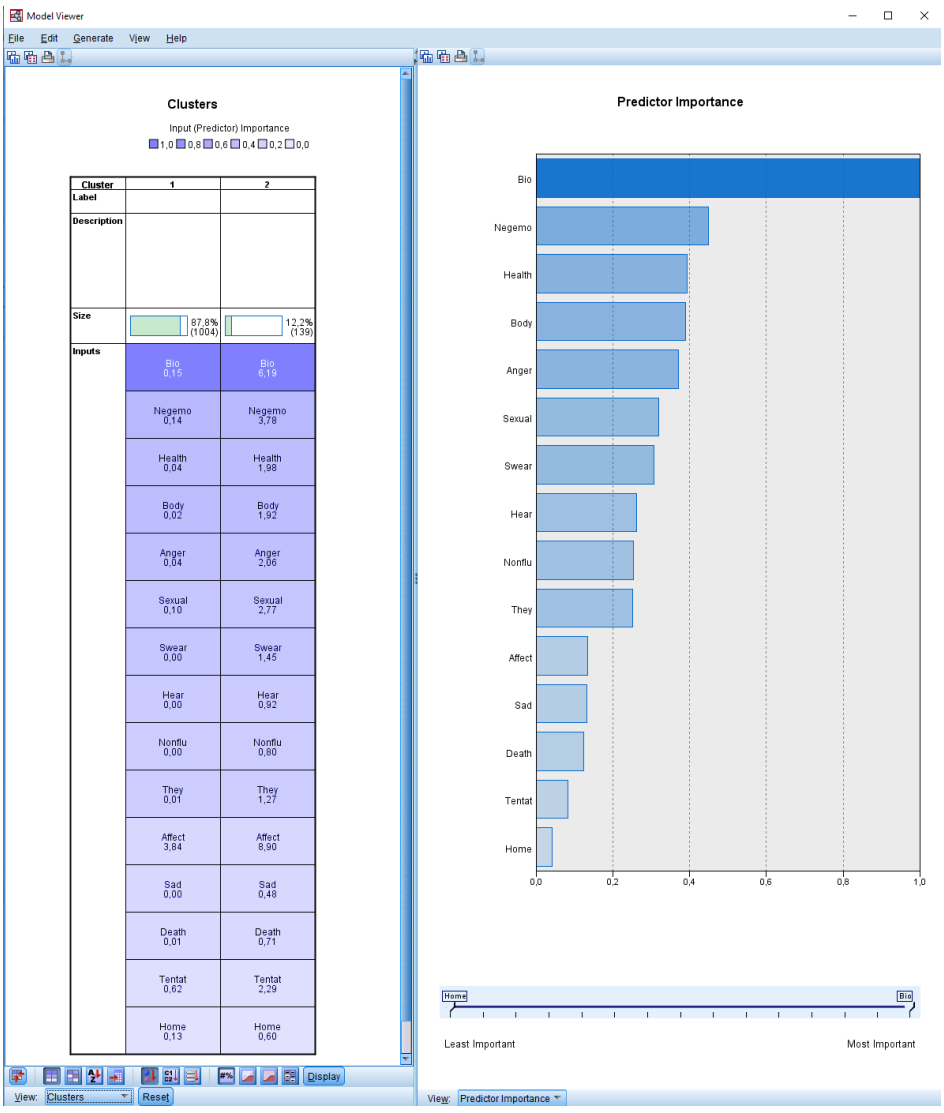
Η βάση δεδομένων της εταιρείας KIA αποτελούνταν από 1143 εγγραφές και 65 μεταβλητές εισόδου του λεξικού LIWC2007. Η εφαρμογή του αλγόριθμου οδήγησε στη δημιουργία ενός μοντέλου 2 συστάδων καλής ποιότητας με συνολικά 15 μεταβλητές εισόδου (Εικόνα), τις: **They, Swear, Affect, Negemo, Anger, Sad, Tentat, Hear, Bio, Body, Health, Sexual, Home, Death** και **Nonflu**. Η πρώτη συστάδα αποτελεί το 12,2% του δείγματος, ενώ η δεύτερη το 87,8% (Εικόνα 144).



Εικόνα 143



Εικόνα 144. Παράθυρο προβολής μοντέλων



Εικόνα 145

Στο ραβδόγραμμα των σημαντικότερων μεταβλητών για τη δημιουργία των συστάδων (Εικόνα 145), παρατηρούμε ότι παρά την καλή ποιότητα του μοντέλου, ένας μεγάλος αριθμός από μεταβλητές όπως οι **Home, Tentat, Death, Sad, Affect, They, Nonflu** και **Hear,** δεν συμμετέχουν σημαντικά στην δημιουργία του μοντέλου. Εφαρμόσαμε και σε αυτή την περίπτωση έναν μη συσχετισμένο έλεγχο t-test για τις 15 μεταβλητές του τελικού μοντέλου (Εικόνα 146). Παρατηρούμε ότι το Levene Test για όλες τις μεταβλητές δίνει p-value = 0 < 0,05 και έτσι κι εδώ η μηδενική υπόθεση για την ισότητα των διακυμάνσεων απορρίπτεται και συνεπώς, η επιλογή διαστήματος εμπιστοσύνης και p-value για τον μη συσχετισμένο έλεγχο t-test θα γίνει από τη δεύτερη γραμμή (Equal variances not assumed) για την κάθε μεταβλητή.

T-Test

Independent Samples Test

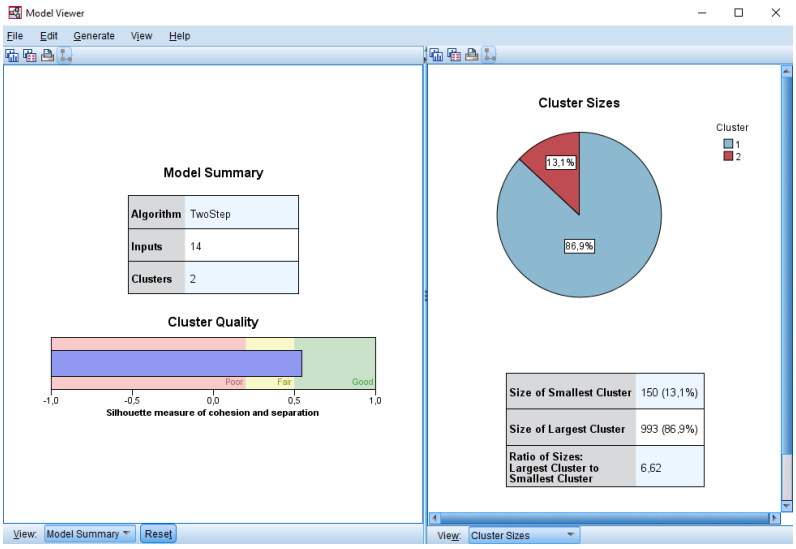
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
They	Equal variances assumed	539,392	,000	-10,646	1141	,000	-1,26480	,11880	-1,49790	-1,03170
	Equal variances not assumed			-3,996	138,125	,000	-1,26480	,31650	-1,89061	-,63899
Swear	Equal variances assumed	649,715	,000	-11,895	1141	,000	-1,45403	,12224	-1,69386	-1,21420
	Equal variances not assumed			-4,414	138,000	,000	-1,45403	,32942	-2,10538	-,80267
Affect	Equal variances assumed	26,974	,000	-7,573	1141	,000	-5,05536	,66753	-6,36509	-3,74563
	Equal variances not assumed			-5,619	155,506	,000	-5,05536	,89971	-6,83259	-3,27812
Negemo	Equal variances assumed	618,202	,000	-14,655	1141	,000	-3,63711	,24819	-4,12406	-3,15015
	Equal variances not assumed			-5,756	138,648	,000	-3,63711	,63185	-4,88642	-2,38779
Anger	Equal variances assumed	785,322	,000	-13,178	1141	,000	-2,02441	,15362	-2,32582	-1,72301
	Equal variances not assumed			-5,053	138,365	,000	-2,02441	,40060	-2,81651	-1,23232
Sad	Equal variances assumed	247,121	,000	-7,527	1141	,000	-,48065	,06385	-,60593	-,35536
	Equal variances not assumed			-2,793	138,000	,006	-,48065	,17208	-,82091	-,14039
Tentat	Equal variances assumed	107,911	,000	-5,787	1141	,000	-1,66454	,28762	-2,22885	-1,10022
	Equal variances not assumed			-2,985	143,128	,003	-1,66454	,55763	-2,76679	-,56228
Hear	Equal variances assumed	595,405	,000	-10,920	1141	,000	-,92165	,08440	-1,08725	-,75606
	Equal variances not assumed			-4,052	138,000	,000	-,92165	,22745	-1,37139	-,47192
Bio	Equal variances assumed	1054,092	,000	-23,330	1141	,000	-6,04720	,25920	-6,55576	-5,53864
	Equal variances not assumed			-9,097	138,560	,000	-6,04720	,66473	-7,36152	-4,73288
Body	Equal variances assumed	862,074	,000	-13,564	1141	,000	-1,90200	,14022	-2,17712	-1,62687
	Equal variances not assumed			-5,108	138,162	,000	-1,90200	,37233	-2,63819	-1,16580
Health	Equal variances assumed	861,391	,000	-13,629	1141	,000	-1,94420	,14265	-2,22409	-1,66431
	Equal variances not assumed			-5,234	138,382	,000	-1,94420	,37144	-2,67862	-1,20977
Sexual	Equal variances assumed	568,582	,000	-12,138	1141	,000	-2,67489	,22037	-3,10727	-2,24251
	Equal variances not assumed			-4,728	138,548	,000	-2,67489	,56574	-3,79349	-1,55629
Home	Equal variances assumed	62,780	,000	-3,985	1141	,000	-,47915	,12023	-,71504	-,24326
	Equal variances not assumed			-1,882	141,382	,062	-,47915	,25454	-,98234	,02405

Εικόνα 146. Πίνακας αποτελεσμάτων ελέγχου ανεξαρτησίας του δείγματος (KIA)

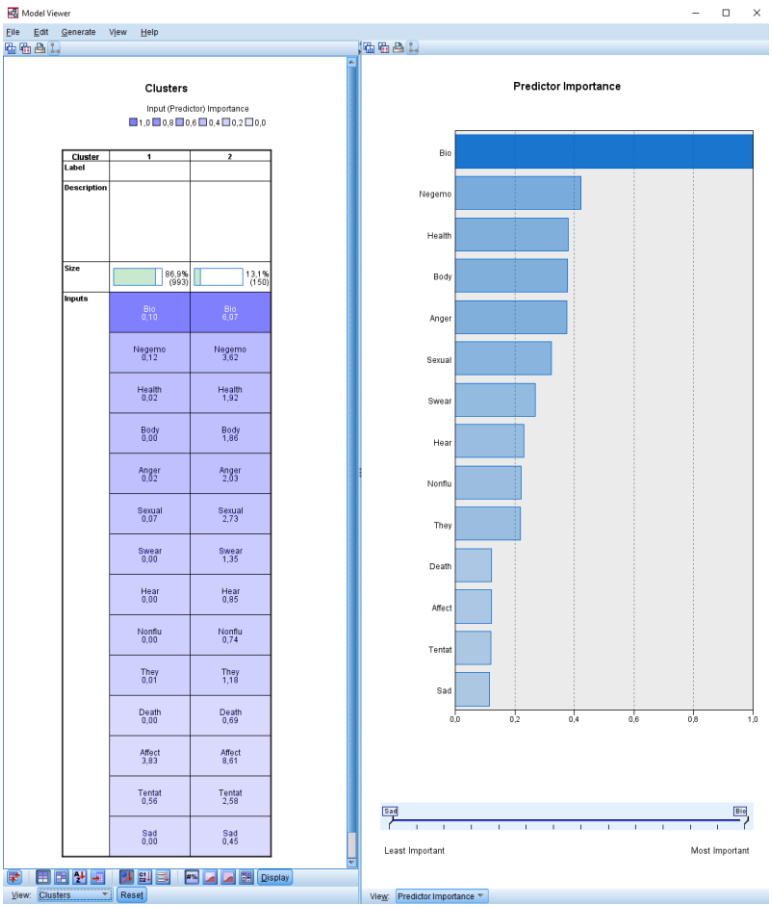
Για τις μεταβλητές **Sad** και **Tentan** το p-value παίρνει τις τιμές 0,006 και 0,003 αντίστοιχα που είναι μικρότερες του επιπέδου σημαντικότητας 0,05 και άρα οι δύο αυτές

ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες, γεγονός που δηλώνει ότι δεν χρειάζεται να τις αφαιρέσουμε από το μοντέλο.

Αντίθετα, η μεταβλητή Home εμφανίζει p-value = 0,062 > 0,05, που σημαίνει ότι υπάρχει στατιστικά σημαντική διαφορά και έτσι την αφαιρέσαμε από το μοντέλο, εφαρμόζοντας εκ νέου τον αλγόριθμο. Τα αποτελέσματα δίνονται στις Εικόνες 147 και 148 .



Εικόνα 147. Παράθυρο προβολής μοντέλων



Εικόνα 148

Επαναλαμβάνοντας τον έλεγχο t-test παρατηρήσαμε πλέον ότι το σύνολο των μεταβλητών που συμμετέχουν στην δημιουργία του μοντέλου δεν εμφανίζουν στατιστικά σημαντική διαφορά (Εικόνα 149).

T-Test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
They	Equal variances assumed	483,184	,000	-10,143	1141	,000	-1,17126	,11548	-1,39783	-,94468
	Equal variances not assumed			-3,977	149,159	,000	-1,17126	,29449	-1,75316	-,58936
Swear	Equal variances assumed	585,397	,000	-11,329	1141	,000	-1,34740	,11893	-1,58074	-1,11406
	Equal variances not assumed			-4,392	149,000	,000	-1,34740	,30675	-1,95355	-,74125
Affect	Equal variances assumed	23,603	,000	-7,401	1141	,000	-4,78729	,64684	-6,05642	-3,51817
	Equal variances not assumed			-5,578	170,142	,000	-4,78729	,85819	-6,48137	-3,09322
Negemo	Equal variances assumed	622,892	,000	-14,568	1141	,000	-3,50293	,24046	-3,97472	-3,03114
	Equal variances not assumed			-5,934	149,721	,000	-3,50293	,59032	-4,66937	-2,33649
Anger	Equal variances assumed	866,876	,000	-13,628	1141	,000	-2,01721	,14802	-2,30763	-1,72680
	Equal variances not assumed			-5,362	149,206	,000	-2,01721	,37624	-2,76066	-1,27377
Sad	Equal variances assumed	224,106	,000	-7,191	1141	,000	-,44540	,06194	-,56692	-,32388
	Equal variances not assumed			-2,788	149,000	,006	-,44540	,15975	-,76107	-,12973
Tentat	Equal variances assumed	169,346	,000	-7,317	1141	,000	-2,01990	,27605	-2,56153	-1,47827
	Equal variances not assumed			-3,666	153,742	,000	-2,01990	,55093	-3,10826	-,93154
Hear	Equal variances assumed	531,232	,000	-10,409	1141	,000	-,85407	,08205	-1,01505	-,69308
	Equal variances not assumed			-4,036	149,000	,000	-,85407	,21163	-1,27225	-,43588
Bio	Equal variances assumed	1092,948	,000	-24,005	1141	,000	-5,96637	,24855	-6,45404	-5,47871
	Equal variances not assumed			-9,598	149,441	,000	-5,96637	,62162	-7,19468	-4,73807
Body	Equal variances assumed	893,507	,000	-13,653	1141	,000	-1,85147	,13561	-2,11754	-1,58541
	Equal variances not assumed			-5,306	149,032	,000	-1,85147	,34897	-2,54104	-1,16190
Health	Equal variances assumed	893,106	,000	-13,711	1141	,000	-1,89165	,13796	-2,16234	-1,62096
	Equal variances not assumed			-5,429	149,297	,000	-1,89165	,34845	-2,58017	-1,20312
Sexual	Equal variances assumed	618,964	,000	-12,528	1141	,000	-2,66239	,21252	-3,07936	-2,24542
	Equal variances not assumed			-5,018	149,468	,000	-2,66239	,53054	-3,71072	-1,61406
Death	Equal variances assumed	241,454	,000	-7,412	1141	,000	-,69160	,09331	-,87468	-,50852
	Equal variances not assumed			-2,874	149,000	,005	-,69160	,24068	-1,16718	-,21602
Nonflu	Equal variances assumed	502,305	,000	-10,219	1141	,000	-,74393	,07280	-,88676	-,60110
	Equal variances not assumed			-3,962	149,000	,000	-,74393	,18776	-1,11496	-,37291

Εικόνα 149

6.3.2.4.1. Ερμηνεία αποτελεσμάτων

Οι 14 μεταβλητές που συμμετείχαν στην δημιουργία των συστάδων μετά την εφαρμογή του αλγόριθμου είναι οι **They, Swear, Affect, Negemo, Anger, Sad, Tentat, Hear, Bio, Body, Health, Sexual, Death** και **Nonflu**.

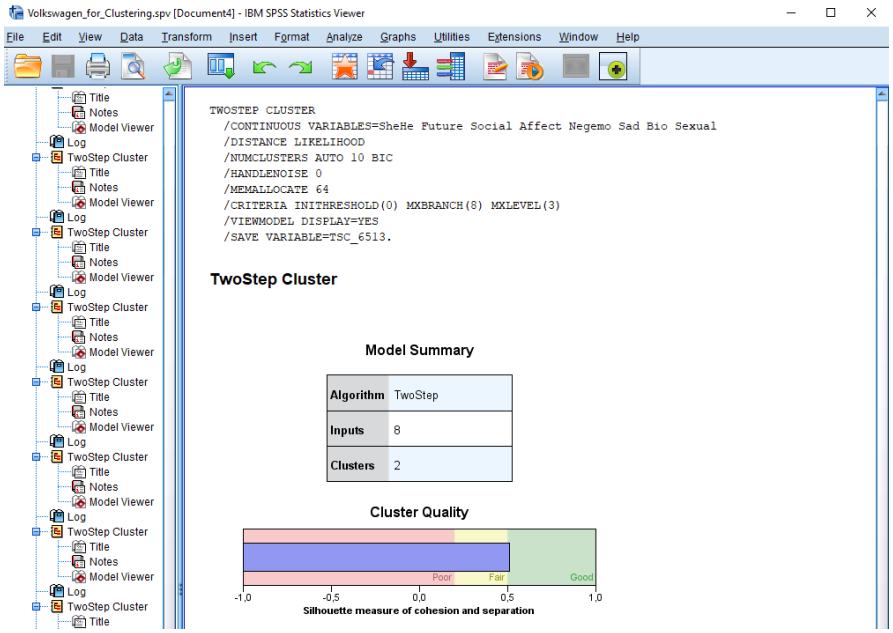
Η μεταβλητή Swear αναφέρεται στη χρήση υβριστικών λέξεων, ενώ η μεταβλητή Tentat αφορά λέξεις που δηλώνουν αβεβαιότητα ή διστακτικότητα, όπως οι λέξεις maybe (ίσως), perhaps (πιθανώς) και guess (εικάζω). Επιπλέον η μεταβλητή Death, η οποία ανήκει στην ευρύτερη κατηγορία των προσωπικών ανησυχιών, αφορά λέξεις που σχετίζονται με τον θάνατο όπως bury (θάβω), coffin (φέρετρο) και kill (θανάτωση). Η μεταβλητή Nonflu ανήκει στις κατηγορίες ομιλίας και περιλαμβάνει «λέξεις» που δεν εμφανίζουν ροή λόγου όπως για παράδειγμα είναι τα er, hm και umm. Αυτού του είδους οι λέξεις δηλώνουν αμφιβολία και χρησιμοποιούνται συνήθως όταν ο ομιλητής προσπαθεί να υπεκφύγει να μιλήσει για ένα συγκεκριμένο θέμα. Επίσης, η μεταβλητή Hear αφορά λέξεις που δηλώνουν ακρόαση, όπως listen (ακούω) και hearing (ακρόαση).

Οι χρήστες που συμμετέχουν στον σχολιασμό της αυτοκινητοβιομηχανίας KIA εμφανίζουν στοιχεία της κατηγορίας τόσο των κριτικών, όσο και των θεατών (spectators) σύμφωνα με την κατηγοριοποίηση της Forrester Research. Οι κριτικοί κατά τον σχολιασμό του χρησιμοποιούν υβριστικές λέξεις (που μπορεί να περιέχουν και σεξουαλικό περιεχόμενο) και εκφράζονται συχνά με θυμό, που αντανακλά ίσως και την ψυχολογική τους κατάσταση. Οι θεατές είναι άτομα που ακούν, διαβάζουν, αλλά και σχολιάζουν το περιεχόμενο που δημοσιεύουν οι άλλοι χρήστες, χρησιμοποιώντας συχνά εκφράσεις χωρίς βαθυστόχαστες σκέψη που συνοδεύονται συχνά από λέξεις της κατηγορίας Nonflu (προέρχεται από τη λέξη nonfluencies).

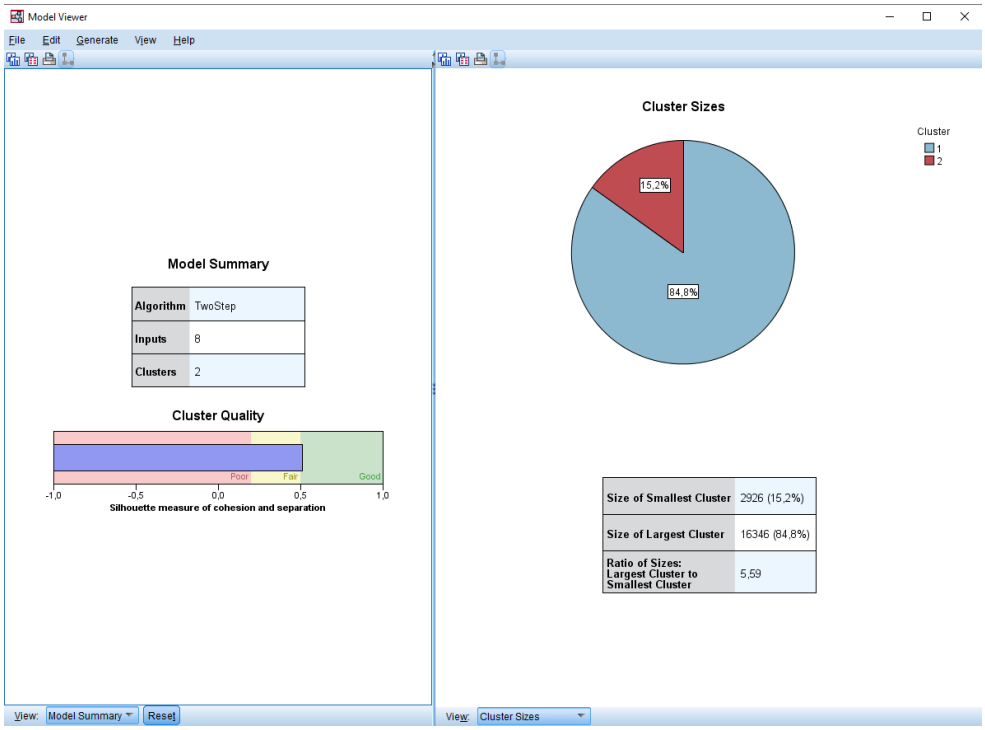
Σύμφωνα με το μοντέλο προσωπικότητας OCEAN, οι χρήστες θα μπορούσαν να ανήκουν στην κατηγορία των νευρωτικών, καθώς οι μεταβλητές από τις οποίες προκύπτουν οι συστάδες εμφανίζουν σημαντική συσχέτιση με τα χαρακτηριστικά γνωρίσματα των χρηστών με αυτή την προσωπικότητα, όπως είναι η σκέψη του θανάτου, η θλίψη και η αβεβαιότητα.

6.3.2.5. *Volkswagen*

Η βάση δεδομένων της εταιρείας Volkswagen αποτελούνταν από 19272 εγγραφές και 65 μεταβλητές εισόδου του λεξικού LIWC2007. Η εφαρμογή του αλγόριθμου οδήγησε στη δημιουργία ενός μοντέλου 2 συστάδων καλής ποιότητας με συνολικά 8 μεταβλητές εισόδου (Εικόνα 150), τις: **SheHe, Future, Social, Affect, Negemo, Sad, Bio και Sexual**. Η πρώτη συστάδα αποτελεί το 84,8% του δείγματος, ενώ η δεύτερη το 15,2% (Εικόνα 151).

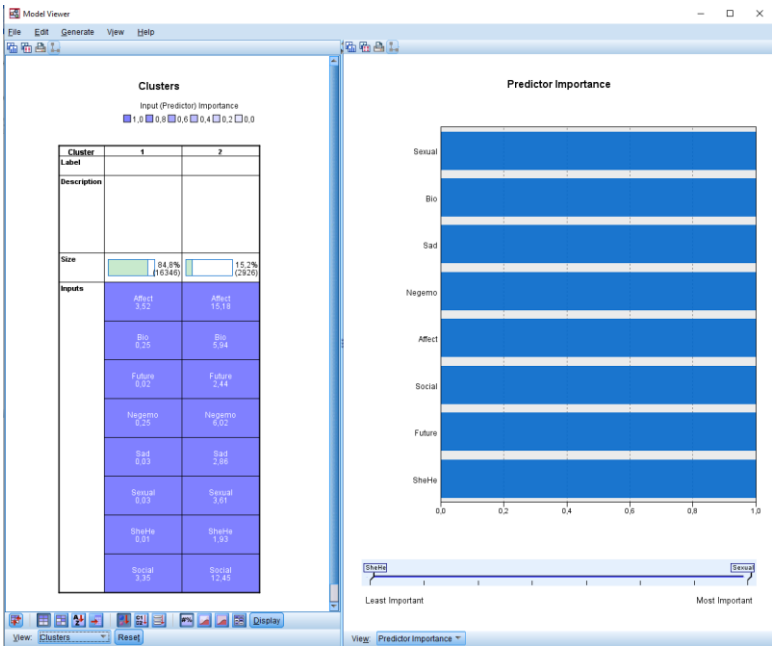


Εικόνα 150



Εικόνα 151. Παράθυρο προβολής μοντέλων

Στο ραβδόγραμμα των σημαντικότερων μεταβλητών για τη δημιουργία των συστάδων (Εικόνα 152), παρατηρούμε ότι το σύνολο των μεταβλητών που συμμετέχουν στην δημιουργία του μοντέλου είναι εξίσου σημαντικές. Από τον μη συσχετισμένο έλεγχο t-test επιβεβαιώσαμε ότι όλες οι ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες (Εικόνα 153).



Εικόνα 152

T-Test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SheHe	Equal variances assumed	8322,238	,000	-43,470	19270	,000	-1,91773	,04412	-2,00420	-1,83126
	Equal variances not assumed			-18,506	2927,464	,000	-1,91773	,10363	-2,12092	-1,71454
Future	Equal variances assumed	19768,615	,000	-62,192	19270	,000	-2,41324	,03880	-2,48930	-2,33718
	Equal variances not assumed			-26,655	2930,122	,000	-2,41324	,09054	-2,59076	-2,23572
Social	Equal variances assumed	2854,828	,000	-53,752	19270	,000	-9,09348	,16917	-9,42507	-8,76188
	Equal variances not assumed			-32,617	3140,268	,000	-9,09348	,27879	-9,64011	-8,54684
Affect	Equal variances assumed	2792,793	,000	-61,492	19270	,000	-11,65465	,18953	-12,02614	-11,28315
	Equal variances not assumed			-35,416	3097,050	,000	-11,65465	,32908	-12,29988	-11,00942
Negemo	Equal variances assumed	10910,574	,000	-63,314	19270	,000	-5,77336	,09119	-5,95209	-5,59463
	Equal variances not assumed			-27,909	2941,669	,000	-5,77336	,20686	-6,17897	-5,36775
Sad	Equal variances assumed	8633,435	,000	-46,886	19270	,000	-2,83566	,06048	-2,95421	-2,71712
	Equal variances not assumed			-19,977	2927,795	,000	-2,83566	,14194	-3,11398	-2,55734
Bio	Equal variances assumed	11531,974	,000	-59,778	19270	,000	-5,68536	,09511	-5,87178	-5,49894
	Equal variances not assumed			-26,383	2942,190	,000	-5,68536	,21550	-6,10790	-5,26282
Sexual	Equal variances assumed	9672,895	,000	-48,969	19270	,000	-3,57250	,07295	-3,71550	-3,42950
	Equal variances not assumed			-20,849	2927,503	,000	-3,57250	,17135	-3,90848	-3,23652

Εικόνα 153. Πίνακας αποτελεσμάτων ελέγχου ανεξαρτησίας του δείγματος (Volkswagen)

6.3.2.5.1. Ερμηνεία αποτελεσμάτων

Οι 8 μεταβλητές που συμμετείχαν στην δημιουργία των συστάδων μετά την εφαρμογή του αλγόριθμου είναι οι **SheHe, Future, Social, Affect, Negemo, Sad, Bio και Sexual**.

Η μεταβλητή Future αναφέρεται σε λέξεις που σχετίζονται με το μέλλον, όπως will (επιθυμία) και gonna (σκοπεύω να). Παρατηρούμε ότι οι μεταβλητές που εμφανίζονται ως σημαντικές για τον καθορισμό των συστάδων περιέχουν λέξεις που δηλώνουν κοινωνικές, συναισθηματικές και βιολογικές διεργασίες, αρνητικότητα και θλίψη, καθώς επίσης και λέξεις με σεξουαλικό περιεχόμενο.

Οι κατηγορίες των λέξεων που εμφανίζουν τα tweets των χρηστών για την αυτοκινητοβιομηχανία Volkswagen, δεν μας δίνουν κάποια αξιοποιήσιμη πληροφορία για τον προσδιορισμό της κατηγορίας των χρηστών των κοινωνικών μέσων. Θα μπορούσαμε ωστόσο να τους κατατάξουμε με ασφάλεια στους συμμετέχοντες, καθώς διατηρούν λογαριασμό στο κοινωνικό μέσο Twitter.

Σύμφωνα με το μοντέλο προσωπικότητας OCEAN, οι χρήστες θα μπορούσαν να είναι εξωστρεφείς λόγω της κοινωνικής τους δραστηριότητας, την εξωτερίκευση των συναισθημάτων του και τις αναφορές τους στο μέλλον. Η ύπαρξη των μεταβλητών Negemo και Sad εκφράζουν επίσης συναισθήματα που ακόμη κι αν δεν είναι θετικά, συχνά εξωτερικεύονται από τους ανθρώπους με αυτό τον τύπο προσωπικότητας.

7. Συμπεράσματα – Προτάσεις

7.1. Γενικά συμπεράσματα

7.1.1. Ανάλυση συναισθήματος

Από την ανάλυση συναισθήματος που πραγματοποιήσαμε στα tweets που αφορούσαν τον κλάδο της αυτοκινητοβιομηχανίας μπορέσαμε να εξάγουμε μια σειρά από συμπεράσματα που αφορούν τόσο την διαδικασία όσο και τα αποτελέσματα αυτής.

Τα μηνύματα που δημοσιεύονται στο κοινωνικό μέσο Twitter περιλαμβάνουν εκτός του κειμένου που εκφράζει την άποψη του χρήστη, μια σειρά από συντομεύσεις που αφορούν λειτουργίες σήμανσης ή/και αναφοράς. Για το λόγο αυτό η ανάλυση συναισθήματος αυτού του είδους κειμένου προϋποθέτει την αρχική επεξεργασία του κάθε tweet με σκοπό την άντληση της χρήσιμης πληροφορίας προς ανάλυση. Η εκκαθάριση της αρχικής βάσης δεδομένων από διπλότυπες εγγραφές και η προεπεξεργασία που ακολούθησε έδωσε την δυνατότητα στον αλγόριθμο ανάλυσης συναισθήματος να εξάγει την πολικότητα και την υποκειμενικότητα του κάθε tweet και στη συνέχεια να προσδιορίσει βάση της πολικότητας το συναίσθημα που εκφράζεται σε αυτό. Με αυτό τον τρόπο τα tweets που αφορούσαν την κάθε αυτοκινητοβιομηχανία κατηγοριοποιήθηκαν σε θετικά, αρνητικά ή ουδέτερα, αναδεικνύοντας το γενικό συναίσθημα των χρηστών αναφορικά με αυτές.

Τα βήματα της προεπεξεργασίας που ακολουθήθηκαν ώστε τα δεδομένα να αποκτήσουν την κατάλληλη μορφή προς ανάλυση, επιλέχθηκαν βάση της διεθνούς βιβλιογραφίας για την προετοιμασία των δεδομένων πριν την εφαρμογή ενός αλγόριθμου ανάλυσης συναισθήματος.

Η ανάλυση των tweets για κάθε αυτοκινητοβιομηχανία πραγματοποιήθηκε σε δύο περιπτώσεις. Στην πρώτη περίπτωση, κατά την διάρκεια της προεπεξεργασίας, εκτός των διαδικασιών της διακριτοποίησης και της αφαίρεσης των λέξεων διακοπής και των ειδικών χαρακτήρων που δεν εμφανίζουν κάποια χρησιμότητα για την εξαγωγή συναισθήματος, εφαρμόσαμε και την διαδικασία της στελέχωσης κατά την οποία γίνεται αντιστοίχιση όλων των διαφορετικών μορφών μιας λέξης στην βασική της λέξη που καλείται λήμμα. Στην δεύτερη περίπτωση, αφαιρέσαμε την διαδικασία της στελέχωσης από την προεπεξεργασία των δεδομένων και επαναλάβαμε την ανάλυση.

Τα αποτελέσματα έδειξαν ότι για τις αυτοκινητοβιομηχανίες Audi και Chevrolet είχαμε μεταστροφή του ποσοστού των θετικών και ουδέτερων tweets μετά την εφαρμογή της διαδικασίας της στελέχωσης. Η μεταστροφή αυτή οφείλεται πιθανότατα στην

αναντιστοιχία των λέξεων που απαρτίζουν τα tweets μετά την στελέχωση και των εκφράσεων/λέξεων του λεξικού[1011] που χρησιμοποιεί η βιβλιοθήκη TextBlob για να εξάγει την πολικότητα και την αντικειμενικότητα.

Τα tweets που επιλέχθηκαν για την διεξαγωγή της ανάλυσης συναισθήματος συλλέχθηκαν κατά την περίοδο διεξαγωγής του Super Bowl, του πιο διάσημου και πολυσυζητημένου ετήσιου ποδοσφαιρικού αγώνα των ΗΠΑ, με μεγάλη εμπορική και διαφημιστική συμμετοχή από εταιρίες κάθε είδους και αντικειμένου δραστηριότητας. Αν και πρόκειται για ένα αθλητικό γεγονός, ήδη από τα πρώτα χρόνια της διεξαγωγής του άρχισε να εμφανίζεται ένα πολιτιστικό φαινόμενο παράλληλα με το παιχνίδι, αυτό της παρουσίασης διαφημίσεων υψηλής κινηματογραφικής ποιότητας, σουρεαλισμού, χιούμορ και χρήσης ειδικών εφέ[177]. Είναι γεγονός πλέον ότι πολλοί από τους θεατές παρακολουθούν το παιχνίδι προκειμένου να δουν τις διαφημίσεις των αγαπημένων τους εταιριών, μέσα από τις οποίες προβάλλονται συνήθως νέα προϊόντα και υπηρεσίες, ενώ συχνά ανακοινώνονται ειδικές προσφορές και εκπτώσεις. Αυτό έχει σαν συνέπεια την αύξηση της συμμετοχής των χρηστών των κοινωνικών μέσων σε συζητήσεις γύρω από τις εταιρίες που διαφημίζονται, εκφράζοντας συναισθήματα, επιθυμίες, γνώμες, αλλά και ασκώντας κριτική για ότι δεν τους αρέσει σε αυτές.

Η ανάλυση συναισθήματος που εφαρμόσαμε στο συγκεκριμένο δείγμα των tweets ανέδειξε το γενικό αίσθημα των χρηστών για την κάθε αυτοκινητοβιομηχανία, δίνοντας την δυνατότητα σε αυτές να καταλήξουν πιθανώς σε χρήσιμα συμπεράσματα σχετικά με το αν τα συγκεκριμένα μοντέλα αυτοκινήτου που παρουσίασαν στις διαφημίσεις τους είχαν θετικό, αρνητικό ή ουδέτερο αντίκτυπο στο κοινό. Η αξία αυτής της πληροφόρησης έγκειται στο γεγονός ότι στα κοινωνικά μέσα οι γνώμες που εκφράζονται είναι σε μεγάλο βαθμό «αφιλτράριστες» κάτι που τις καθιστά περισσότερο υποκειμενικές. Μέσα από την σφυγμομέτρηση αυτή θα μπορούσαν πιθανόν να καθοριστούν στη συνέχεια οι επόμενες διαφημιστικές και εμπορικές κινήσεις των αυτοκινητοβιομηχανιών. Μια θετική στάση των χρηστών, όπως στην περίπτωση των εταιριών Audi, KIA και Chevrolet, επιβεβαιώνει σε μεγάλο βαθμό την στρατηγική μιας εταιρίας, ενώ μια αρνητική σηματοδοτεί πιθανές αλλαγές και βελτιώσεις που πρέπει να γίνουν. Τέλος, μια ουδέτερη

[10] Αγγλικό λεξικό της βιβλιοθήκης TextBlob για την Επεξεργασία Φυσικής Γλώσσας: https://raw.githubusercontent.com/sloria/TextBlob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-lexicon.txt

[11] Λεξικό Υποκειμενικότητας για Αγγλικά Επίθετα (Subjectivity Lexicon For English Adjectives): https://github.com/sloria/TextBlob/blob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-sentiment.xml

στάση, όπως στην περίπτωση των εταιριών Chrysler και Volkswagen, μπορεί να αποτελεί ένδειξη στασιμότητας, οπότε και θα πρέπει να βρεθούν τρόποι αύξησης της παρουσίας τους στο συγκεκριμένο κλάδο, βελτίωσης της εικόνας τους, προσέλκυσης νέων πελατών και ενίσχυσης του αισθήματος δέσμευσης των ήδη υπαρχόντων πελατών.

7.1.2. Εξόρυξη προφίλ χρηστών

Για την εξόρυξη του προφίλ των χρηστών του Twitter έγινε χρήση των αποτελεσμάτων της ανάλυσης που είχε προηγηθεί σε tweets που αφορούσαν τον κλάδο της αυτοκινητοβιομηχανίας από το λογισμικό LIWC2007. Από την ανάλυση αυτή προέκυψε για το κάθε tweet μια βαθμολογία/ποσοστό για κάθε μία από τις μεταβλητές του ενσωματωμένου λεξικού που διαθέτει. Κάθε μεταβλητή αντιστοιχεί σε μια συγκεκριμένη κατηγορία λέξεων καλύπτοντας το σύνολο σχεδόν των βασικών λέξεων που θα μπορούσαν να χρησιμοποιηθούν σε ένα κείμενο. Σύμφωνα με τις μελέτες των Tausczik και Pennebaker (2010)[81] και την εργασία των Schwarts et al. [137] φαίνεται να υπάρχει συσχέτιση μεταξύ των κατηγοριών των λέξεων του λεξικού LIWC με το φύλο, την ηλικία και το μοντέλο OCEAN.

Στα πλαίσια της ανάλυσης μας και προκειμένου να εξάγουμε συμπεράσματα σχετικά με το προφίλ των χρηστών των αυτοκινητοβιομηχανιών Audi, Chevrolet, Chrysler , KIA και Volkswagen, εφαρμόσαμε τον αλγόριθμο συσταδοποίησης δύο βημάτων με την βοήθεια του προγράμματος SPSS. Το αποτέλεσμα σε κάθε περίπτωση ήταν η δημιουργία συστάδων βάση τη σημαντικότητας της κάθε μεταβλητής του λεξικού LIWC2007. Καθώς οι ποσοτικές μεταβλητές σχετίζονται με τις λέξεις που χρησιμοποιούνται για την σύνταξη των tweets, η χρήση συγκεκριμένων λέξεων από τους χρήστες φανερώνουν συγκεκριμένα μοτίβα προσωπικότητας αυτών. Οι μεταβλητές που πετύχαιναν την δημιουργία ενός μοντέλου συστάδων με καλή ποιότητα μας έδωσαν την δυνατότητα να σκιαγραφήσουμε το προφίλ των χρηστών, σύμφωνα με τα αποτελέσματα των παραπάνω εργασιών.

Τα αποτελέσματα της μελέτης που διεξαγάγαμε οδήγησαν στην κατηγοριοποίηση των χρηστών του twitter σε δύο κυρίως κατηγορίες προσωπικότητας σύμφωνα με το μοντέλο OCEAN, στους εξωστρεφείς και τους νευρωτικούς. Η διάκριση μεταξύ αυτών των δύο τύπων προσωπικότητας εμφανίζει δυσκολίες, όπως έχουν επισημάνει και στην μελέτη του οι Argamon et al. (2005)[132].

Πιο συγκεκριμένα οι χρήστες, τα σχόλια των οποίων αφορούσαν τις αυτοκινητοβιομηχανίες Chevrolet, Chrysler και KIA, εμφάνιζαν αρνητικά συναισθήματα στα tweets τους, όπως νευρικότητα, θυμό και θλίψη, ενώ έντονη ήταν και η χρήση πρώτου προσώπου και υβριστικών λέξεων με σεξουαλικό περιεχόμενο. Τα χαρακτηριστικά αυτά σκιαγραφούν τον τύπο των νευρωτικών προσωπικοτήτων.

Επίσης, τα tweets των χρηστών που συμμετείχαν στον σχολιασμό των εταιριών Audi και Volkswagen, περιείχαν λέξεις που υποδηλώνουν τόσο κοινωνικές διεργασίες, όσο και αρνητικά συναισθήματα, γεγονός που θα μπορούσε να τους κατατάξει σε δύο τύπους προσωπικότητας, τόσο των εξωστρεφών, όσο και των νευρωτικών. Η δυσκολία στην διάκριση μεταξύ αυτών των δύο τύπων προσωπικότητας οφείλεται πιθανότατα στην έλλειψη που παρουσιάζει το λεξικό LIWC2007 στην αναγνώριση την γλώσσας που χρησιμοποιείται στα κοινωνικά μέσα (netspeak) αλλά και των γνωστών emoticons, μέσω των οποίων εκφράζονται συναισθήματα με έντονο συνήθως τρόπο. Την δυνατότητα αυτή παρέχει η νέα έκδοση του λεξικού LIWC-22.

Σύμφωνα με την έρευνα του Mooradian (1996) [178], οι καταναλωτές διαφέρουν στον τρόπο με τον οποίο ανταποκρίνονται στις συναισθηματικές εκκλήσεις που συνοδεύουν συνήθως τις διαφημίσεις, ανάλογα με τον τύπο της προσωπικότητάς τους. Οι Larsen και Ketelaar (1991)[179] καταλήγουν στο συμπέρασμα ότι οι νευρωτικοί καταναλωτές είναι πιο πιθανό να επικεντρωθούν σε αρνητικά μηνύματα με έντονο το αίσθημα της τιμωρίας, ενώ οι εξωστρεφείς τείνουν να επικεντρώνονται σε μηνύματα ανταμοιβής.

Η γνώση του τύπου προσωπικότητας που παρέχει η ανάλυσή μας στα tweets των χρηστών για τις παραπάνω αυτοκινητοβιομηχανίες θα μπορούσε να χρησιμοποιηθεί από το τμήμα μάρκετινγκ της εκάστοτε εταιρίας, ώστε να καθοριστούν οι παράμετροι, το είδος και το ύφος των μελλοντικών διαφημίσεων, αλλά και τα μηνύματα που αυτές θέλουν θα περάσουν στους καταναλωτές τους, ώστε να ανταποκριθούν με τον καλύτερο δυνατό τρόπο στις απαιτήσεις και τις προσδοκίες τους, αποφεύγοντας παράλληλα την δημιουργία πιθανόν αρνητικών σχολίων για τις ίδιες και τα προϊόντα τους.

Μια επιπλέον κατηγοριοποίηση των χρηστών επιχειρήθηκε με βάση τα 7 επίπεδα της «Σκάλας των κοινωνικών τεχνολογικών συμπεριφορών»[7] που δημοσιεύτηκε ως αποτέλεσμα της έρευνας της Forrester Research το 2010. Η μελέτη μας έδειξε ότι οι χρήστες εμφανίζουν κυρίως χαρακτηριστικά των συνομιλητών, των κριτικών, των θεατών και των συμμετεχόντων, καθώς τα tweets τους έχουν περιεχόμενο σχολιασμού

και συζητήσεων σχετικά με τα προϊόντα, τις διαφημίσεις και τα τεκταινόμενα γενικά των αυτοκινητοβιομηχανιών.

Πιο συγκεκριμένα τα tweets των χρηστών για τις εταιρίες Audi και Volkswagen, τους κατατάσσουν στην κατηγορία των συμμετεχόντων καθώς συμμετέχουν ενεργά στα κοινωνικά μέσα εκφράζοντας έντονα συναισθήματα. Αυτοί οι χρήστες είναι και εκείνοι που σύμφωνα με το μοντέλο OCEAN ανήκουν στους εξωστρεφείς ή νευρωτικούς, χαρακτηριστικά των οποίων παρατηρούμε ότι συναντώνται και στους συμμετέχοντες. Τα tweets που αφορούν την εταιρία Chevrolet χαρακτηρίζουν τους χρήστες που ανήκουν στην κατηγορία των συνομιλητών, καθώς χρησιμοποιούν συχνά προσωπικές αντωνυμίες, εκφράζουν έντονα την άποψή τους, συνοδεύοντας την άλλοτε με αρνητικού περιεχομένου λέξεις και άλλοτε με ευγενικές εκφράσεις. Επίσης, τα tweets των εταιριών Chrysler και KIA εμφανίζουν στοιχεία που κατατάσσουν του χρήστες στην κατηγορία των κριτικών καθώς χαρακτηρίζονται από συχνή χρήση προσωπικών αντωνυμιών και συναισθηματικών λέξεων.

Τα αποτελέσματα της παραπάνω κατηγοριοποίησης δίνουν χρήσιμες πληροφορίες που θα μπορούσαν να αξιοποιηθούν από τις αυτοκινητοβιομηχανίες, προκειμένου να καθορίσουν την στρατηγική παρουσίας τους στα κοινωνικά μέσα, ώστε να συνάδουν με τα χαρακτηριστικά των χρηστών που συμμετέχουν σε αυτά, προωθώντας με τον βέλτιστο τρόπο τα προϊόντα και τις υπηρεσίες τους.

Αξίζει να σημειώσουμε ότι η ερμηνεία των αποτελεσμάτων στηρίχθηκε σε σημαντικό βαθμό στην υποκειμενικότητα του παρατηρητή, καθώς η επιχείρηση εξόρυξης της προσωπικότητας ενός ανθρώπου μέσω του γραπτού του λόγου σε ένα κοινωνικό μέσο, εμφανίζει εν γένει περιορισμούς και πιθανές παρερμηνείες, λόγω της ελεύθερης και χωρίς πολλές σκέψεις εκφοράς του λόγου.

7.2. Περιορισμοί της έρευνας και προτάσεις για περαιτέρω έρευνα

Η ανάλυση συναισθήματος των tweets πραγματοποιήθηκε με τη βοήθεια της βιβλιοθήκης TextBlob της γλώσσας προγραμματισμού Python. Η βιβλιοθήκη αυτή περιέχει ένα μεγάλο αριθμό αλγορίθμων επεξεργασίας φυσικής γλώσσας με δυνατότητα παραμετροποίησης των διαδικασιών και σημαντικά ακριβή αποτελέσματα.

Η ανάλυση συναισθήματος από το TextBlob γίνεται εξετάζοντας λέξεις και φράσεις στις οποίες εκχωρεί τιμή πολικότητας και υποκειμενικότητας βάση του ενσωματωμένου λεξικού που διαθέτει. Η βαθμολογία της πολικότητας είναι αυτή που χρησιμοποιείται στη συνέχεια για τον προσδιορισμό του συναισθήματος. Ωστόσο, η αδυναμία του συγκεκριμένου εργαλείου έγκειται στο ότι κατά την ανάλυση συναισθήματος ενός κειμένου αγνοεί τις λέξεις που δεν γνωρίζει μη εκχωρώντας βαθμολογία, με αποτέλεσμα να περιορίζεται η ακρίβεια της ανάλυσης.

Επιπλέον, η διαδικασία της εκκαθάρισης των δεδομένων πριν την εφαρμογή του αλγορίθμου, θα μπορούσε να περιλαμβάνει πιο λεπτομερή επιλογή των χαρακτήρων που περιέχονταν στα tweets, και πιο συγκεκριμένα εκείνων που ο συνδυασμός τους οδηγούν στην εμφάνιση των λεγόμενων emoticons. Αν και έγινε προσπάθεια προσδιορισμού των συγκεκριμένων χαρακτήρων και απομάκρυνσή τους από το κείμενο, το αποτέλεσμα δεν ήταν ικανοποιητικό.

Για την επικύρωση των αποτελεσμάτων της ανάλυσης συναισθήματος θα μπορούσαν να χρησιμοποιηθούν και άλλες βιβλιοθήκες επεξεργασίας φυσικής γλώσσας της Python, όπως είναι η βιβλιοθήκη Vader Sentiment (Valance aware dictionary for sentiment reasoning)[12] και η βιβλιοθήκη Flair[13].

Ο προσδιορισμός της προσωπικότητας των χρηστών του κοινωνικού μέσου Twitter που επιχειρήθηκε στην παρούσα διπλωματική βασίστηκε εξ' ολοκλήρου στα αποτελέσματα της συσταδοποίησης δύο βημάτων που εφαρμόσαμε στις βαθμολογίες των 65 μεταβλητών του λεξικού LIWC2007 μετά την ανάλυση των tweets. Ωστόσο, για την εφαρμογή του συγκεκριμένου αλγορίθμου υπάρχουν μια σειρά από προϋποθέσεις που πρέπει να πληρούνται σχετικά με το δείγμα στο οποίο πραγματοποιείται η ανάλυση και οι οποίες λόγω της φύσης του δείγματος δεν ήταν δυνατόν να εκπληρωθούν στο σύνολό τους. Ένας επιπλέον καθοριστικός παράγοντας, από τον οποίο εξαρτώνται σε σημαντικό βαθμό τα αποτελέσματα της έρευνάς μας είναι το γεγονός ότι το σύνολο σχεδόν των λέξεων που απαρτίζουν την κάθε κατηγορία του λεξικού LIWC2007, προέρχονται από

[12]https://pypi.org/project/vaderSentiment/
[13] https://github.com/flairNLP/flair

δείγματα κειμένων, όπως δοκίμια, άρθρα, ερωτηματολόγια, συνεντεύξεις, κ.α αλλά και από λέξεις της καθομιλουμένης γλώσσας. Μία πιο ακριβής ανάλυση θα μπορούσε να επιτευχθεί αξιοποιώντας τα αποτελέσματα της ανάλυσης των tweets από την νεότερη έκδοση της εφαρμογής LIWC-22, η οποία ενσωματώνει την δυνατότητα ανάγνωσης της γλώσσας «netspeak» που είναι η κοινή γλώσσα επικοινωνίας στις αναρτήσεις των κοινωνικών μέσων Twitter και Facebook, όπως και των σύντομων μηνυμάτων των υπηρεσιών όπως Snapchat. Επίσης, θα πρέπει να λάβουμε υπόψιν μας και το γεγονός ότι το λεξικό LIWC2007 αγνοεί γλωσσικές συμπεριφορές, όπως την ειρωνεία, τον αυτοσαρκασμό κ.α, καθώς επίσης και τις ιδιωματικές εκφράσεις, καθιστώντας εκτός ανάλυσης ένα αρκετά μεγάλο αριθμό λέξεων και εκφράσεων με σημαντική συμβολή στην προσπάθεια ανίχνευσης και ανάδειξης της προσωπικότητας ενός χρήστη των κοινωνικών μέσων.

Τέλος, ο περιορισμός των λέξεων που επιβάλλεται στη σύνταξη ενός tweet έχει ως αποτέλεσμα οι χρήστες να εκφράζουν «εν συντομία» οποιαδήποτε άποψη ή συναίσθημα, γεγονός που θέτει επιπλέον δυσκολίες στην προσπάθεια ανάλυσής του για την εύρεση κάποιου συναισθήματος και την αναγνώριση του τύπου προσωπικότητας του συντάκτη του μηνύματος.

8. Βιβλιογραφία

[1] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life.," *J Pers Soc Psychol*, vol. 90, no. 5, p. 862, 2006.

[2] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of computer-mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[3] "Wikipedia," *https://el.wikipedia.org/wiki/Μέσα_κοινωνικής_δικτύωσης*.

[4] C. T. Carr and R. A. Hayes, "Social Media: Defining, Developing, and Divining," *Atl J Commun*, vol. 23, no. 1, pp. 46–65, Jan. 2015, doi: 10.1080/15456870.2015.972282.

[5] Merriam-Webster, "Definition of social media," *https://www.merriam-webster.com/dictionary/social%20media*, 2019.

[6] C. Mayfield, G. Perdue, and K. Wooten, "Investment management and personality type," *Financial Services Review*, vol. 17, Jan. 2008.

[7] C. Li, "Groundswell. Winning in a World Transformed by Social Technologies," *Strategic Direction*, vol. 26, Jun. 2010, doi: 10.1108/sd.2010.05626hae.002.

[8] Aliza Rosen, "Tweeting Made Easier," *https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier*, Nov. 07, 2017.

[9] Twitter, "'About Twitter,'" *https://web.archive.org/*, Apr. 24, 2014. https://about.twitter.com/en/who-we-are/our-company

[10] Twitter, "'Twitter turns six,'" *Twitter*, Mar. 21, 2012. https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html

[11] Twitter Search, "'The Engineering Behind Twitter's New Search Experience,'" *Twitter Engineering Blog. Twitter.* , May 31, 2011. https://web.archive.org/web/20140325080255/https://blog.twitter.com/2011/engineering-behind-twitter's-new-search-experience

[12] Leslie D'Monte, "'Swine flu's tweet tweet causes online flutter,'" *Business Standard*, Jan. 19, 2013. https://www.business-standard.com/article/technology/swine-flu-s-tweet-tweet-causes-online-flutter-109042900097_1.html

[13] B. M. BRETT MOLINA, "Twitter overcounted active users since 2014, shares surge on profit hopes," *USA Today*. https://eu.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/

[14] N. Carlson, "'10% Of Twitter Users Account For 90% Of Twitter Activity,'" *Business Insider* , Jun. 02, 2009. https://www.cbsnews.com/news/stunning-new-numbers-on-who-uses-twitter/

[15] A. H. STEFAN WOJCIK, "Sizing Up Twitter Users," *Pew Research Center*, Apr. 25, 2019. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[16] M. H. L. Isaac, "'Musk's deal for Twitter is worth about $44 billion,'" *The New York Times*, Apr. 25, 2022. https://www.nytimes.com/live/2022/04/25/business/elon-musk-twitter?smid=url-copy#musks-deal-for-twitter-is-worth-about-44-billion

[17] L. Feiner, "'Twitter accepts Elon Musk's buyout deal,'" *CNBC*, Apr. 25, 2022. https://www.cnbc.com/2022/04/25/twitter-accepts-elon-musks-buyout-deal.html

[18] M. A. Russell and M. Klassen, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More*. O'Reilly Media, 2018. [Online]. Available: https://books.google.gr/books?id=DXJ9DwAAQBAJ

[19] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS One*, vol. 5, no. 11, pp. e14118-, Nov. 2010, [Online]. Available: https://doi.org/10.1371/journal.pone.0014118

[20] R. Merchant, S. Elmer, and N. Lurie, “Integrating Social Media Into Emergency-Preparedness Efforts,” *N Engl J Med*, vol. 365, pp. 289–291, Jul. 2011, doi: 10.1056/NEJMp1103591.

[21] R. Procter, F. Vis, and A. Voss, “Reading the riots on Twitter: Methodological innovation for the analysis of big data,” *Int J Soc Res Methodol*, vol. 16, May 2013, doi: 10.1080/13645579.2013.774172.

[22] K. Lachlan, P. Spence, X. Lin, K. Najarian, and M. Greco, “Social Media and Crisis Management: CERC, Search Strategies, and Twitter Content,” *Comput Human Behav*, vol. 54, May 2015, doi: 10.1016/j.chb.2015.05.027.

[23] T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, and B. Adini, “Twitter in the Cross Fire—The Use of Social Media in the Westgate Mall Terror Attack in Kenya,” *PLoS One*, vol. 9, p. e104136, Aug. 2014, doi: 10.1371/journal.pone.0104136.

[24] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, “Using Social Media to Enhance Emergency Situation Awareness,” *IEEE Intell Syst*, vol. 27, no. 6, pp. 52–59, 2012, doi: 10.1109/MIS.2012.6.

[25] M. Cameron, R. Power, B. Robinson, and J. Yin, “Emergency situation awareness from twitter for crisis management,” Apr. 2012, doi: 10.1145/2187980.2188183.

[26] K. Woodfield, *The ethics of online research*. Emerald Group Publishing, 2017.

[27] W. Ahmed, P. Bath, and G. Demartini, “Chapter 4: Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges,” 2017, pp. 79–107. doi: 10.1108/S2398-601820180000002004.

[28] B. Liu, “Sentiment analysis: Mining opinions, sentiments, and emotions,” *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, no. May, pp. 1–367, 2015, doi: 10.1017/CBO9781139084789.

[29] T. Nasukawa and J. Yi, *Sentiment analysis: Capturing favorability using natural language processing*. 2003. doi: 10.1145/945645.945658.

[30] K. Dave, S. Lawrence, and D. Pennock, “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*, vol. 775152, Oct. 2003, doi: 10.1145/775152.775226.

[31] J. Wiebe, “Learning Subjective Adjectives from Corpora,” May 2000.

[32] S. Das and M. Chen, “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web,” *Manage Sci*, vol. 53, pp. 1375–1388, Sep. 2007, doi: 10.1287/mnsc.1070.0704.

[33] R. Tong, “An Operational System for Detecting and Tracking Opinions in On-line Discussions,” in *Working Notes of the SIGIR Workshop on Operational Text Classification*, 2001, pp. 1–6.

[34] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, *Mining product reputations on the Web*. 2002. doi: 10.1145/775047.775098.

[35] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Jul. 2002, pp. 79–86. doi: 10.3115/1118693.1118704.

[36] P. Turney, “Thumbs Up or Thumbs Down? {S}emantic Orientation Applied to Unsupervised Classification of Reviews,” *Computing Research Repository - CORR*, pp. 417–424, Dec. 2002, doi: 10.3115/1073083.1073153.

[37] V. Hatzivassiloglou and K. McKeown, “Predicting the semantic orientation of adjectives,” in *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, 1997, pp. 174–181.

[38] T. P. Minka and R. W. Picard, “Interactive learning with a ‘society of models,’” *Pattern Recognit*, vol. 30, no. 4, pp. 565–581, 1997, doi: https://doi.org/10.1016/S0031-3203(96)00113-6.

[39] J. Wiebe, R. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 1999, pp. 246–253.

[40] Debasish Kalita, "A Comprehensive Overview of Sentiment Analysis," *https://www.analyticsvidhya.com/blog/2022/04/a-comprehensive-overview-of-sentiment-analysis/*, Apr. 01, 2022.

[41] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment.," in *AAAI spring symposium: computational approaches to analyzing weblogs*, 2006, pp. 155–158.

[42] E. Sadikov, A. Parameswaran, and P. Venetis, "Blogs as predictors of movie success," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2009, vol. 3, no. 1, pp. 304–307.

[43] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: a sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 607–614.

[44] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[45] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, 2010, vol. 1, pp. 492–499.

[46] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," in *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, 2010, pp. 293–296.

[47] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," 2010.

[48] A. Bermingham and A. Smeaton, "On using Twitter to monitor political sentiment and predict election results," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2011, pp. 2–10.

[49] J. E. Chung and E. Mustafaraj, "Can collective sentiment expressed on twitter predict political elections?," 2011.

[50] D. Gayo-Avello, P. Metaxas, and E. Mustafaraj, "Limits of electoral predictions using twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011, vol. 5, no. 1, pp. 490–493.

[51] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 1195–1198.

[52] E. T. K. Sang and J. Bos, "Predicting the 2011 dutch senate election results with twitter," in *Proceedings of the workshop on semantic analysis in social media*, 2012, pp. 53–60.

[53] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2010, vol. 4, no. 1, pp. 178–185.

[54] T. Yano and N. A. Smith, "What's worthy of comment? content and comment volume in political blogs," 2010.

[55] B. Chen, L. Zhu, D. Kifer, and D. Lee, "What is an opinion about? exploring political standpoints using opinion scoring model," 2010.

[56] C. S. Khoo, A. Nourbakhsh, and J. Na, "Sentiment analysis of online news text: A case study of appraisal theory," *Online Information Review*, 2012.

[57] K.-Y. Ho, Y. Shi, and Z. Zhang, “Does news matter in China’s foreign exchange market? Chinese RMB volatility and public information arrivals,” *International Review of Economics & Finance*, vol. 52, pp. 302–321, 2017.

[58] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J Comput Sci*, vol. 2, no. 1, pp. 1–8, 2011.

[59] W. Zhang and S. Skiena, “Trading strategies to exploit blog and news sentiment,” 2010.

[60] M. McGlohon, N. Glance, and Z. Reiter, “Star quality: Aggregating reviews to rank products and merchants,” 2010.

[61] Y. Hong and S. Skiena, “The wisdom of bookies? sentiment analysis versus. the nfl point spread,” 2010.

[62] J. Sallet *et al.*, “Social network size affects neural circuits in macaques,” *Science (1979)*, vol. 334, no. 6056, pp. 697–700, 2011.

[63] P. Sakunkoo and N. Sakunkoo, “Analysis of social influence in online book reviews,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2009, vol. 3, no. 1, pp. 308–310.

[64] G. Groh and J. Hauffa, “Characterizing social relations via nlp-based sentiment analysis,” 2011.

[65] Y. Chen and J. Xie, “Online consumer review: Word-of-mouth as a new element of marketing communication mix,” *Manage Sci*, vol. 54, no. 3, pp. 477–491, 2008.

[66] D.-H. Park, J. Lee, and I. Han, “The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement,” *International journal of electronic commerce*, vol. 11, no. 4, pp. 125–148, 2007.

[67] A. Ghose, P. Ipeirotis, and A. Sundararajan, “Opinion mining using econometrics: A case study on reputation systems,” in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 416–423.

[68] C. Dellarocas, X. Zhang, and N. F. Awad, “Exploring the value of online product reviews in forecasting sales: The case of motion pictures,” *Journal of Interactive marketing*, vol. 21, no. 4, pp. 23–45, 2007.

[69] N. Archak, A. Ghose, and P. G. Ipeirotis, “Show me the money! Deriving the pricing power of product features by mining consumer reviews,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 56–65.

[70] M. Louwerse, Z. Cai, X. Hu, M. Ventura, and P. Jeuniaux, “Cognitively inspired NLP-based knowledge representations: Further explorations of Latent Semantic Analysis,” *International Journal on Artificial Intelligence Tools*, vol. 15, no. 06, pp. 1021–1039, 2006.

[71] N. Jindal and B. Liu, “Mining comparative sentences and relations,” in *Aaai*, 2006, vol. 22, no. 13311336, p. 9.

[72] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[73] V. Vapnik, “The Support Vector Method of Function Estimation,” in *Nonlinear Modeling: Advanced Black-Box Techniques*, J. A. K. Suykens and J. Vandewalle, Eds. Boston, MA: Springer US, 1998, pp. 55–85. doi: 10.1007/978-1-4615-5703-6_3.

[74] H. Zhang, L. Jiang, and J. Su, “The optimality of naive bayes,” in *In Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, 2004, pp. 562–567.

[75] S. Wang and C. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jul. 2012, pp. 90–94. [Online]. Available: https://aclanthology.org/P12-2018

[76] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” May 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf

[77] A. Das and B. Gambäck, “Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality,” in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, 2012, pp. 38–46.

[78] P. J. Stone, D. C. Dunphy, and M. S. Smith, “The general inquirer: A computer approach to content analysis.,” 1966.

[79] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis,” *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011, doi: 10.1162/COLI_a_00049.

[80] P. D. Turney and M. L. Littman, “Measuring praise and criticism: Inference of semantic orientation from association,” *acm Transactions on Information Systems (tois)*, vol. 21, no. 4, pp. 315–346, 2003.

[81] Y. R. Tausczik and J. W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” *J Lang Soc Psychol*, vol. 29, no. 1, pp. 24–54, Dec. 2009, doi: 10.1177/0261927X09351676.

[82] P. Gonçalves, F. Benevenuto, and M. Cha, “PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter,” Aug. 2013.

[83] J. Bollen, A. Pepe, and H. Mao, “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,” *arXiv preprint arXiv:0911.1583*, 2009.

[84] X. Hu, J. Tang, H. Gao, and H. Liu, “Unsupervised Sentiment Analysis with Emotional Signals,” in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 607–618. doi: 10.1145/2488388.2488442.

[85] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, “Comparing and combining sentiment analysis methods,” in *Proceedings of the first ACM conference on Online social networks*, 2013, pp. 27–38.

[86] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” *J Lang Soc Psychol*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.

[87] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[88] C. Strapparava and A. Valitutti, “Wordnet affect: an affective extension of wordnet.,” in *Lrec*, 2004, vol. 4, no. 1083–1086, p. 40.

[89] E. Cambria, R. Speer, C. Havasi, and A. Hussain, “Senticnet: A publicly available semantic resource for opinion mining,” 2010.

[90] M. Ibrahim, I. S. Bajwa, R. Ul-Amin, and B. Kasi, “A neural network-inspired approach for improved and true movie recommendations,” *Comput Intell Neurosci*, vol. 2019, 2019.

[91] X. Li, H. Xie, R. Y. K. Lau, T.-L. Wong, and F.-L. Wang, “Stock prediction via sentimental transfer learning,” *IEEE Access*, vol. 6, pp. 73110–73118, 2018.

[92] L. Qi, C. Zhang, A. Sukul, W. Tavanapong, and D. A. M. Peterson, “Automated coding of political video ads for political science research,” in *2016 IEEE International Symposium on Multimedia (ISM)*, 2016, pp. 7–13.

[93] C. van Hee, E. Lefever, and V. Hoste, “We usually don’t like going to the dentist: Using common sense to detect irony on Twitter,” *Computational Linguistics*, vol. 44, no. 4, pp. 793–832, 2018.

[94] S. Noferesti and M. Shamsfard, “Using Linked Data for polarity classification of patients’ experiences,” *J Biomed Inform*, vol. 57, pp. 6–19, 2015.

[95] T. Dang *et al.*, “Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 27–35.

[96] H. Almeida, M. Queudot, and M.-J. Meurs, “Automatic triage of mental health online forum posts: CLPsych 2016 system description,” in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 183–187.

[97] A. Joshi, X. Dai, S. Karimi, R. Sparks, C. Paris, and C. R. MacIntyre, “Shot or not: Comparison of NLP approaches for vaccination behaviour detection,” in *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 2018, pp. 43–47.

[98] M. O. Kelly and E. F. Risko, “Journal of Applied Research in Memory and Cognition”.

[99] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *Journal of the American society for information science and technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[100] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word–emotion association lexicon,” *Comput Intell*, vol. 29, no. 3, pp. 436–465, 2013.

[101] J. Brooke, M. Tofiloski, and M. Taboada, “Cross-linguistic sentiment analysis: From English to Spanish,” in *Proceedings of the international conference RANLP-2009*, 2009, pp. 50–54.

[102] P. S. Dodds and C. M. Danforth, “Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents,” *J Happiness Stud*, vol. 11, no. 4, pp. 441–456, 2010, doi: 10.1007/s10902-009-9150-9.

[103] M. M. Bradley and P. J. Lang, “Affective norms for English words (ANEW): Instruction manual and affective ratings,” Technical report C-1, the center for research in psychophysiology …, 1999.

[104] M. E. Francis and J. W. Pennebaker, “Putting stress into words: The impact of writing on physiological, absentee, and self-reported emotional well-being measures,” *American Journal of Health Promotion*, vol. 6, no. 4, pp. 280–287, 1992.

[105] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: LIWC 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[106] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, “The development and psychometric properties of LIWC2015,” 2015.

[107] H. Lane, H. Hapke, and C. Howard, *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications, 2019. [Online]. Available: https://books.google.gr/books?id=UyHgswEACAAJ

[108] A. M. TURING, “I.—COMPUTING MACHINERY AND INTELLIGENCE,” *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[109] P. Simon, *Too big to ignore: the business case for big data*, vol. 72. John Wiley & Sons, 2013.

[110] A. Categorical, “Glossary of terms,” *Mach Learn*, vol. 30, no. 2, pp. 271–274, 1998.

[111] Oxford Lexico, “Oxford,” *https://www.lexico.com/definition/user_profile*, 1960. https://www.lexico.com/definition/user_profile

[112] B. Kerin, A. Caputo, and S. Lawless, “Temporal word embeddings for dynamic user profiling in Twitter,” 2019.

[113] F. Alam, E. A. Stepanov, and G. Riccardi, “Personality traits recognition on social network-facebook,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2013, vol. 7, no. 2, pp. 6–9.

[114] S. Rothmann and E. P. Coetzer, “The big five personality dimensions and job performance,” *SA Journal of Industrial Psychology*, vol. 29, no. 1, 2003, doi: 10.4102/sajip.v29i1.88.

[115] A. Poropat, “A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance,” *Psychol Bull*, vol. 135, pp. 322–338, Apr. 2009, doi: 10.1037/a0014996.

[116] P. E. Shrout and S. T. Fiske, *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske*. Psychology Press, 2014.

[117] G. W. Allport and H. S. Odbert, “Trait-names: A psycho-lexical study.,” *Psychol Monogr*, vol. 47, no. 1, p. i, 1936.

[118] R. M. Bagby, M. B. Marshall, and S. Georgiades, “Dimensional personality traits and the prediction of DSM-IV personality disorder symptom counts in a nonclinical sample,” *J Pers Disord*, vol. 19, no. 1, pp. 53–67, 2005.

[119] E. C. Tupes and R. E. Christal, “Recurrent personality factors based on trait ratings.(pp. 61-97): USAF ASD Tech.” Rep, 1961.

[120] W. T. Norman, “Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings.,” *The journal of abnormal and social psychology*, vol. 66, no. 6, p. 574, 1963.

[121] J. M. Digman, “Personality structure: Emergence of the five-factor model,” *Annu Rev Psychol*, vol. 41, no. 1, pp. 417–440, 1990.

[122] L. R. Goldberg, “The structure of phenotypic personality traits.,” *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[123] B. P. O’Connor, “A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories,” *Assessment*, vol. 9, no. 2, pp. 188–203, 2002.

[124] B. Mershon and R. L. Gorsuch, “Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria?,” *J Pers Soc Psychol*, vol. 55, no. 4, p. 675, 1988.

[125] S. v Paunonen and M. C. Ashton, “Big five factors and facets and the prediction of behavior.,” *J Pers Soc Psychol*, vol. 81, no. 3, p. 524, 2001.

[126] S. Roccas, L. Sagiv, S. H. Schwartz, and A. Knafo, “The big five personality factors and personal values,” *Pers Soc Psychol Bull*, vol. 28, no. 6, pp. 789–801, 2002.

[127] C. G. DeYoung, L. C. Quilty, and J. B. Peterson, “Between facets and domains: 10 aspects of the Big Five.,” *J Pers Soc Psychol*, vol. 93, no. 5, p. 880, 2007.

[128] G. Toegel and J.-L. Barsoux, “How to become a better leader,” *MIT Sloan Manag Rev*, vol. 53, no. 3, pp. 51–60, 2012.

[129] G. Farnadi *et al.*, “How are you doing? : emotions and personality in Facebook,” in *EMPIRE2014 : 2nd workshop on emotions and personality in personalized services*, 2014, pp. 45–56.

[130] V. Benet-Martínez and O. P. John, “Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English.,” *J Pers Soc Psychol*, vol. 75, no. 3, p. 729, 1998.

[131] J. W. Pennebaker and L. A. King, “Linguistic styles: language use as an individual difference.,” *J Pers Soc Psychol*, vol. 77, no. 6, p. 1296, 1999.

[132] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, “Lexical predictors of personality type,” in *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*, 2005, pp. 1–16.

[133] T. Yarkoni, “Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers,” *J Res Pers*, vol. 44, no. 3, pp. 363–373, 2010.

[134] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*. 2011. doi: 10.1109/PASSAT/SocialCom.2011.26.

[135] L. Qiu, H. Lin, J. Ramsay, and F. Yang, “You are what you tweet: Personality expression and perception on Twitter,” *J Res Pers*, vol. 46, no. 6, pp. 710–718, 2012.

[136] R. Wald, T. Khoshgoftaar, and C. Sumner, “Machine prediction of personality from Facebook profiles,” in *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, 2012, pp. 109–115.

[137] H. A. Schwartz *et al.*, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PLoS One*, vol. 8, no. 9, p. e73791, 2013.

[138] J. Mahmud, J. Chen, and J. Nichols, “Why are you more engaged? predicting social engagement from word use,” *arXiv preprint arXiv:1402.6690*, 2014.

[139] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, “Psychological targeting as an effective approach to digital mass persuasion,” *Proc Natl Acad Sci U S A*, vol. 114, no. 48, pp. 12714–12719, 2017, doi: 10.1073/pnas.1710966114.

[140] Z. Xu, L. Ru, L. Xiang, and Q. Yang, “Discovering user interest on twitter with a modified author-topic model,” in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, vol. 1, pp. 422–429.

[141] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Process*, vol. 25, no. 2–3, pp. 259–284, 1998.

[142] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *J Intell Inf Syst*, vol. 17, no. 2, pp. 107–145, 2001.

[143] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, “Internal versus external cluster validation indexes,” *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.

[144] SPSS, “TwoStep Cluster Analysis,” *https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=edition-twostep-cluster-analysis*, Feb. 28, 2021.

[145] DAVID GRINER, “Debunking the Super Bowl myths,” *https://archive.kitsapsun.com/opinion/david-griner--debunking-the-super-bowl-myths-ep-907964635-354954361.html*, Feb. 01, 2015.

[146] ADAM WELLS, “Super Bowl Commercials 2015: Ad Costs, Value Review Before Patriots vs. Seahawks,” *https://bleacherreport.com/articles/2349506-super-bowl-commercials-2015-ad-costs-value-review-before-patriots-vs-seahawks*, Feb. 01, 2015.

[147] A. Kedia and M. Rasu, *Hands-On Python Natural Language Processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications*. Packt Publishing, 2020. [Online]. Available: https://books.google.gr/books?id=1AbuDwAAQBAJ

[148] Anaconda Inc, “Anaconda (Python distribution),” *https://www.anaconda.com/*, 2012.

[149] Travis Oliphant, “NumPy,” *https://numpy.org/*, 2006.

[150] Wes McKinney, “pandas,” *https://pandas.pydata.org/*, 2008.

[151] David Cournapeau, “scikit-learn,” *https://scikit-learn.org/*, 2007.

[152] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. “ O’Reilly Media, Inc.,” 2009.

[153] Νίκος Αβούρης, “Εισαγωγή στη γλώσσα προγραμματισμού Python. Μάθημα 19. Διαχείριση κειμένου με regular expressions,” Πάτρα. [Online]. Available: https://mathesis.cup.gr

[154] A.M. Kuchling, “Regular Expression HOWTO,” *https://docs.python.org/3.8/howto/regex.html#regex-howto*.

[155] Steven Loria *et al.*, “TextBlob: Simplified Text Processing,” *https://textblob.readthedocs.io/en/dev/*.

[156] Matthew Honnibal and Ines Montani, “spaCy,” *https://spacy.io/usage/spacy-101*, 2015. https://spacy.io

[157] John D. Hunter, “Matplotlib,” *https://matplotlib.org/*, 2003.

[158] Michael Waskom, “seaborn,” *https://seaborn.pydata.org/*, 2012.

[159] Andreas Mueller, “WordCloud,” *https://amueller.github.io/word_cloud/*, 2020.

[160] Python, “Common string operations,” *https://docs.python.org/3/library/string.html*. https://docs.python.org/3/library/string.html

[161] “IBM SPSS Statistics,” *https://www.ibm.com/products/spss-statistics*.

[162] A. P. Rovai, J. D. Baker, and M. K. Ponton, *Social science research design and statistics: A practitioner’s guide to research methods and IBM SPSS*. Watertree Press LLC, 2013.

[163] Wikipedia, “SPSS,” *https://en.wikipedia.org/wiki/SPSS#cite_note-10*.

[164] SPSS Predictive Analytics, “‘What’s New in SPSS Statistics 25 & Subscription - SPSS Predictive Analytics,’” *https://www.ibm.com/products/spss-statistics*, Jul. 18, 2017.

[165] Priya Pedamkar, “What is SPSS,” *https://www.educba.com/what-is-spss/*.

[166] Ruben Geert van den Berg, “SPSS – What Is It?,” *https://www.spss-tutorials.com/spss-what-is-it/*.

[167] J. Perkins, *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd, 2014.

[168] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O’Reilly Media, 2020.

[169] S. Chatterjee and M. Krystyanczuk, *Python Social Media Analytics*. Packt Publishing, 2017. [Online]. Available: https://books.google.gr/books?id=8eZDDwAAQBAJ

[170] D. Kushwah, “ What is difference between stemming and lemmatization? ,” *Quora*, May 16, 2019.

[171] Shahul ES, “Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch,” *https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair*, Jul. 21, 2022.

[172] D. Şchiopu, “Applying TwoStep cluster analysis for identifying bank customers’ profile,” *Buletinul*, vol. 62, no. 3, pp. 66–75, 2010.

[173] SPSS, “TwoStep Cluster Analysis,” *https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=features-twostep-cluster-analysis*, Mar. 22, 2021.

[174] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[175] SPSS, “Model Summary View,” *https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=viewer-model-summary-view#clusterviewer_modelsummary_panel*, Mar. 22, 2021.

[176] Γιώργος Βλαχόπουλος and Κωνσταντίνος Κουτσογιάννης, *Βιοστατιστική - Εφαρμογή με το SPSS και το R Project*. Πάτρα: Εκδόσεις Αλγόριθμος, 2012.

[177] Wikipedia, “Super Bowl commercials,” *https://en.wikipedia.org/wiki/Super_Bowl_commercials*, Jul. 03, 2022.

[178] T. A. Mooradian, “Personality and ad-evoked feelings: The case for extraversion and neuroticism,” *J Acad Mark Sci*, vol. 24, no. 2, pp. 99–109, 1996.

[179] R. J. Larsen and T. Ketelaar, “Personality and susceptibility to positive and negative emotional states.,” *J Pers Soc Psychol*, vol. 61, no. 1, p. 132, 1991.