ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Business Analytics and Data Science

Πρόγραμμα Μεταπτυχιακών Σπουδών στην

ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

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

του Πέτρου Αμοιρίδη του Αναστασίου

Επιβλέπων Καθηγητής: Λεωνίδας Χατζηθωμάς,

Επίκουρος Καθηγητής Πανεπιστημίου Μακεδονίας

Δομή Παρουσίασης

- Εισαγωγή
- Σκοπός της έρευνας
- Μέσα Κοινωνικής Δικτύωσης
- Ανάλυση Συναισθήματος (Sentiment Analysis)
- Εξόρυξη Προφίλ Χρηστών (User Profiling)
- Μεθοδολογία και εργαλεία ανάλυσης
- Ερευνητική διαδικασία
- Συμπεράσματα έρευνας

Εισαγωγή

- Η ανάλυση συναισθήματος (ή εξόρυξη γνώμης) έχει ως στόχο την ανάδειξη του συναισθήματος που εκφράζεται μέσω των δημοσιευμένων μηνυμάτων των χρηστών στα κοινωνικά μέσα και η κατηγοριοποίησή τους σε θετικό, αρνητικό ή ουδέτερο.
- Η ανάλυση συναισθήματος αποτελεί αναπόσπαστο κομμάτι της έρευνας στο αντικείμενο της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing)
- Δύο βασικοί τύποι μεθόδων ανάλυσης συναισθήματος
  - Μηχανική μάθηση
  - Χρήση λεξικών (π.χ LIWC)
- 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).

Εισαγωγή

- 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]
- 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.
- 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.

Σκοπός της έρευνας

- Ανάδειξη της χρησιμότητας διεξαγωγής αναλύσεων σε μηνύματα των χρηστών των κοινωνικών μέσων κατά την διάρκεια διεξαγωγής σημαντικών γεγονότων → 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 συνοδεύονται από πληθώρα προωθητικών και διαφημιστικών ενεργειών από τις εταιρίες που συμμετέχουν σε αυτά ως χορηγοί → Πλούσια πηγή πληροφοριών, καθώς τότε υπάρχει πιο έντονη και ενθουσιώδης συμμετοχή του καταναλωτικού κοινού σε συζητήσεις για αυτές τις εταιρίες και τα πεπραγμένα τους
- Ανάλυση συναισθήματος των tweets των χρηστών του Twitter που αφορούν προϊόντα και εταιρίες του κλάδου της αυτοκινητοβιομηχανίας → Εκμαίευση του γενικού αισθήματος που τα συνοδεύει,
- Διερεύνηση και ανάδειξη του προφίλ των χρηστών (user profiling) με την βοήθεια των αποτελεσμάτων της ανάλυσης των tweets που αφορούν τις αυτοκινητοβιομηχανίες, από την εφαρμογή LIWC2007.

Μέσα Κοινωνικής Δικτύωσης

- Τα κοινωνικά δίκτυα ορίζονται ως βασισμένες στο διαδίκτυο (διαδικτυακές) υπηρεσίες που

✓ επιτρέπουν στα άτομα να δημιουργήσουν ένα δημόσιο ή ημι-δημόσιο προφίλ μέσα σε ένα οριοθετημένο σύστημα,

✓ να επικοινωνήσουν με μια λίστα από άλλους χρήστες με τους οποίους μοιράζονται μια μορφή σύνδεσης και

✓ να δουν και να διανείμουν την δικιά τους λίστα των συνδέσεων και αυτών που φτιάχτηκαν από άλλους μέσα στο σύστημα» (Boyd & Ellison, 2007)

Μέσα Κοινωνικής Δικτύωσης – Βασικά χαρακτηριστικά

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

- ✓ ***Συμμετοχή (Participation)***: Ενθάρρυνση συνεισφοράς και σχολιασμού από τους ενδιαφερομένους χρήστες.
- ✓ ***Διαφάνεια (Openness)***: Υπηρεσίες των κοινωνικών μέσων ανοιχτές σε ανατροφοδότηση και συμμετοχή.
- ✓ ***Συνομιλία (Conversation)***: Δυνατότητα αμφίδρομης επικοινωνίας, σχολιασμού και ανταλλαγής απόψεων.
- ✓ ***Κοινότητα (Community)***: Εύκολη και άμεση δημιουργία κοινοτήτων που μοιράζονται κοινά ενδιαφέροντα.
- ✓ ***Συνεκτικότητα (Connectedness)***: Κοινή χρήση συνδέσεων με άλλες ιστοσελίδες, πόρους και ανθρώπους.

Κατηγορίες χρηστών των κοινωνικών μέσων (Forrester Research, 2010)

- Σύμφωνα με την διαδικτυακή έρευνα της Forrester Research, Inc., που πραγματοποιήθηκε το 2010 στην Αμερική σε δείγμα ενήλικων προέκυψαν 7 διαφορετικές ομάδες χρηστών
- Κάθε ομάδα εμφανίζει μια λίστα από δραστηριότητες στις οποίες συμμετέχουν οι χρήστες που ανήκουν σε αυτήν 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» :

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.

Κατηγορίες χρηστών των κοινωνικών μέσων (Forrester Research, 2010)

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.

Twitter

- Αμερικανική υπηρεσία μικρο-ιστολογίου (micro-blogging) και κοινωνικής δικτύωσης (social network)
- Δημιουργήθηκε από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams τον Μάρτιο του 2006
- Οι χρήστες δημοσιεύουν και αλληλεπιδρούν με μηνύματα γνωστά ως «tweets» («τιτιβίσματα»).
- 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.
- Μέχρι το 2012, περισσότεροι από 100 εκατομμύρια χρήστες δημοσίευσαν 340 εκατομμύρια tweets την ημέρα
- Το 2013, ήταν ένας από τους δέκα ιστότοπους με τις περισσότερες επισκέψεις
- Από το πρώτο τρίμηνο του 2019, το Twitter είχε περισσότερους από 330 εκατομμύρια μηνιαίους ενεργούς χρήστες
- Στην πράξη, η συντριπτική πλειοψηφία των tweets γράφεται από μια μειονότητα χρηστών
- 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.

Twitter – Αξία και δημοτικότητα των tweets

- Δημοσιευμένα σχόλια, μηνύματα και κριτικές στα κοινωνικά μέσα αποτελούν τα τελευταία 20 χρόνια αντικείμενο έρευνας τόσο της ακαδημαϊκής κοινότητας, όσο και των εταιριών ή και σε αρκετές περιπτώσεις σε συνεργασία μεταξύ τους.
- Το περιεχόμενο των κοινωνικών μέσων, παρουσιάζει στους ακαδημαϊκούς ερευνητές νέες σημαντικές ευκαιρίες μελέτης μιας σειράς θεμάτων σε ένα φυσικό περιβάλλον.
- Τα κοινωνικά μέσα αλλάζουν τον τρόπο επικοινωνίας των ανθρώπων, τόσο στην καθημερινή τους ζωή, όσο και σε ακραίες συνθήκες
- Η έρευνα στο Twitter καλύπτει ένα ευρύ φάσμα, όπως η ανάλυση tweets που σχετίζονται με εξεγέρσεις, φυσικές καταστροφές και κρίσιμα γεγονότα
- 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

Twitter – Αξία και δημοτικότητα των tweets

- 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.
- 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), η επιλογή των emoticons για την σύνταξη των tweets από έναν χρήστη αποκαλύπτει τόσο την ψυχολογική του κατάσταση εκείνη τη χρονική στιγμή, όσο και στοιχεία του χαρακτήρα 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.

Twitter – Αξία και δημοτικότητα των tweets

- Σύμφωνα με τους Ερευνητές της κοινότητας των Νέων Κοινωνικών Μέσων Νέα Κοινωνική Επιστήμη (New Social Media New Social Science - NSMNSS), οι λόγοι για τους οποίους το Twitter έχει προσελκύσει περισσότερη ακαδημαϊκή έρευνα σε σύγκριση με άλλες πλατφόρμες κοινωνικών μέσων είναι:

1. Το Twitter API (application programming interface) είναι πιο ανοιχτό και προσβάσιμο σε σύγκριση με άλλες πλατφόρμες κοινωνικών μέσων.

2. Διευκολύνει την εύρεση και την παρακολούθηση συνομιλιών, καθώς διαθέτει λειτουργία αναζήτησης tweets, ενώ και τα tweets εμφανίζονται επίσης στα αποτελέσματα αναζήτησης στο Google.

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.

4. Χρησιμοποιείται ευρέως από δημοσιογράφους, τόσο για τον εντοπισμό γεγονότων που αποτελούν ειδήσεις όσο και για τη διανομή έκτακτων ειδήσεων. Επίσης λαμβάνει πολύ περισσότερη προσοχή από τα μέσα ενημέρωσης επειδή διασημότητες, πολιτικοί και αθλητές δημοσιεύουν tweets 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.

Ανάλυση Συναισθήματος (Sentiment Analysis)

- 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 (|B. Liu, 2015).
- 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.
- 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)

Ανάλυση Συναισθήματος (Sentiment Analysis)

- 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..
- 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.
- 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.
- 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.

Ανάλυση Συναισθήματος (Sentiment Analysis)

Τύποι ανάλυσης συναισθήματος

- Το συνολικό συναίσθημα που εκφράζεται από μια παράγραφο, φράση ή λέξη αναφέρεται ως πολικότητα και μπορεί να μετρηθεί χρησιμοποιώντας μια «βαθμολογία συναισθήματος»

- 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.
  - ✓ Ανάλυση συναισθήματος βάσει διαστάσεων (Aspect-based Sentiment Analysis ABSA): Ανάλυση που σχετίζεται με μια συγκεκριμένη ιδιότητα ή χαρακτηριστικό που περιγράφεται στο κείμενο. Αυτά τα χαρακτηριστικά αναφέρονται ως «θέματα» (themes) στο Θεματικό (Thematic).
  - ✓ Ανίχνευση συναισθημάτων (Emotion detection): Ανίχνευση συγκεκριμένων συναισθημάτων. παραδείγματα.
  - ✓ Βάσει πρόθεσης (Intent based Sentiment Analysis): Διάκριση μεταξύ γεγονότων και απόψεων σε ένα κείμενο.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Εφαρμογές

Τομείς εφαρμογής της Ανάλυσης Συναισθήματος:

- Καταναλωτικά προϊόντα: Η γνώση της άποψης άλλων καταναλωτών για τα προϊόντα βοηθούν στην σωστή και πιο συμφέρουσα επιλογή.
- Κυβερνητικές υπηρεσίες: Ανακάλυψη δημοσίων συναισθημάτων και ανησυχιών των πολιτών που βοηθάει στην λήψη αποφάσεων.
- Box – Office: Πρόβλεψη επιτυχίας ταινιών από το θετικό συναίσθημα που εκφράζεται μέσω των φημών που κυκλοφορούν
- Εκλογές πολιτικών κομμάτων: Χρήση δεδομένων του Twitter για την πρόβλεψη πολιτικών εκλογών
- Χρηματιστήριο: Πρόβλεψη των κινήσεων των δεικτών χρηματιστηρίου μέσα από τον εντοπισμό θετικών και αρνητικών δημόσιων τοποθετήσεων στο Twitter

Ανάλυση Συναισθήματος (Sentiment Analysis) - Μέθοδοι

Δύο βασικοί τύποι μεθόδων ανάλυσης συναισθήματος:

- Μέθοδοι Μηχανικής Μάθησης (machine-learning based)
  - ✓ 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.
  - ✓ Οι ταξινομητές ταξινομούν τα κείμενα ανάλογα με το συνολικό θετικό ή αρνητικό συναίσθημα που εκφράζουν.
  - ✓ Συχνά χρησιμοποιούμενες μέθοδοι μηχανικής μάθησης:
    - Μηχανή διανυσμάτων υποστήριξης (Support Vector Machine - SVM)
    - Μέθοδος Naïve Bayes
  - ✓ 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.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Μέθοδοι

- Μέθοδοι με τη χρήση λεξικών
  - ✓ Βασίζονται σε λεξικά (stentiment lexicon)
  - ✓ Χρήση 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, είτε με την χρήση αυτόματων συσχετισμών λέξεων, είτε ημιαυτόματα αντλώντας τιμές συναισθημάτων από πηγές.
  - ✓ 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.
  - ✓ 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.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Μέθοδοι

Μειονεκτήματα της εφαρμογής των μεθόδων ανάλυσης συναισθήματος με την χρήση λεξικών:

1. Μια θετική ή αρνητική λέξη συναισθημάτων μπορεί να έχει αντίθετους προσανατολισμούς ή πολικότητες σε διαφορετικούς τομείς εφαρμογών ή περιεχόμενο προτάσεων. 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.
3. Οι σαρκαστικές προτάσεις με ή χωρίς συναισθηματικά λόγια είναι δύσκολο να ερμηνευτούν σωστά.
4. Πολλές προτάσεις χωρίς συναισθηματικές λέξεις μπορεί να υποδηλώνουν θετικά ή αρνητικά συναισθήματα ή απόψεις των συγγραφέων τους.

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

Ανάλυση Συναισθήματος (Sentiment Analysis) - Λεξικά

Γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά:

1. SentiWordNet
   - ✓ Λεξικό συναισθήματος που προέκυψε από τον εμπλουτισμό του αγγλικού λεξικού WordNet με πληροφορίες συναισθήματος (Esuli και Sebastiani, 2006)
   - ✓ Αποτελείται από περισσότερες από 38000 πολικές και αρκετές άλλες αυστηρά αντικειμενικές λέξεις.
   - ✓ 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, τις οποίες συσχετίζει 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, επιτρέποντας να έχει πολλαπλά συναισθήματα που αντιστοιχούν σε κάθε έννοια.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Λεξικά

Γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά:

2. WordNet-Affect
   - ✓ 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
3. SenticNet
   - ✓ Μέθοδος εξόρυξης απόψεων και ανάλυσης συναισθημάτων που ερευνά τεχνικές τεχνητής νοημοσύνης και τεχνικές σημασιολογικού Ιστού
   - ✓ Είναι ένας συνδυασμός του λεξικού WordNet-Affect και του ConceptNet.
   - ✓ 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.
   - ✓ Εχει χρησιμοποιηθεί σε πολλές εργασίες εκτός από την ανίχνευση πολικότητας, όπως π.χ. συστήματα σύστασης, χρηματιστηριακή πρόβλεψη, πολιτικές προβλέψεις, ανίχνευση ειρωνείας, μέτρηση αποτελεσματικότητας φαρμάκων, ανίχνευση κατάθλιψης, διαλογή ψυχικής υγείας, ανίχνευση συμπεριφοράς εμβολιασμού, ψυχολογικές μελέτες και πολλά άλλα.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Λεξικά

Γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά:

4. Senti-Strength
   - ✓ 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 που χρησιμοποιείται με την βοήθεια του λεξικού αναζητά λέξεις με παρόμοια ρίζα.
   - ✓ Οι βαθμολογίες μεταξύ 1 και 5 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 ή την αποδυνάμωση των συναισθημάτων, μια λίστα από emoticons, ενώ έγινε και χρήση επαναλαμβανόμενων σημείων στίξης 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.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Λεξικά

Γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά:

5. Emo-Lexicon
   - ✓ 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.
   - ✓ Διαδικασία δημιουργίας ανοιχτή στο ευρύ κοινό → Ποιοτικός έλεγχος αποτελεί μια πρόκληση
6. SO-CAL (Sentiment Orientation CALculator)
   - ✓ Αναπτύχθηκε από τον Brooke το 2009
   - ✓ Βασίζεται σε ένα χειροκίνητα κατασκευασμένο αποθετήριο ακατέργαστων λέξεων.
   - ✓ Αποτελείται από περίπου 5000 λέξεις κάθε μία από τις οποίες 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.
   - ✓ Μεγάλη ακρίβεια αποτελεσμάτων καθώς βασίζεται στη χρήση λεπτομερών χαρακτηριστικών που χειρίζονται το συναίσθημα σε διάφορες περιπτώσεις με τρόπους που συμμορφώνονται με γλωσσικά φαινόμενα.

Ανάλυση Συναισθήματος (Sentiment Analysis) - Λεξικά

Γνωστά και ευρέως χρησιμοποιούμενα εργαλεία ανάλυσης συναισθήματος βασισμένα σε λεξικά:

7. Happiness Index & ANEW (Affective Norms for English Words)
   - ✓ Κλίμακα συναισθημάτων που βασίζεται στο λεξικό ANEW (Affective Norms for English Word).
   - ✓ Συλλογή από 2447 λέξεις που έχουν βαθμολογηθεί από προπτυχιακούς φοιτητές ως προς τρεις συναισθηματικές διαστάσεις, την δυναμικότητα/σθένος (valence), την διέγερση (arousal) και την κυριαρχίας τους (dominance).

Ανάλυση Συναισθήματος (Sentiment Analysis)
Εφαρμογή LIWC

8. LIWC (Linguistic Inquiry and Word Count)

- ✓ 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.
- ✓ 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.
- ✓ Δυνατότητα ανάλυσης μεμονωμένων ή πολλαπλών αρχείων κειμένου
- ✓ 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.

Ανάλυση Συναισθήματος (Sentiment Analysis)
Εφαρμογή LIWC

- ✓ 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.
- ✓ Έχει πρόσβαση σε ένα μόνο αρχείο ή ομάδα αρχείων και αναλύει το καθένα διαδοχικά, γράφοντας την έξοδο σε ένα μοναδικό αρχείο
- ✓ 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 2η έκδοση (LIWC2001) σε 90 στην 4η 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

Ανάλυση Συναισθήματος (Sentiment Analysis)
Εφαρμογή LIWC2007 – Κατηγορίες λέξεων

	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

Ανάλυση Συναισθήματος (Sentiment Analysis)
Επεξεργασία Φυσικής Γλώσσας (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.
- 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.

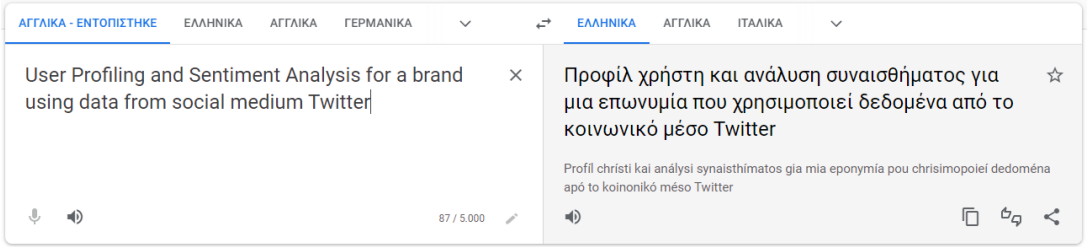
Ανάλυση Συναισθήματος (Sentiment Analysis)
Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)

Εφαρμογές

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

Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) Εφαρμογές

- Machine Translation – Μηχανική μετάφραση
  - ✓Τεχνική που βοηθά στην παροχή συντακτικών, γραμματικών και σημασιολογικά σωστών μεταφράσεων μεταξύ δύο γλωσσών



- Συστήματα αναγνώρισης ομιλίας
  - ✓Η αναγνώριση ομιλίας είναι από τις πιο δύσκολες και απαιτητικές εφαρμογές της NLP και των συστημάτων τεχνητής νοημοσύνης.
  - ✓Βρίσκουν εφαρμογή σε πολλούς τομείς, από τους υπολογιστές και τα κινητά τηλέφωνα, έως τα συστήματα εικονικής βοήθειας ή/και εξυπηρέτησης.

- 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).
  - ✓Ασχολείται κυρίως με τη δημιουργία ισχυρών και κλιμακωτών συστημάτων που παρέχουν απαντήσεις σε ερωτήσεις που δίνονται από χρήστες σε μορφή φυσικής γλώσσας.
  - ✓Πετυχημένα μοντέλα εξατομικευμένης βοήθειας είναι η Siri της Apple και η Cortana της Microsoft

Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) Εφαρμογές

- Αναγνώριση συμφραζόμενων και Ανάλυση Συσχέτισης
  - ✓ Καλύπτει έναν ευρύ τομέα στην κατανόηση της φυσικής γλώσσας, η οποία περιλαμβάνει συντακτική και σημασιολογική συλλογιστική.
  - ✓ Η συσχέτιση εμφανίζεται όταν δύο ή περισσότερες όροι/εκφράσεις σε ένα σώμα κειμένου αναφέρονται στην ίδια οντότητα (πρόσωπο ή πράγμα).

- Σύνοψη κειμένου (Text summarization)
  - ✓ Λήψη ενός συνόλου εγγράφων κειμένου και μείωση του περιεχόμενο κατάλληλα προς δημιουργία μιας περίληψης διατηρώντας τα βασικά σημεία των εγγράφων αυτών.

- Κατηγοριοποίηση κειμένου (Text categorization)
  - ✓ Προσδιορισμός κατηγορίας ή κλάσης που θα πρέπει να τοποθετηθεί ένα συγκεκριμένο έγγραφο με βάση το περιεχόμενο του εγγράφου.
  - ✓ Έχει βοηθήσει στη δημιουργία πολλών επιτυχημένων και πρακτικών εφαρμογών, συμπεριλαμβανομένων φίλτρων ανεπιθύμητης αλληλογραφίας και κατηγοριοποίησης άρθρων ειδήσεων.

- Ανάλυση κειμένου (Text Analysis)
  - ✓ Μεθοδολογία και διαδικασία που ακολουθείται για την εξαγωγή ποιοτικών και εφαρμόσιμων πληροφοριών και πληροφοριών από δεδομένα κειμένου.
  - ✓ Χρήση τεχνικών επεξεργασίας φυσικής γλώσσας, ανάκτησης πληροφοριών και μηχανικής μάθησης
  - ✓ Περιλαμβάνει μια συλλογή τεχνικών μηχανικής μάθησης, γλωσσολογίας και στατιστικής που χρησιμοποιούνται για τη μοντελοποίηση και εξαγωγή πληροφοριών από το κείμενο που υπόκεινται σε ανάλυση,

Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) Εφαρμογές

- 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)

- 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

Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) Εφαρμογές

- Μηχανική Μάθηση (Machine Learning - ML)
  - ✓ 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»
  - ✓ Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα 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.
  - ✓ Οι τρεις κύριες κατηγορίες τεχνικών μηχανικής μάθησης περιλαμβάνουν εποπτευόμενους, μη επιτηρημένους αλγορίθμους, καθώς και αλγορίθμους ενισχυτικής μάθησης.

- Βαθιά Μάθηση (Deep Learning - DL)
  - ✓ Υποπεδίο της μηχανικής μάθησης που ειδικεύεται σε μοντέλα και αλγορίθμους, οι οποίοι έχουν εμπνευστεί από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου.
  - ✓ Η βαθιά μάθηση ή τα βαθιά νευρωνικά δίκτυα συνήθως χρησιμοποιούν πολλαπλά στρώματα μη γραμμικών μονάδων επεξεργασίας, επίσης γνωστά ως νευρώνες, ή προτιμότερα «μονάδες επεξεργασίας».

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»:
  1. 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.
  2. Είναι η συλλογή πληροφοριών ή δεδομένων σχετικά με τις συνήθειες, τις προτιμήσεις κ.λπ, ενός χρήση, ιδίως για ένα προϊόν ή μια υπηρεσία.

- 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.

- Είναι η διαδικασία απόκτησης δεδομένων, εκτέλεσης οποιασδήποτε απαιτούμενης επεξεργασίας σε αυτά για την παραγωγή ενός ολοκληρωμένου μοντέλου ή αναπαράστασης ενός χρήστη ή μιας ομάδας χρηστών.

Big 5 ή OCEAN (Openess-Conscientiousness-extraversion-agreebleness-neuroticism)

- Η προσωπικότητα είναι ένα σύνολο ιδιαίτερων χαρακτηριστικών που διαφοροποιούν το ένα άτομο από το άλλο.
- 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
- Ορίστηκε από διάφορες ανεξάρτητες ομάδες ερευνητών που χρησιμοποίησαν την ανάλυση παραγόντων των λεκτικών περιγραφέων (verbal descriptors) της ανθρώπινης συμπεριφοράς.
- Προτάθηκε από τους Ernest Tupes και Raymond Christal το 1961, αλλά απέτυχε να προσεγγίσει ένα ακαδημαϊκό κοινό μέχρι τη δεκαετία του 1980.

Big 5 ή OCEAN (Openess-Conscientiousness-extraversion-agreebleness-neuroticism)

- Το 1990, ο J.M. Digman πρότεινε το μοντέλο προσωπικότητας πέντε παραγόντων, το οποίο ο Lewis Goldberg επέκτεινε σε υψηλότερο επίπεδο οργάνωσης.
- 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.
- 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.

Big 5 ή OCEAN (Openess-Conscientiousness-extraversion-agreebleness-neuroticism)

- 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/άμελος)

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]

Εξόρυξη Προφίλ Χρηστών (User Profiling) – Σχετικές εργασίες

- Pennebaker (1999)
  - ✓ Οι κατηγορίες των λέξεων που χρησιμοποιούνται σε καθημερινές συζητήσεις συνδέονται με τα OCEAN χαρακτηριστικά προσωπικότητας.

- Argamon et al. (2005)
  - ✓ Μελέτη της δυσκολίας πρόβλεψης του τύπου της εξωστρέφειας, σε σχέση με τον τύπο του νευρωτισμού.

- Tausczik και Pennebaker (2010b)
  - ✓ Εξήγηση του τρόπου 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.
  - ✓ Κατά προσέγγιση πρόβλεψη της κατάστασης ψυχικής υγείας των ανθρώπων.

- Yarkoni (2010)
  - ✓ Συσχέτισης μεταξύ των προτιμήσεων στη χρήση λέξεων και τον τύπο προσωπικότητας ενός ατόμου.
  - ✓ 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.

Εξόρυξη Προφίλ Χρηστών (User Profiling) – Σχετικές εργασίες

- ✓ 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.
- ✓ Μια προσωπικότητα είναι ένας σημαντικός παράγοντας που επηρεάζει είτε τη συμπεριφορά ενός ατόμου στον εικονικό κόσμο όσο και τη συμπεριφορά του στον πραγματικό κόσμο.

- D. Quercia et al. (2011)
  - ✓ 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.

- Qiu et al. (2012)
  - ✓ Μέτρηση OCEAN χαρακτηριστικών ερευνώντας τη σχέση μεταξύ αυτών και ορισμένων γλωσσικών χαρακτηριστικών που εμφανίζονται σε tweets.
  - ✓ Γλωσσική ανάλυση για την πρόβλεψη της προσωπικότητας χρησιμοποιώντας την εφαρμογή LIWC2007.
  - ✓ 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,
  - ✓ Οι ανοιχτοί άνθρωποι (openness) συσχετίζεται αρνητικά με τις βρισιές, την επιρροή και τις λέξεις χωρίς ευχέρεια, αλλά συσχετίζεται έντονα με τη χρήση προθέσεων.

Εξόρυξη Προφίλ Χρηστών (User Profiling) – Σχετικές εργασίες

- R. Wald et al. (2012)
  - ✓ 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.

- Schwartz et al. (2013)
  - ✓ 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.
  - ✓ 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).
  - ✓ 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.

Εξόρυξη Προφίλ Χρηστών (User Profiling) – Σχετικές εργασίες

- Schwartz et al. (2013)
  - ✓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.

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*

Εξόρυξη Προφίλ Χρηστών (User Profiling) – Σχετικές εργασίες

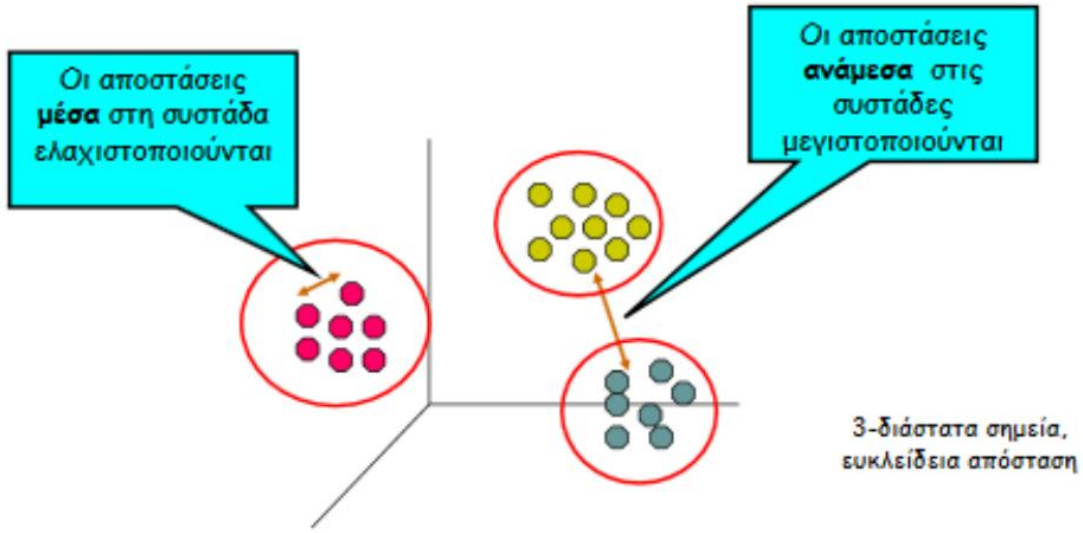
- Mahmud et al. (2014)
  - ✓ Δημιουργία «έξυπνου συστήματος συλλογής πληροφοριών» που παράγει ορισμένες ερωτήσεις για να λάβει τις επιθυμητές πληροφορίες
  - ✓ Σκοπός ήταν να επιλεγούν «οι κατάλληλοι χρήστες την κατάλληλη στιγμή» που είναι πιο πιθανό να δώσουν τις απαραίτητες πληροφορίες.
  - ✓ Το σύστημα 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.
  - ✓ Έγινε χρήση της εφαρμογής LIWC-2001.

- Οι Matz et al. (2017)
  - ✓ Ανάλυση της προσωπικότητα των χρηστών του Facebook ως μέσο μαζικής πειθούς για την αποτελεσματικότερη ανάπτυξη της αγοράς.
  - ✓ Καθορισμός OCEAN χαρακτηριστικών της προσωπικότητας.

- Οι Z. Xu et al. (2011)
  - ✓ Σύνολο δεδομένων αποτελούμενο από 200 tweets, 200 retweets, 200 links και 200 replies
  - ✓ Χαρακτηρισμός με μη αυτόματο τρόπο καθενός από 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.

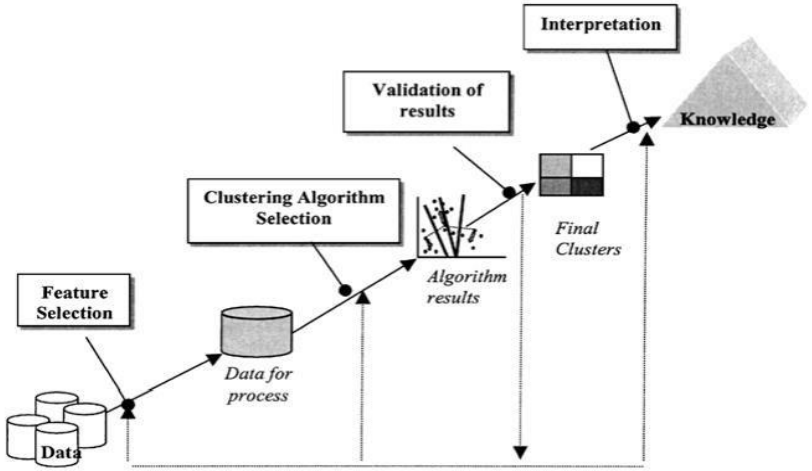
Συσταδοποίηση

- 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).



Συσταδοποίηση – Βήματα συσταδοποίησης

- Τα βασικά βήματα της διαδικασίας συσταδοποίησης σύμφωνα με την Μαρία Χαλκίδη (2001) φαίνονται στην παρακάτω εικόνα:



- **Επιλογή χαρακτηριστικών γνωρισμάτων:** Επιλογή κατάλληλων χαρακτηριστικών γνωρισμάτων βάση των οποίων θα εκτελεστεί με επιτυχία η συσταδοποίηση → Απαραίτητη η προεπεξεργασία των δεδομένων

- **Αλγόριθμος συσταδοποίησης:** Επιλογή του κατάλληλου αλγορίθμου συσταδοποίησης → Εξαρτάται από τα μορφή των δεδομένων και τις ανάγκες τις εκάστοτε εφαρμογής → Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης είναι αυτά που κυρίως χαρακτηρίζουν έναν αλγόριθμο συσταδοποίησης, όπου:
  1. Γειτνίαση: Ομοιότητα μεταξύ των στοιχείων
  2. Κριτήριο συσταδοποίησης: Εκφράζεται συνήθως μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων

Συσταδοποίηση – Βήματα συσταδοποίησης

- **Επικύρωση των αποτελεσμάτων ή εγκυρότητα συσταδοποίησης (cluster validity)**

  Τρεις προσεγγίσεις για τη διερεύνηση της εγκυρότητας:

  ✓*Εξωτερικά κριτήρια*: Έλεγχος των σημείων του συνόλου δεδομένων αν είναι τυχαία δομημένα ή όχι → →δείκτης Rand, συντελεστής Jaccard, εντροπία, καθαρότητα.

  ✓*Εσωτερικά κριτήρια*: Αξιολόγηση αποτελεσμάτων σε σχέση με τις πληροφορίες που είναι εγγενείς μόνο στα δεδομένα → Δείκτης Silhouette, δείκτης Davies-Bouldin (DB), δείκτης Calinski-Harabasz (CH), δείκτης Dunn (Eréndira Rendón, 2011).

  ✓*Σχετικά κριτήρια*: Αξιολοόγηση ποιότητα συστάδας συγκρίνοντας την με άλλα σχήματα συσταδοποίησης, που προκύπτουν από τον ίδιο αλγόριθμο αλλά με διαφορετικές τιμές παραμέτρων.

- **Ερμηνεία των αποτελεσμάτων:** Ενοποίηση αποτελεσμάτων συσταδοποίησης με άλλα πειραματικά στοιχεία και αναλύσεις για εξαγωγή σωστών συμπερασμάτων.

Μέθοδοι κατηγοριοποίησης
Συσταδοποίηση Δύο Βημάτων (TwoStep Cluster Analysis)

- 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 έχει αρκετά επιθυμητά χαρακτηριστικά που τον διαφοροποιούν από τις παραδοσιακές τεχνικές ομαδοποίησης:
  1. 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.
- Χρήση μιας μέτρηση απόστασης πιθανότητας που υποθέτει ότι οι μεταβλητές στο μοντέλο συστάδας είναι ανεξάρτητες.
- Κάθε συνεχής μεταβλητή θεωρείται ότι έχει κανονική (Γκαουσιανή - Gaussian) κατανομή
- Κάθε κατηγορική μεταβλητή θεωρείται ότι έχει μια πολυωνυμική κατανομή.
- Διαδικασία αρκετά ισχυρή σε παραβιάσεις τόσο της υπόθεσης της ανεξαρτησίας όσο και των υποθέσεων κατανομής.

Μεθοδολογία και εργαλεία ανάλυσης

- Ανάλυση συναισθήματος των tweets → Χρήση γλώσσας προγραμματισμού Python και κατάλληλων βιβλιοθηκών επεξεργασίας φυσικής γλώσσας.
- Εκκαθάριση και κανονικοποίηση των δεδομένων → «Διάβασμα» κάθε tweet να από τον αλγόριθμο ανάλυσης συναισθήματος.
- Αξιοποίηση δεδομένα ανάλυσης των 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.

Μεθοδολογία και εργαλεία ανάλυσης

Περιγραφή δεδομένων

- Tweets που αφορούν τον κλάδο της αυτοκινητοβιομηχανίας
- Περίοδος διεξαγωγής του Super Bowl (Φεβρουάριο, 2014)
- Super Bowl
  - ✓ Σημαντικότερος και δημοφιλέστερος ετήσιος αγώνας του πρωταθλήματος του αμερικανικού ποδοσφαίρου και είναι το πρώτο σε τηλεθέαση γεγονός κάθε χρόνο στις ΗΠΑ
  - ✓ Δεύτερο παγκοσμίως σε ετήσια αθλητικά γεγονότα μετά τον τελικό του ΟΥΕΦΑ Τσάμπιονς Λιγκ
  - ✓ Μεγάλη δημοσιότητα → Σημαντικό βήμα προβολής νέων προϊόντων → Προσέλκυση νέων αγοραστών και αύξηση των πωλήσεων
  - ✓ Έρευνα της εταιρείας τεχνολογίας BrandAds (2014):
    - ➢ Αύξηση πιθανότητας των θεατών να αγοράσουν ένα προϊόν κατά 6.6 %, όταν προβάλλεται κατά την διάρκεια του Super Bowl.
- 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).

Μεθοδολογία και εργαλεία ανάλυσης

Περιγραφή δεδομένων

- Tweets που αφορούσαν τις εξής αυτοκινητοβιομηχανίες:
  - ✓ Audi
  - ✓ Chevrolet
  - ✓ Chrysler
  - ✓ 4KIA
  - ✓ Volkswagen

- Ανάλυσή τους με την εφαρμογή LIWC2007
  - ✓ Αντιστοίχηση κάθε λέξης του κάθε tweet (target words) σε μία λέξη λεξικού (dictionary words), αυξάνοντας την βαθμολογία της κάθε μιας από τις 64 μεταβλητές εξόδου.
  - ✓ Αποθήκευση σε αρχεία csv (comma-separated values).
  - ✓ Μετατροπή των αρχείων csv σε αρχεία excel με ταυτόχρονο μετασχηματισμό σε μορφή κατάλληλη για την ανάλυση που ακολούθησε.

Μεθοδολογία και εργαλεία ανάλυσης

Python

- Απλή αλλά ταυτόχρονα ισχυρή γλώσσα προγραμματισμού
- Κατάλληλη για επεξεργασία γλωσσικών δεδομένων
- Εύκολη στην εκμάθησή της
- Διαφανής σύνταξη και σημασιολογία
- Καλή λειτουργικότητα στον χειρισμό συμβολοσειρών
- Εξαιρετικά ευανάγνωστη
- 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 .
- Ως ερμηνευτική γλώσσα, 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.

Μεθοδολογία και εργαλεία ανάλυσης

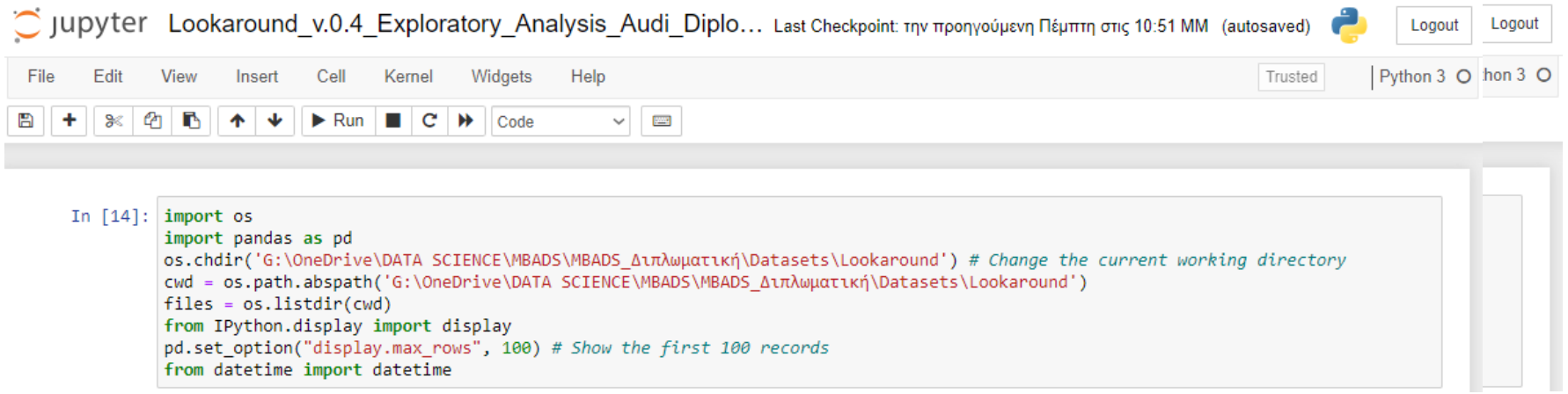
Python

- Με την εγκατάσταση της, εγκαθίσταται και μια εκτεταμένη βασική βιβλιοθήκη που περιέχει στοιχεία (components) για γραφικό προγραμματισμό, αριθμητική επεξεργασία και συνδεσιμότητα ιστού.
- Πολυάριθμες βιβλιοθήκες για κάθε είδους προγραμματιστική εργασία με εύκολη εγκατάσταση, εισαγωγή και ανάπτυξη.
- Μεγάλη κοινότητα προγραμματιστών ανοιχτού κώδικα
- Δημοφιλής στον τομέα της επιστήμης των δεδομένων, της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας.
- Για την παρούσα διπλωματική χρησιμοποιήθηκε η έκδοση 3.8 της Python

Μεθοδολογία και εργαλεία ανάλυσης

Jupyter Notebook

- Επεξεργασία των δεδομένων και ανάπτυξη κώδικα της ανάλυσης συναισθήματος → Jupyter Notebook → Διαδραστικό περιβάλλον ανάπτυξης κώδικα (Interactice Development Environments – IDE) που χαρακτηρίζεται από ευελιξία και ευκολία στη χρήση του.
- Κομμάτι της διανομής Python Anaconda.
- Η Python Anaconda διαθέτει μια σειρά από πολύ χρήσιμες βιβλιοθήκες και επιτρέπει την εύκολη διαχείριση/εγκατάσταση/επεγκατάσταση επιπλέον εφαρμογών και βιβλιοθηκών.



Μεθοδολογία και εργαλεία ανάλυσης

Python - Βιβλιοθήκες

- **NumPy**

✓ Θεμελιώδες πακέτο για επιστημονικούς υπολογισμούς
✓ 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.
✓ Παρέχει μερικές από τις εξαιρετικά βελτιστοποιημένες δομές δεδομένων, τα n-darrays.

- **pandas**

✓ Βιβλιοθήκη ανοιχτού κώδικα για ανάλυση και χειρισμό δεδομένων, που βασίζεται στην NumPy.
✓ 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.

Μεθοδολογία και εργαλεία ανάλυσης

Python - Βιβλιοθήκες

- **scikit-learn**

✓ Ανοιχτού κώδικα βιβλιοθήκη μηχανική μάθησης

✓ Υποστηρίζει αλγόριθμους τόσο επιβλεπόμενης όσο και μη επιβλεπόμενη μάθησης, όπως ταξινόμηση, παλινδρόμηση, συσταδοποίηση

✓ Διαθέτει εργαλεία για προεπεξεργασία δεδομένων κειμένου, εξαγωγή χαρακτηριστικών και κανονικοποίηση. Έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες NumPy και SciPy.

- **NLTK**

✓ 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)

Μεθοδολογία και εργαλεία ανάλυσης

Python - Βιβλιοθήκες

- **re (RegEx – Regular Expression)**

✓ 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.

✓ Έλεγχος συμβολοσειράς (πχ. κείμενο) εάν 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.

- **TextBlob**

✓ Βιβλιοθήκη για την επεξεργασία δεδομένων κειμένου.

✓ Επεξεργασίας φυσικής γλώσσας (NLP), όπως προσθήκη μέρους-του-λόγου ετικετών (part-of-speech tagging), εξαγωγή ονοματικής φράσης , ανάλυση συναισθήματος, ταξινόμηση, μετάφραση και πολλά άλλα.

Μεθοδολογία και εργαλεία ανάλυσης

Python - Βιβλιοθήκες

- **spaCy**

✓ Δωρεάν βιβλιοθήκη ανοιχτού κώδικα για προηγμένη επεξεργασία φυσικής γλώσσας.

✓ Δημιουργία συστημάτων εξαγωγής πληροφοριών ή κατανόησης φυσικής γλώσσας ή για την προεπεξεργασία κειμένου για βαθιά μάθηση.

- **matplotlib**

✓ Βασική βιβλιοθήκη για την δημιουργία γραφημάτων

✓ Δημιουργία διαφορετικών τύπων αναφορών απεικόνισης, όπως γραφήματα γραμμών, γραφήματα διασποράς, ιστογράμματα, γραφήματα ράβδων, γραφήματα πίτας, γραφήματα πλαισίων, καθώς επίσης και υποστήριξη τρισδιάστατης σχεδίασης.

Μεθοδολογία και εργαλεία ανάλυσης

Python - Βιβλιοθήκες

- **Seaborn**

✓ Βιβλιοθήκη απεικόνισης δεδομένων που βασίζεται στο matplotlib.

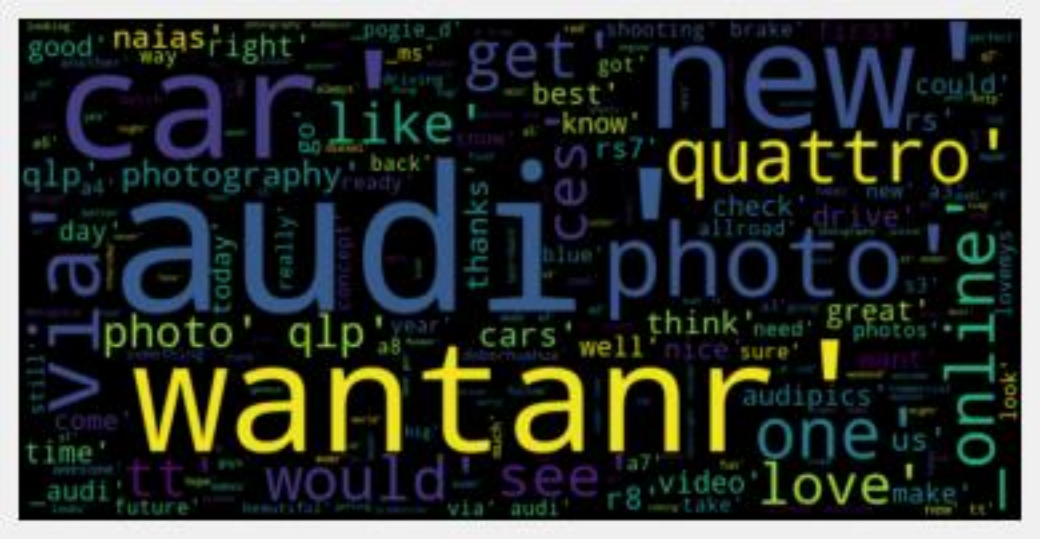
✓ Σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών.

- **Wordcloud**

✓ Απεικόνισης λέξεων που εμφανίζονται μέσα σε ένα κείμενο, σε δομές που μοιάζουν με «σύννεφο».

✓ Απεικόνιση με διαφορετικά χρώματα και μεγέθη, ανάλογα με τη συχνότητά τους ή τη σημαντικότητά τους

✓ Γρήγορος τρόπος να αποκτήσουμε μια εικόνα του περιεχομένου ενός κειμένου.



Μεθοδολογία και εργαλεία ανάλυσης
SPSS (Statistical Package for the Social Sciences)

- Επεξεργασία και στατιστική ανάλυση όλων των ειδών δεδομένων
- Υποστήριξη όλων των μορφών αρχείων δομημένων δεδομένων όπως υπολογιστικά φύλλα από το MS Excel ή το OpenOffice, αρχεία απλού κειμένου (.txt ή .csv), σχεσιακές βάσεις δεδομένων (SQL), αρχεία Stata και SAS.
- 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.
- Ανάλυση δεδομένων για περιγραφική στατιστική, προβλέψεις αριθμητικών αποτελεσμάτων και προσδιορισμό ομάδων (ομαδοποίηση).

Μεθοδολογία και εργαλεία ανάλυσης
SPSS (Statistical Package for the Social Sciences)

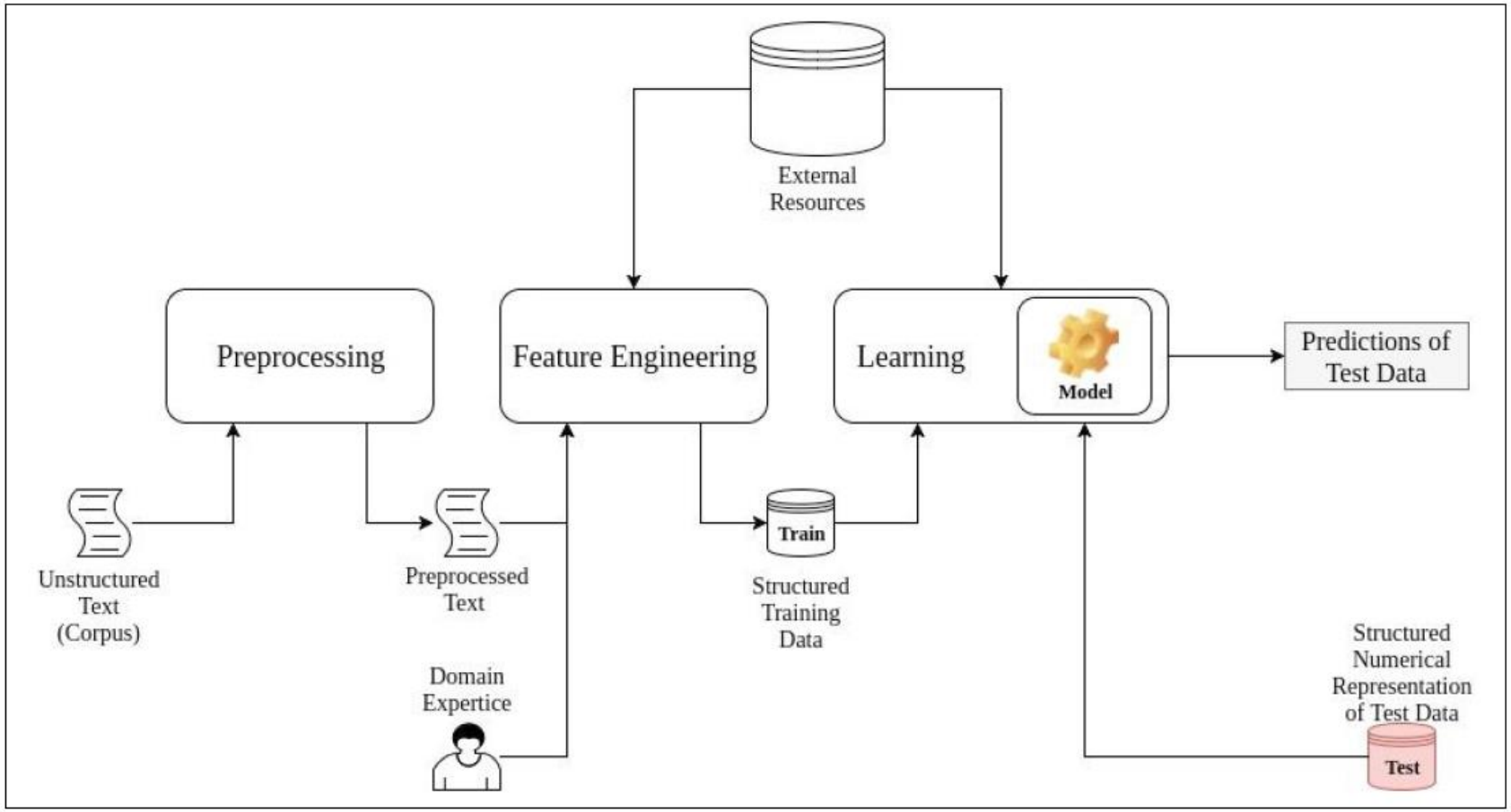
- Μετασχηματισμός δεδομένων, δημιουργία γραφημάτων, διαχείριση δεδομένων (επιλογή περιπτώσεων, αναδιαμόρφωση αρχείων, δημιουργία παράγωγων δεδομένων) και τεκμηρίωση δεδομένων (ένα λεξικό μεταδεδομένων αποθηκεύεται στο αρχείο δεδομένων).

- Διαβάζει και 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 ODBC και SQL.

- 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.

Ερευνητική διαδικασία

Ανάλυση Συναισθήματος – Στάδια επεξεργασίας φυσικής γλώσσας



Ερευνητική διαδικασία

Βήματα Επεξεργασίας Φυσικής Γλώσσας (NLP Pipeline)

Προετοιμασία δεδομένων

- Συλλογή/εξόρυξη δεδομένων και καθαρισμός
- Συλλογή δεδομένων με τη βοήθεια του Twitter API.
- Επεξεργασία από την εφαρμογή LIWC → Ανάλυση και εξαγωγή βαθμολογιών των λέξεων του κάθε tweet σύμφωνα με την αντιστοίχιση στην κάθε μεταβλητή του λεξικού του LIWC.
- Εξαγωγή σε αρχεία csv → Μετατροπή σε αρχεία excel για την περαιτέρω επεξεργασία τους στην Python και το SPSS.
- Αρχική άναρχη δομή δεδομένων → Βήματα μετατροπής της βάσης δεδομένων σε κατάλληλη μορφή για την ανάλυση συναισθήματος:

✓ Εισαγωγή της βιβλιοθήκης pandas

✓ Εισαγωγή των δεδομένων από αρχείο excel που περιλαμβάνει τα tweets που αφορούν την αντίστοιχη αυτοκινητοβιομηχανία

✓ Αφαίρεση εγγραφών που δεν περιείχαν καθόλου δεδομένα (αναφέρονται ως NaN – Not a Number).

✓ Χρήση των δεδομένων κειμένου των tweets → Δημιουργία νέου dataframe που περιείχε μόνο την στήλη Tweet.

✓ Αφαίρεση πολλαπλών εγγραφών των ίδιων tweets.

✓ Αφαίρεση κανονικών εκφράσεων

Ερευνητική διαδικασία
Βήματα Επεξεργασίας Φυσικής Γλώσσας (NLP Pipeline)
Προεπεξεργασία

- Η προεπεξεργασία αποτελείται συνήθως από τις εξής διαδικασίες:

✓ **Διακριτοποίηση (Tokenization)**

- 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.
- Λειτουργία TweetTokenizer: Ανάλυση της άτυπης γλώσσας των μηνυμάτων του κοινωνικού μέσου Twitter (ετικέτες, emoticons, hashtags, σύντομες εκφράσεις κ.α) με στόχο την μετατροπή σε όσο το δυνατόν πιο φυσιολογική και περισσότερο κατανοητή μορφή.
- Αποθήκευση στοιχείων (elements) του κάθε tweet σε λίστα

✓ **Αφαίρεση λέξεων διακοπής (Stopwords removal)**

Ερευνητική διαδικασία
Βήματα Επεξεργασίας Φυσικής Γλώσσας (NLP Pipeline)
Προεπεξεργασία

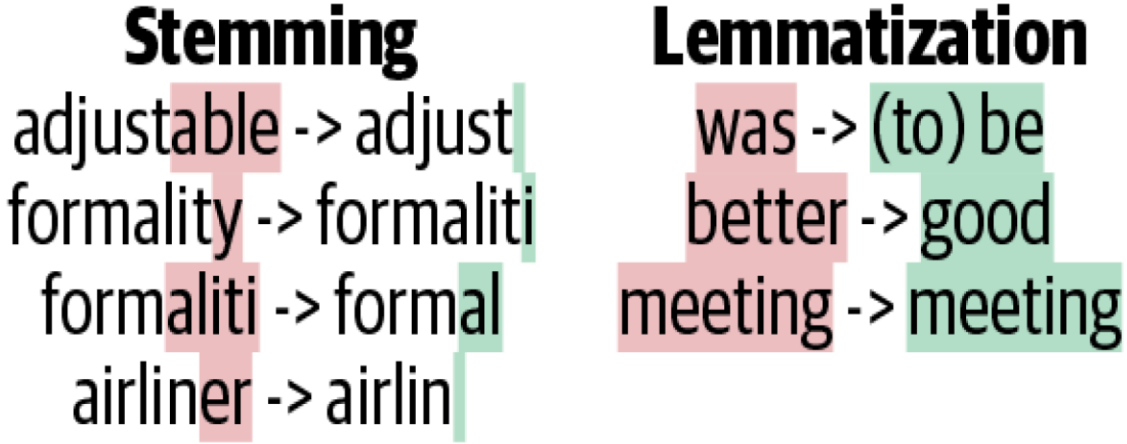
- ✓ **Αφαίρεση ειδικών χαρακτήρων (Remove special words)**
  - Σημεία στίξης
  - Ψηφία
  - Χαρακτήρες ASCII

- ✓ **Στελέχωση ή αποκοπή καταλήξεων (Stemming)**
  - Διαδικασία αφαίρεσης καταλήξεων και σμίκρυνση μιας λέξης σε κάποια βασική μορφή
  - 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

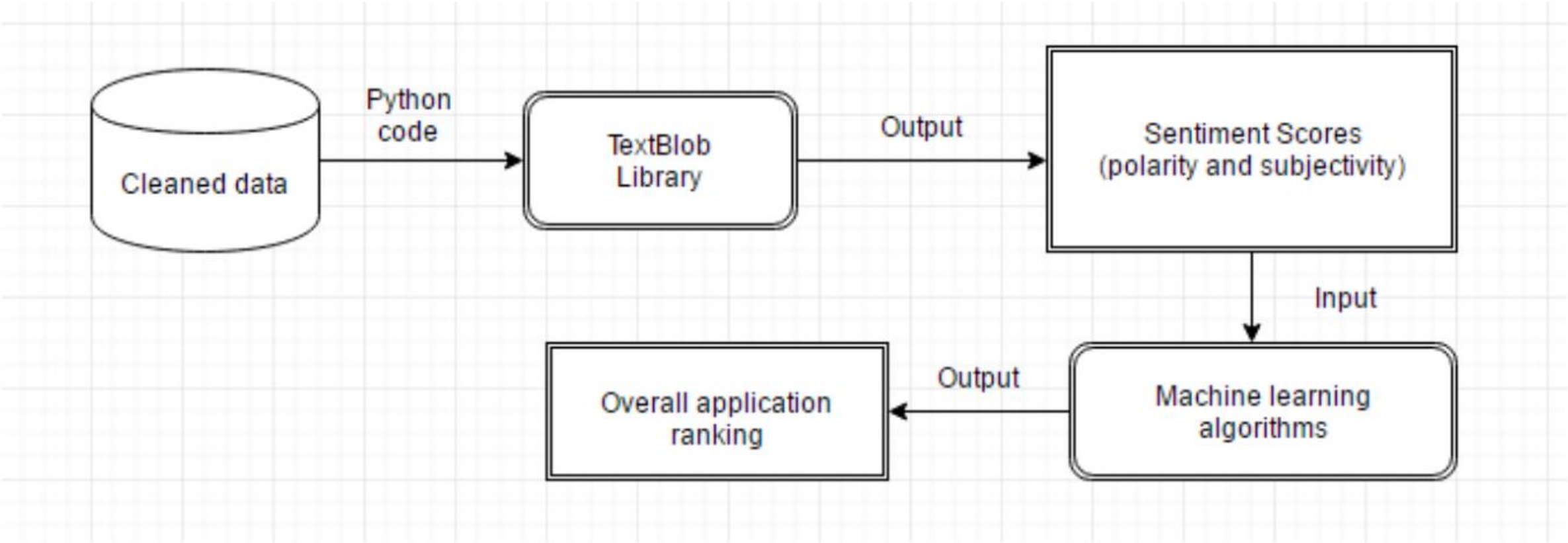
Ερευνητική διαδικασία
Βήματα Επεξεργασίας Φυσικής Γλώσσας (NLP Pipeline)
Προεπεξεργασία

✓ **Λημματοποίηση (lemmatizaion)**

- Διαδικασία αντιστοίχισης όλων των διαφορετικών μορφών μιας λέξης στη βασική της λέξη, ή λήμμα (lemma).
- 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.



Ερευνητική διαδικασία
Ανάλυση Συναισθήματος



Ερευνητική διαδικασία
Ανάλυση Συναισθήματος – TextBlob

- **Λειτουργία 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** 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.
  - Αγνοεί λέξεις που δεν γνωρίζει και εξετάζει λέξεις και φράσεις στις οποίες μπορεί να εκχωρήσει πολικότητα και μέσους όρους για να εξάγει την τελική βαθμολογία.

Ερευνητική διαδικασία

Ανάλυση Συναισθήματος – TextBlob (Αυτοκινητοβιοχανία Audi)

- Προεπεξεργασία δεδομένων → Μετατροπή του κάθε tweet σε λίστα στοιχείων → Εφαρμογή διαδικασιών διακριτοποίησης, αφαίρεσης ειδικών χαρακτήρων και στελέχωσης.
- Μετατροπή των tweets σε μορφή συμβολοσειράς, αποθηκεύοντάς τα στη στήλη Tweet_RFSA (**R**eady **F**or **S**entiment **A**nalysis).

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 chopperðÿˆž @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

Ερευνητική διαδικασία
Ανάλυση Συναισθήματος – TextBlob

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

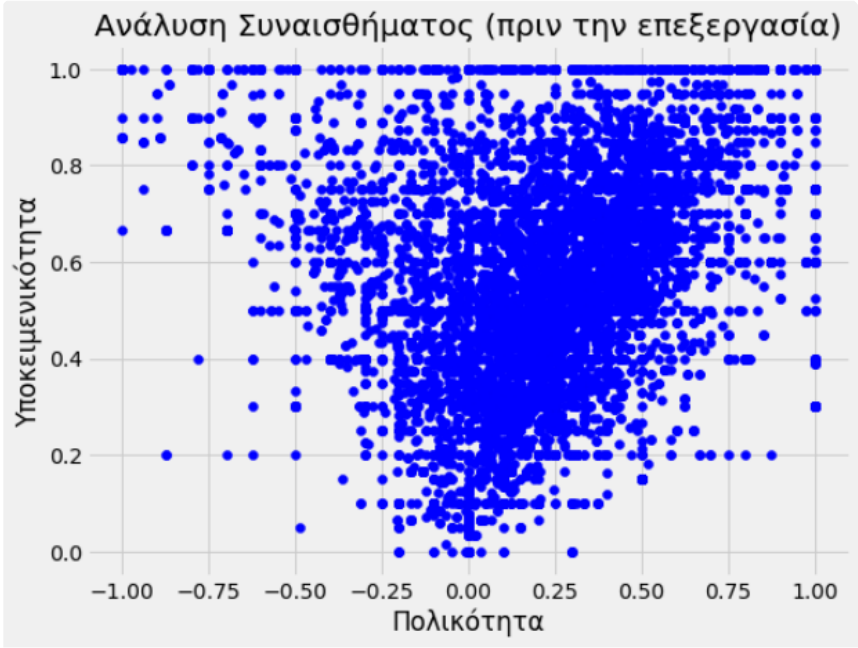
20541 rows × 2 columns

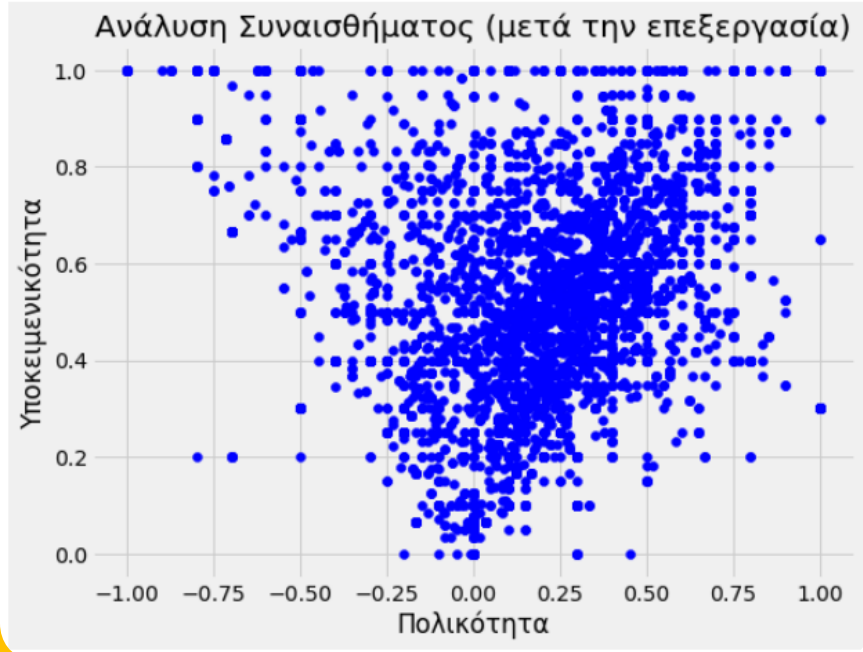
Ανάλυση Συναισθήματος σε Tweets *πριν και μετά την επεξεργασία* Αυτοκινητοβιομηχανία Audi (με στελέχωση)

```
# 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
```

	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





Ανάλυση Συναισθήματος

Διάγραμμα διασποράς – Audi (με στελέχωση)

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

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

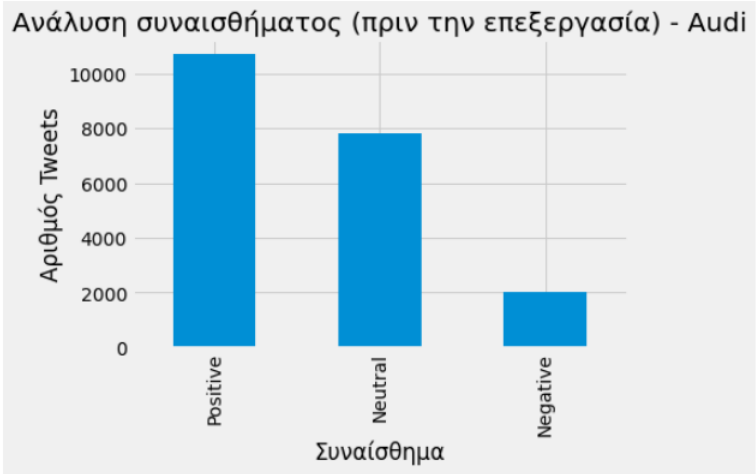
```
Out[41]: Positive    10734
         Neutral      7808
         Negative     1999
         Name: Analysis_Tweet, dtype: int64
```

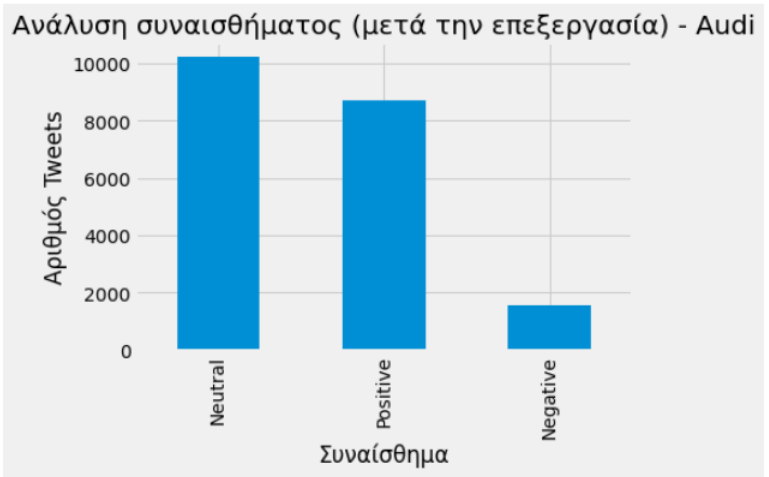
```
In [42]: # Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

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

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

Ανάλυση Συναισθήματος – Ραβδογράμματα πλήθους θετικών, αρνητικών και ουδέτερων Tweets – Αυτοκινητοβιομηχανία Audi (με στελέχωση)

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_Twe

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()
```





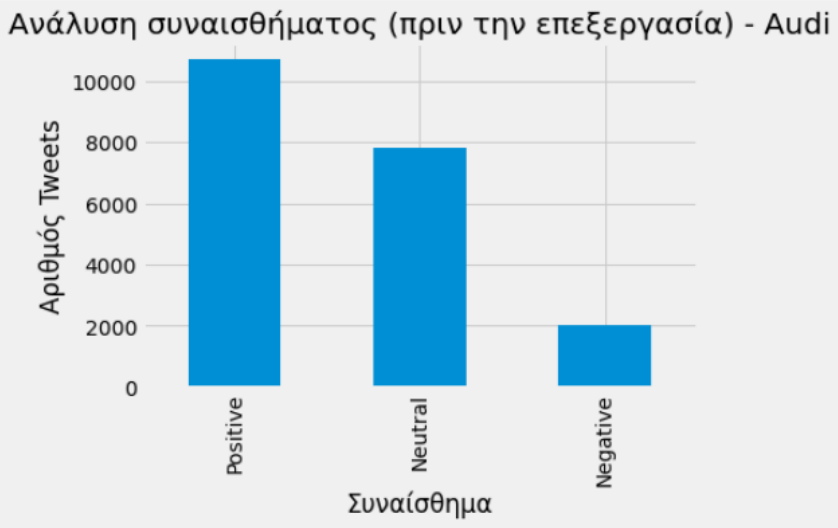
Ανάλυση Συναισθήματος – Πίνακας αποτελεσμάτων- Audi (πριν και μετά την επεξεργασία) (με στελέχωση)

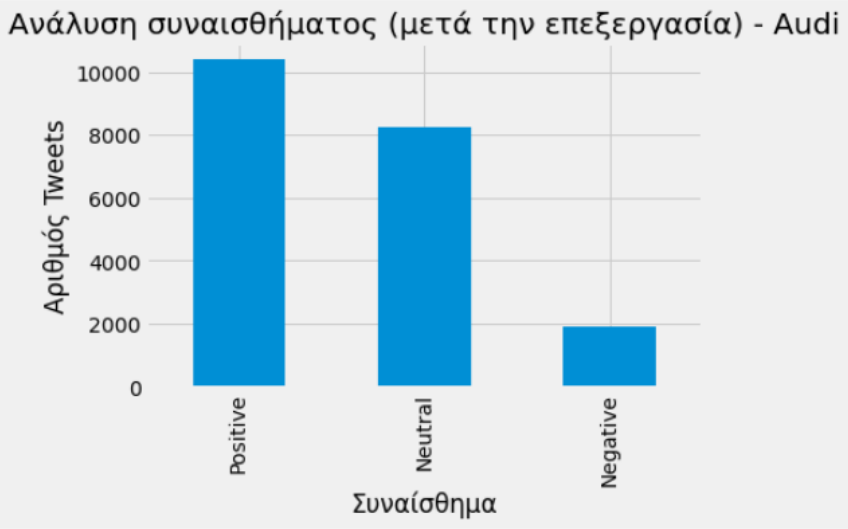
- Η μεταστροφή αυτή στα ποσοστά παρατηρήθηκε στην περίπτωση που κατά την διάρκεια της επεξεργασίας εφαρμόσαμε την διαδικασία της στελέχωσης και πιθανότατα οφείλεται στην αναντιστοιχία των λέξεων που απαρτίζουν τα tweets μετά την στελέχωση και των εκφράσεων/λέξεων του λεξικού που χρησιμοποιεί η βιβλιοθήκη TextBlob για να εξάγει την πολικότητα και την αντικειμενικότητα.

```
# 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

Ανάλυση Συναισθήματος – Audi (χωρίς στελέχωση)






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

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

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

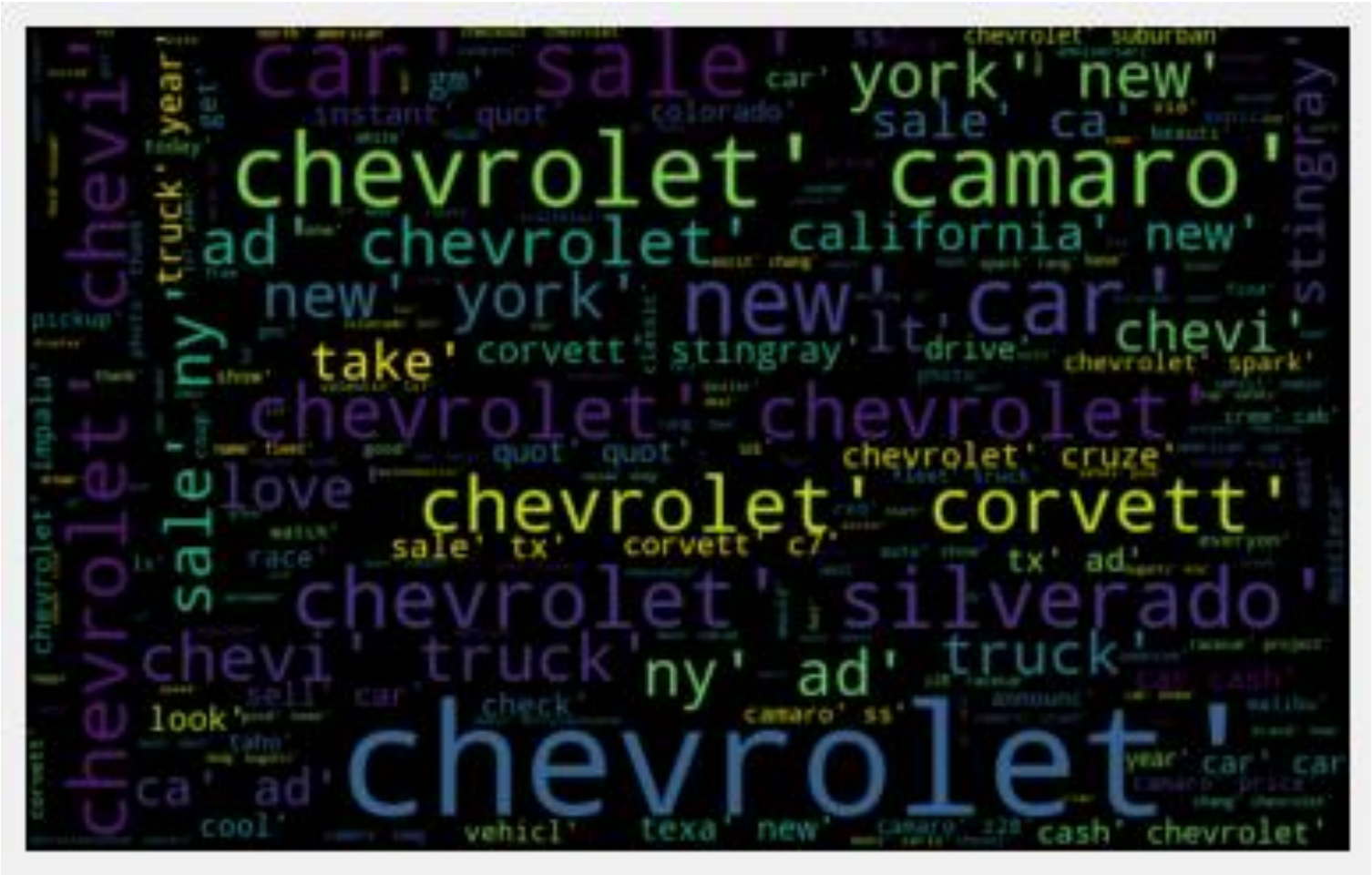
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

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

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

```
# 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

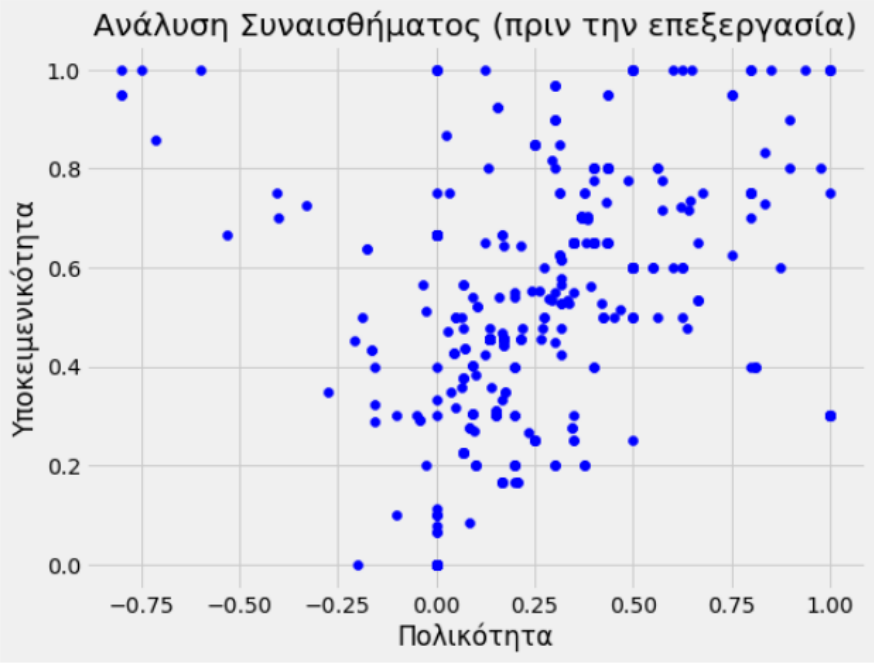
Ανάλυση Συναισθήματος σε Tweets – Σύννεφο λέξεων Αυτοκινητοβιομηχανία Chevrolet (με στελέχωση)

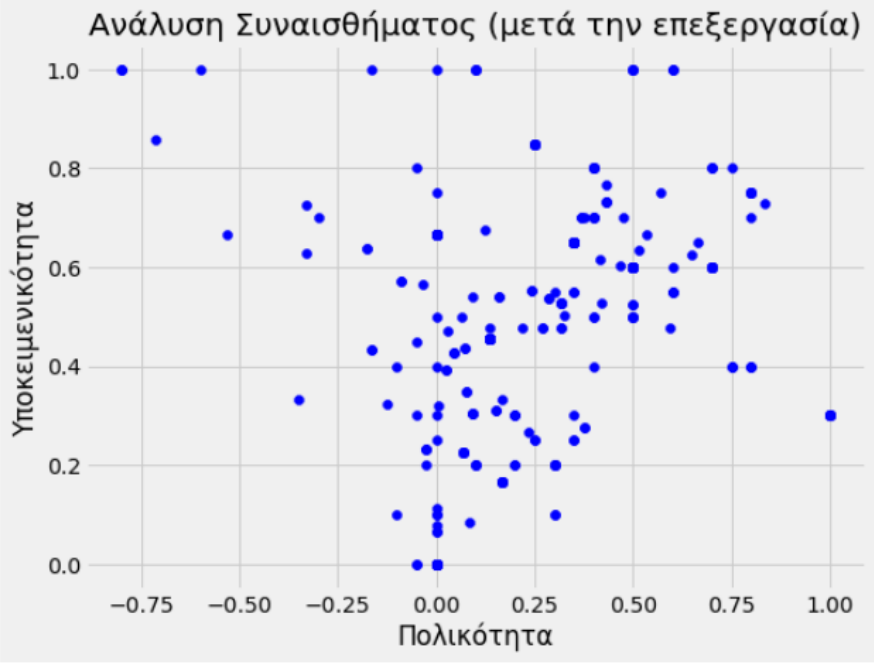


Ανάλυση Συναισθήματος σε Tweets *πριν και μετά την επεξεργασία*
Αυτοκινητοβιομηχανία Chevrolet (με στελέχωση)

	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





Ανάλυση Συναισθήματος

Διάγραμμα διασποράς – Chevrolet (με στελέχωση)

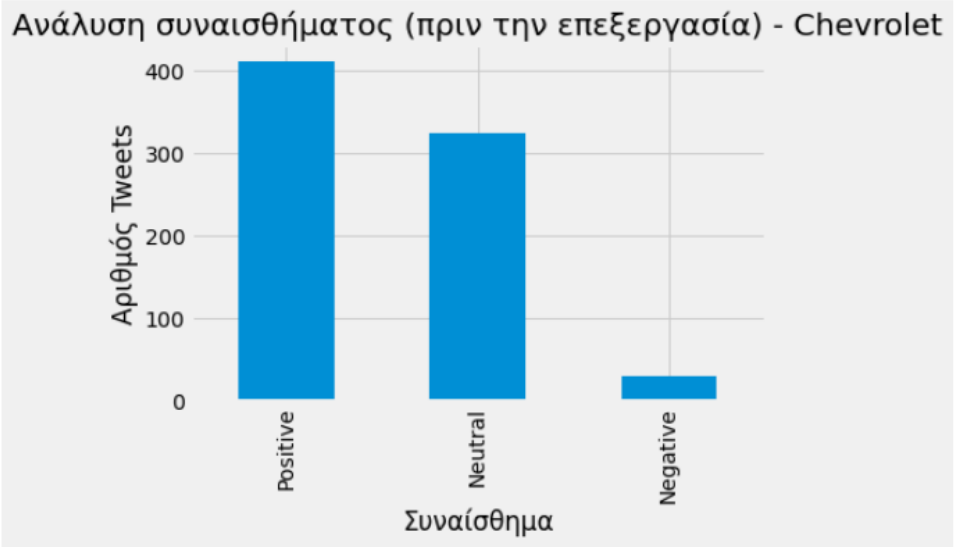
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

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

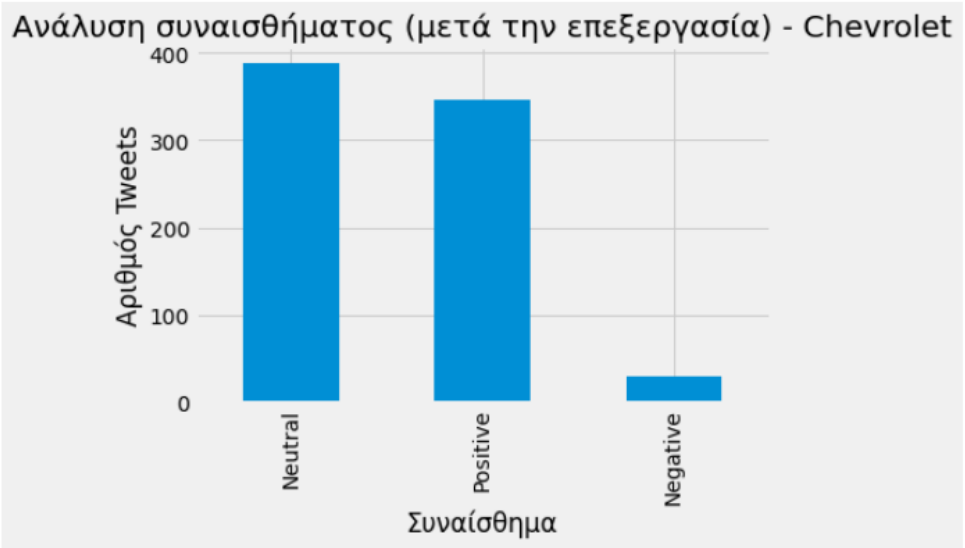
```
Positive    412
Neutral     325
Negative     29
Name: Analysis_Tweet, dtype: int64
```

```
# 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
```

Ανάλυση Συναισθήματος – Ραβδογράμματα πλήθους θετικών, αρνητικών και ουδέτερων Tweets – Αυτοκινητοβιομηχανία Chevrolet (με στελέχωση)

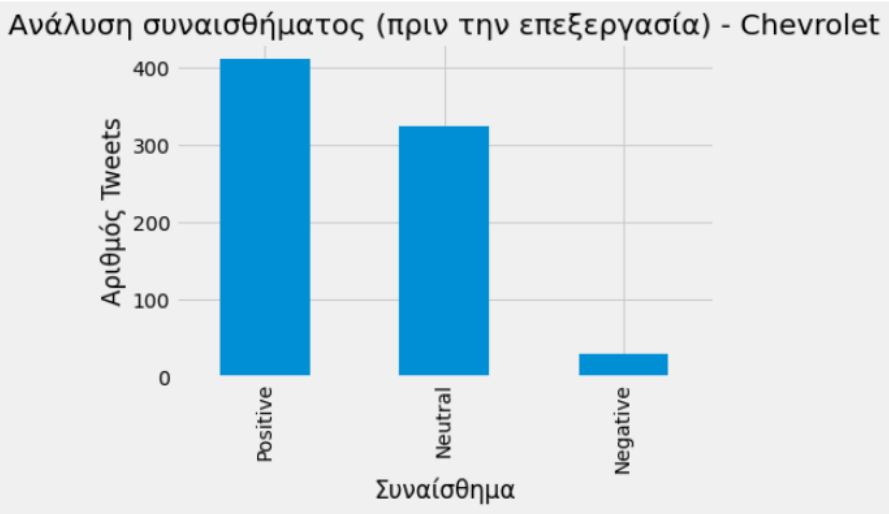


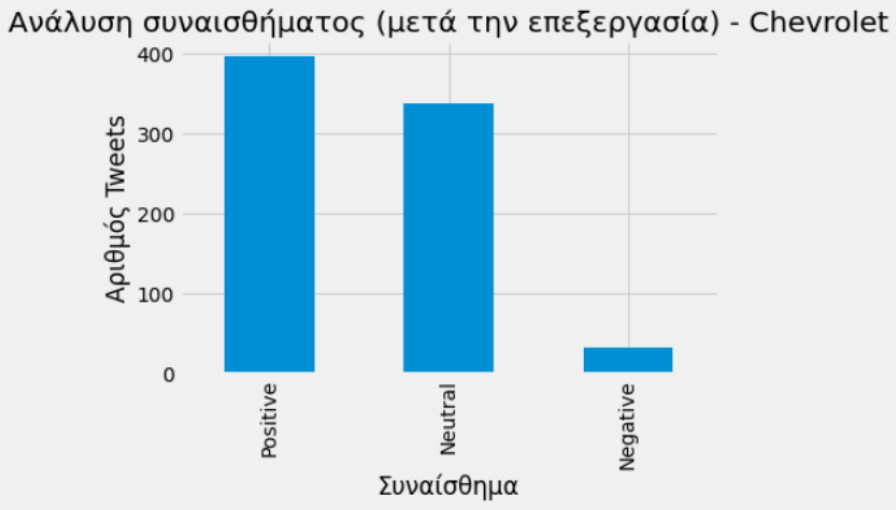


Πίνακας ποσοστών

	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

Ανάλυση Συναισθήματος – Chevrolet (χωρίς στελέχωση)






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

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

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

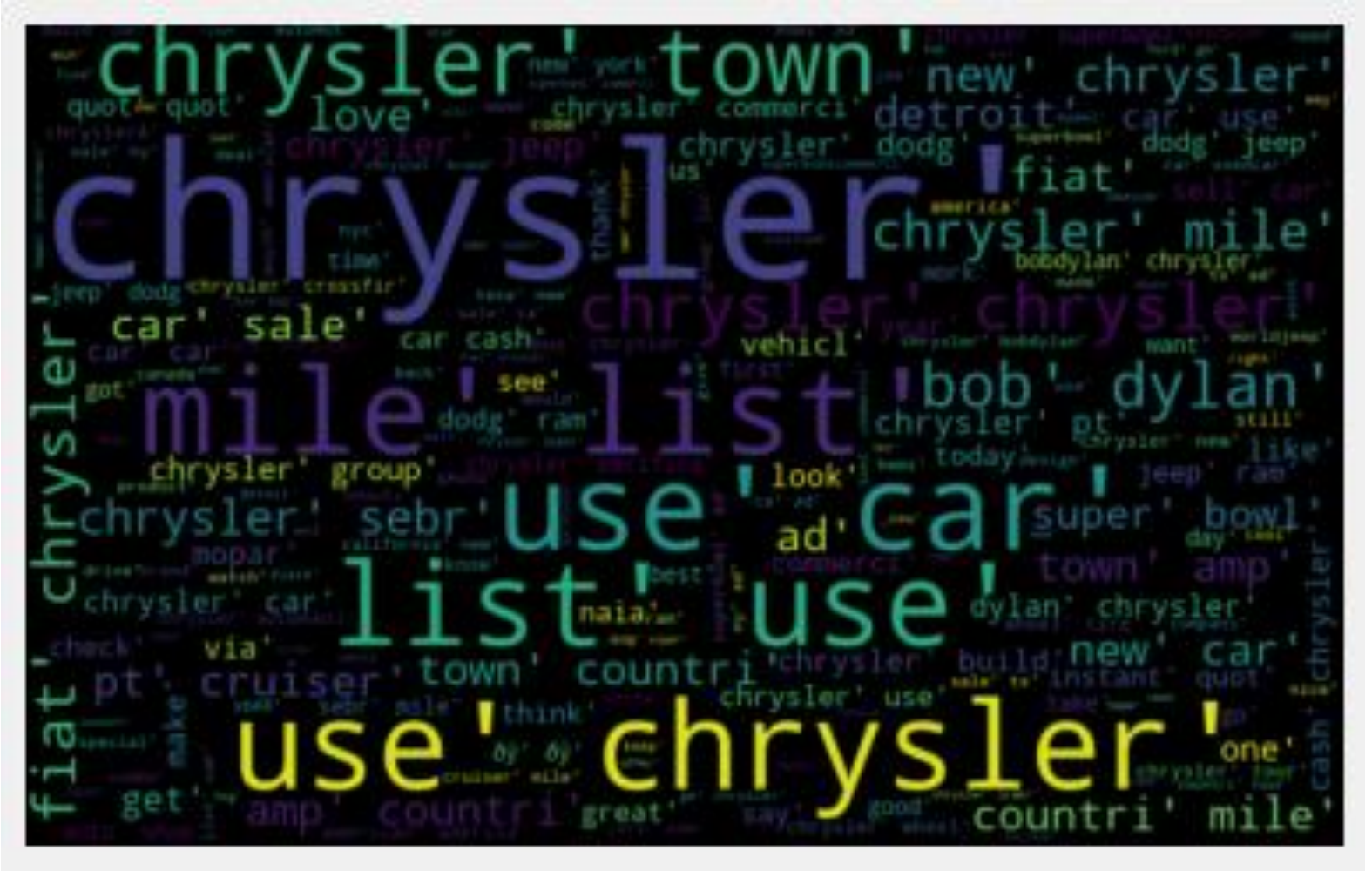
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

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

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

	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

Ανάλυση Συναισθήματος σε Tweets – Σύννεφο λέξεων

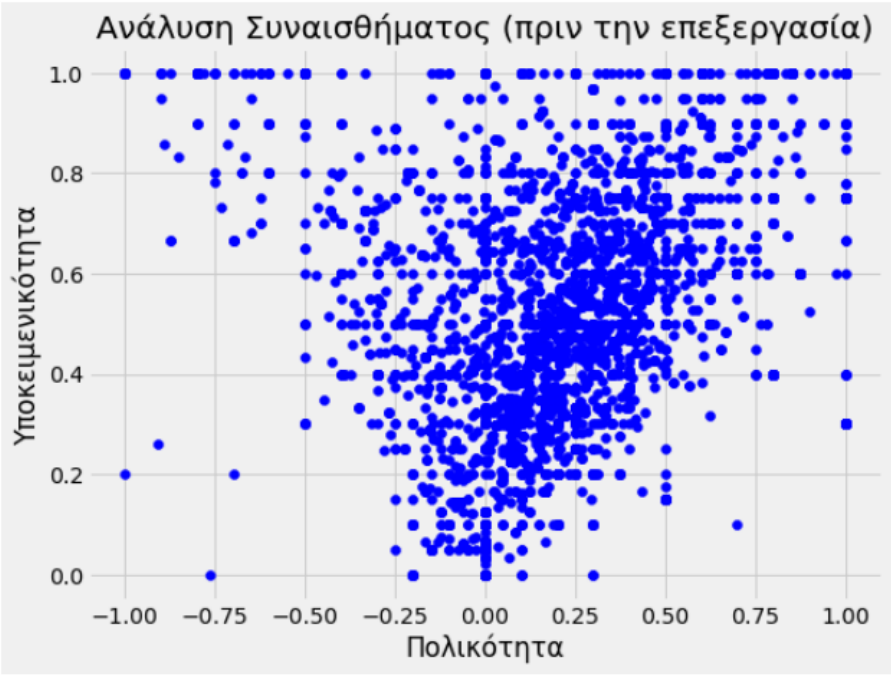
Αυτοκινητοβιομηχανία Chrysler (με στελέχωση)

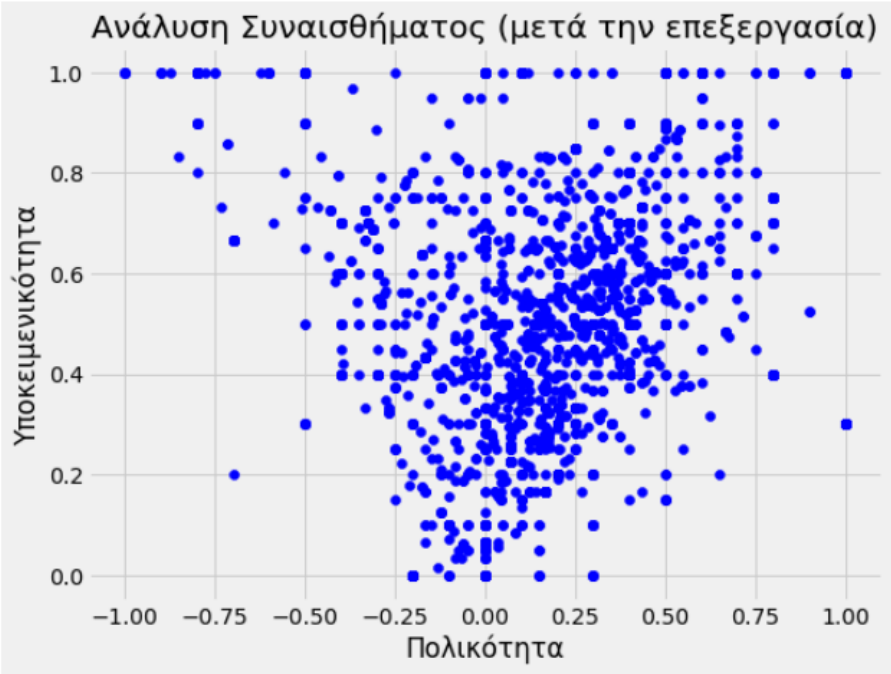


Ανάλυση Συναισθήματος σε Tweets *πριν και μετά την επεξεργασία*
Αυτοκινητοβιομηχανία Chrysler (με στελέχωση)

	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





Ανάλυση Συναισθήματος

Διάγραμμα διασποράς – Chrysler (με στελέχωση)

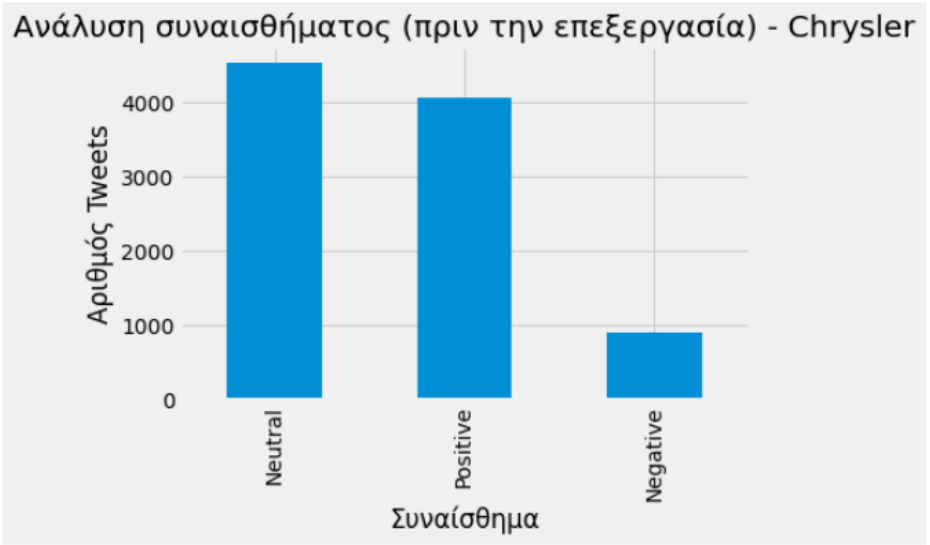
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

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

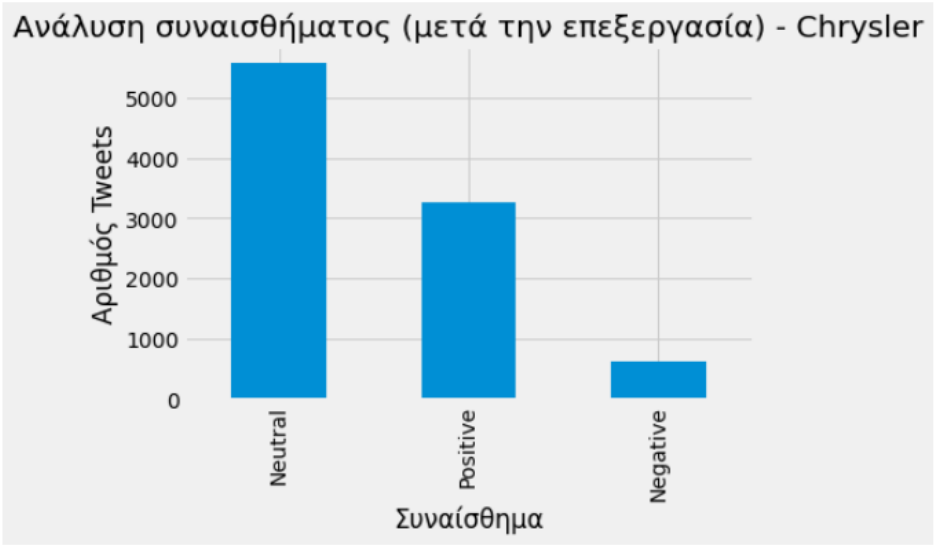
```
Neutral     4529
Positive    4056
Negative     899
Name: Analysis_Tweet, dtype: int64
```

```
# 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
```

Ανάλυση Συναισθήματος – Ραβδογράμματα πλήθους θετικών, αρνητικών και ουδέτερων Tweets – Αυτοκινητοβιομηχανία Chrysler (με στελέχωση)

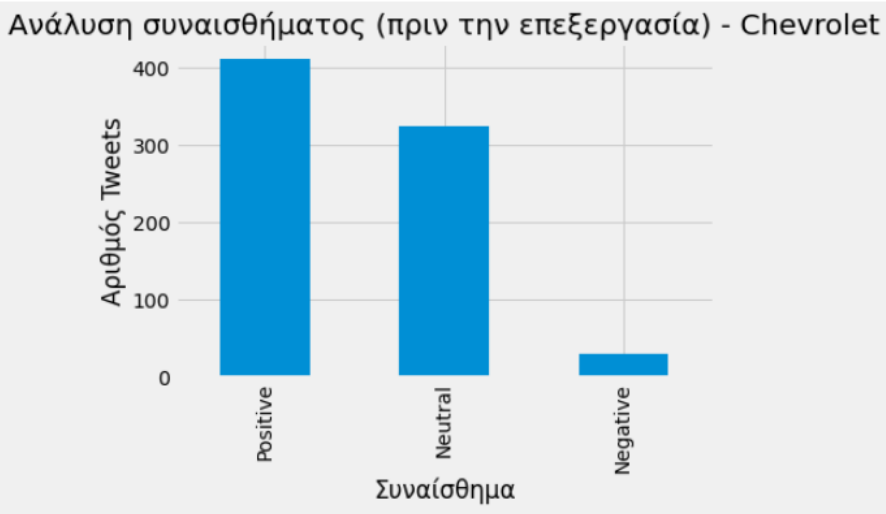


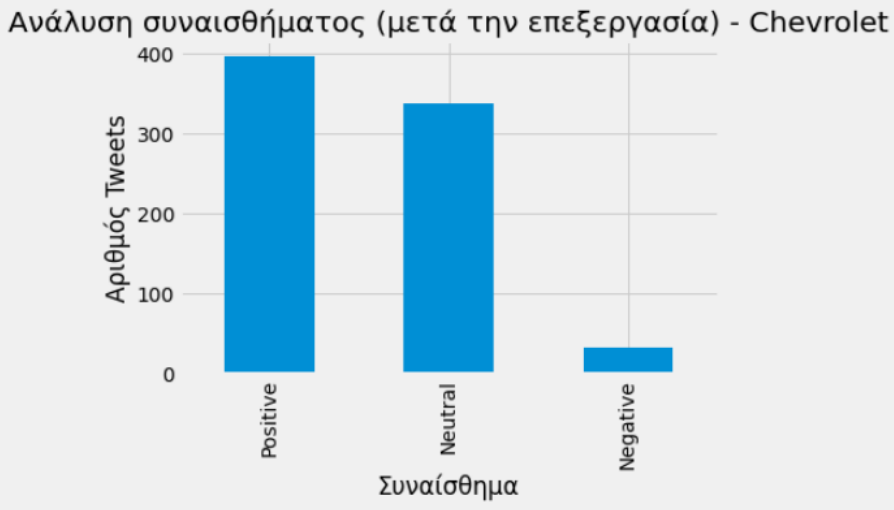


Πίνακας ποσοστών

	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

Ανάλυση Συναισθήματος – Chrysler (χωρίς στελέχωση)






```
# 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
```

	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

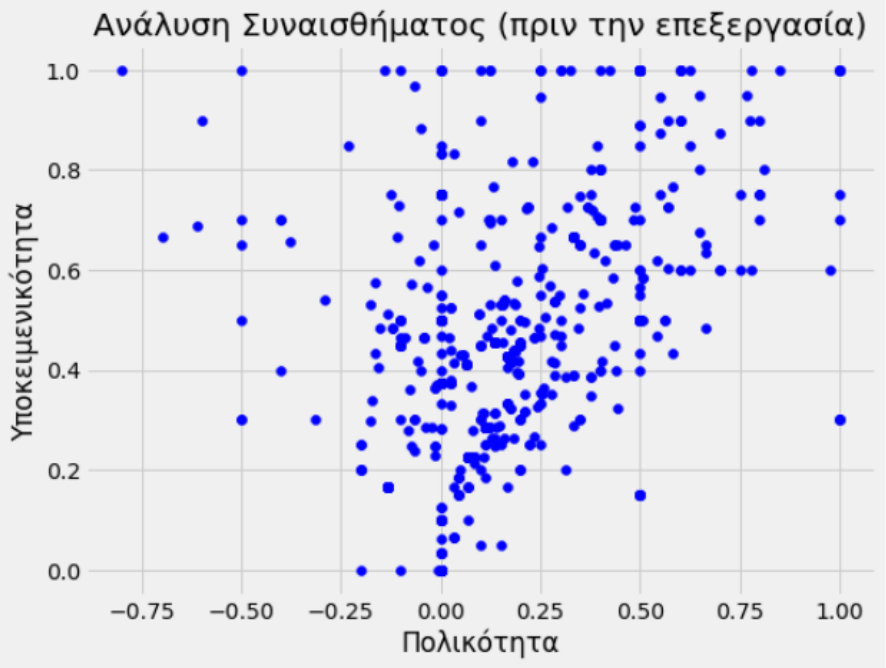
Ανάλυση Συναισθήματος σε Tweets – Σύννεφο λέξεων

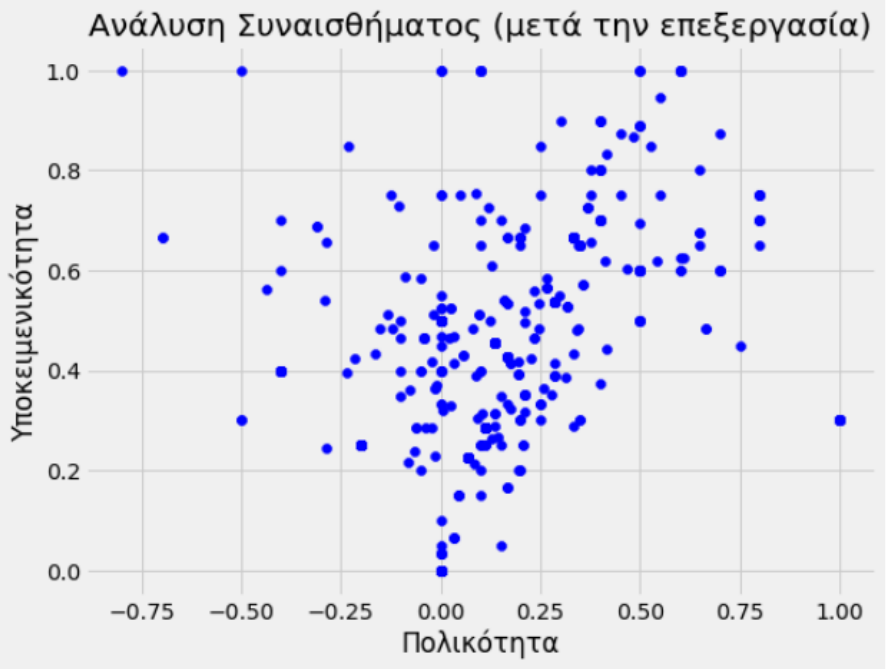
Αυτοκινητοβιομηχανία KIA (με στελέχωση)

Ανάλυση Συναισθήματος σε Tweets *πριν και μετά την επεξεργασία*
Αυτοκινητοβιομηχανία KIA (με στελέχωση)

	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





Ανάλυση Συναισθήματος
Διάγραμμα διασποράς – KIA
(με στελέχωση)

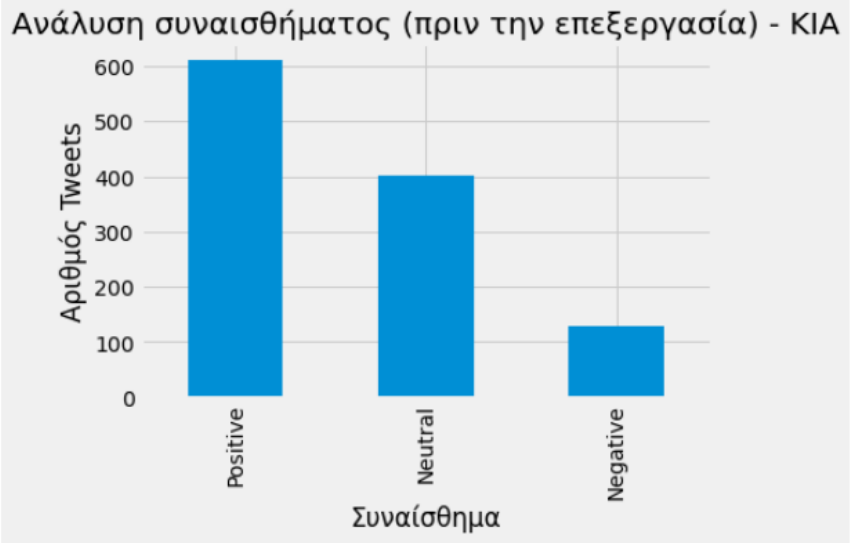
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

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

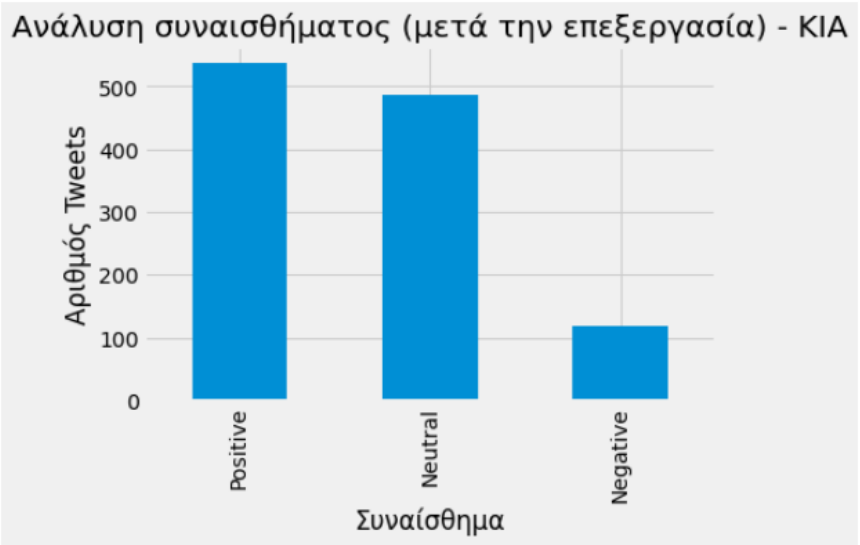
```
Positive    612
Neutral     403
Negative    128
Name: Analysis_Tweet, dtype: int64
```

```
# 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
```

Ανάλυση Συναισθήματος – Ραβδογράμματα πλήθους θετικών, αρνητικών και ουδέτερων Tweets – Αυτοκινητοβιομηχανία KIA (με στελέχωση)

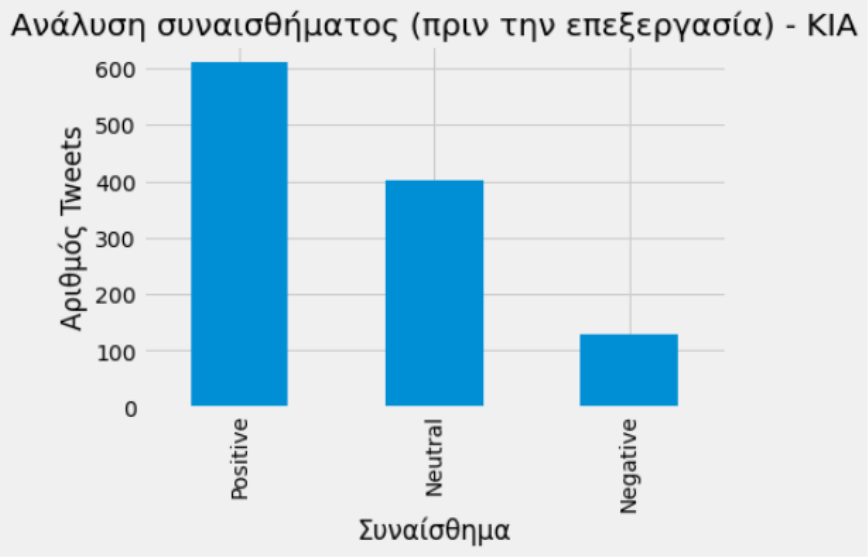


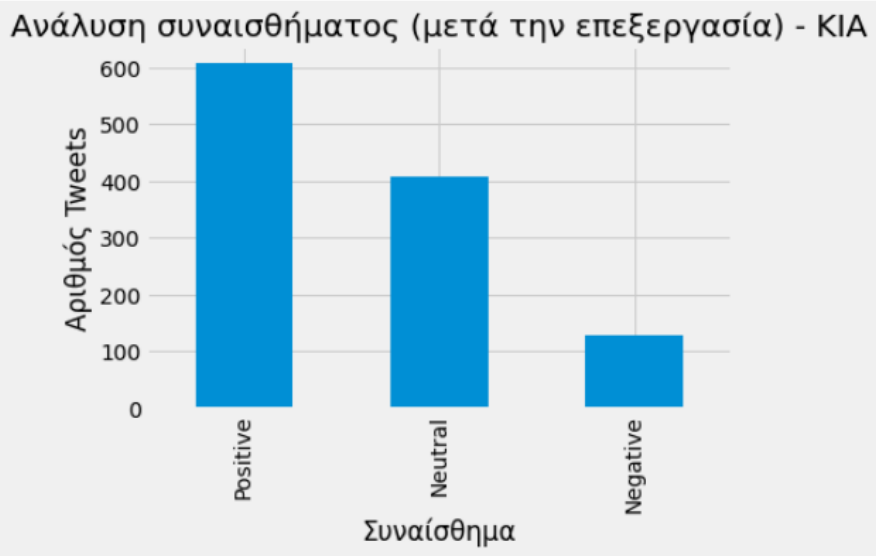


Πίνακας ποσοστών

	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

Ανάλυση Συναισθήματος – ΚΙΑ (χωρίς στελέχωση)






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

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

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

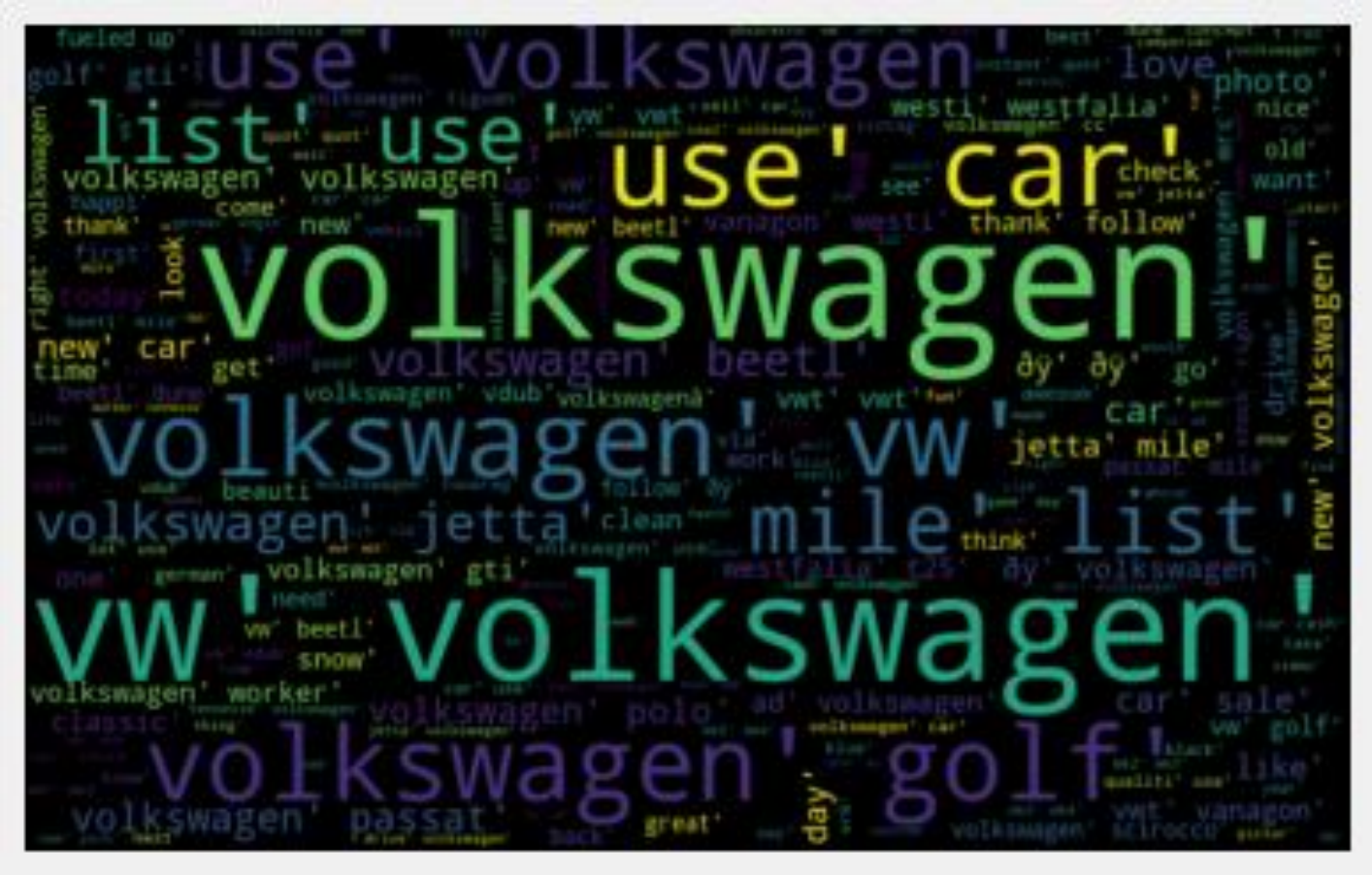
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

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

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

	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

Ανάλυση Συναισθήματος σε Tweets – Σύννεφο λέξεων

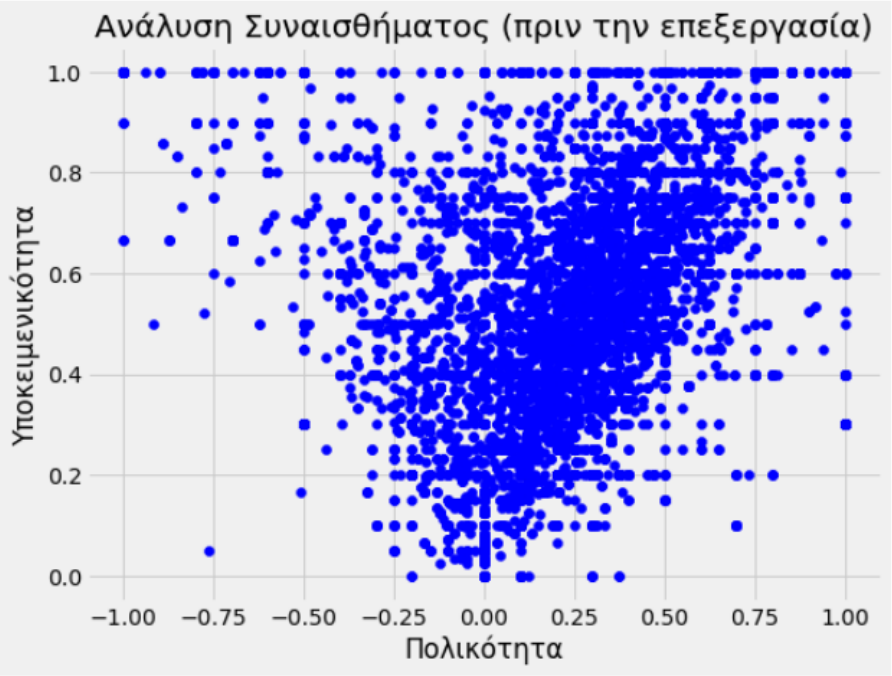
Αυτοκινητοβιομηχανία Volkswagen (με στελέχωση)

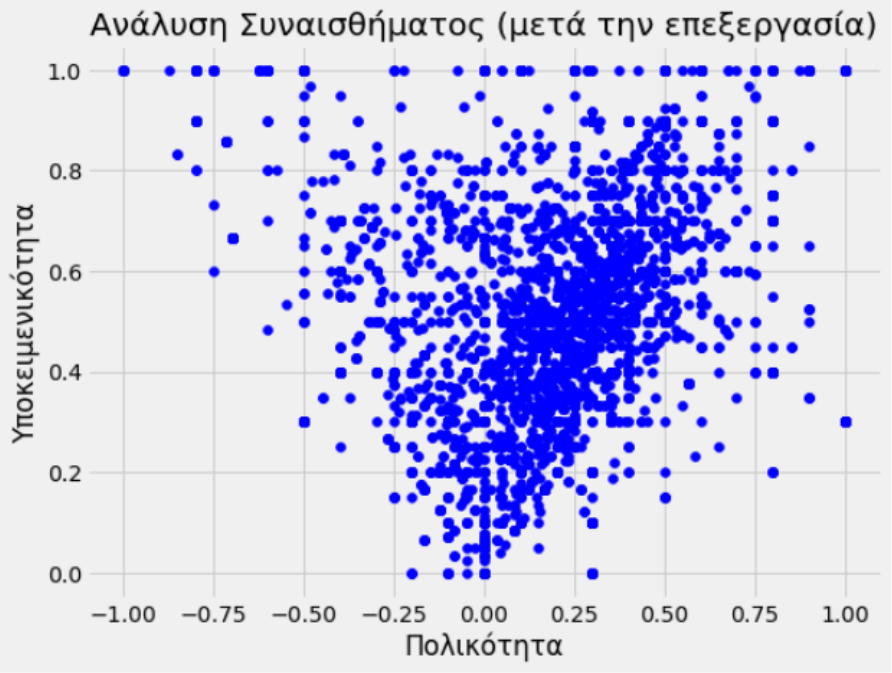


Ανάλυση Συναισθήματος σε Tweets *πριν και μετά την επεξεργασία*
Αυτοκινητοβιομηχανία Volkswagen (με στελέχωση)

	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





Ανάλυση Συναισθήματος

Διάγραμμα διασποράς – Volkswagen (με στελέχωση)

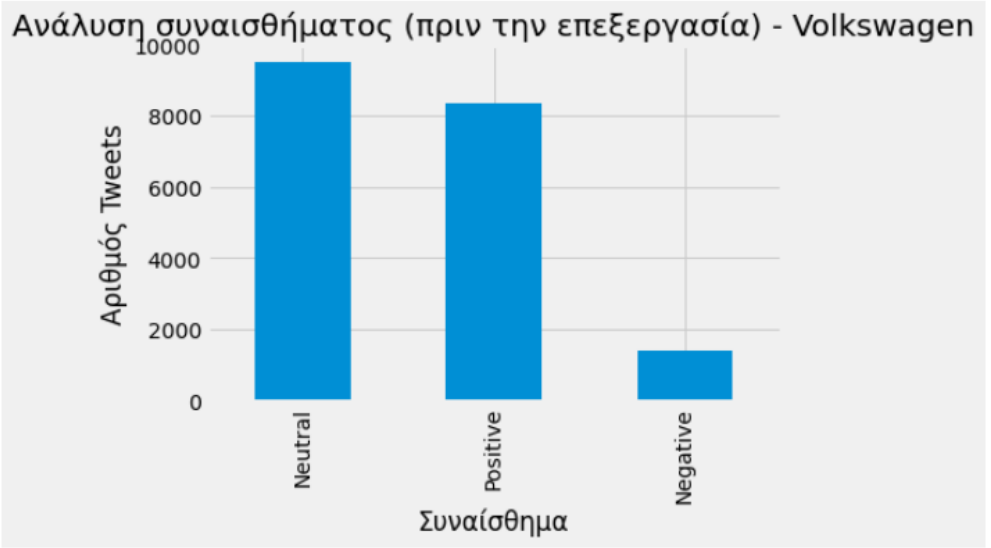
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων αρχικών tweets

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

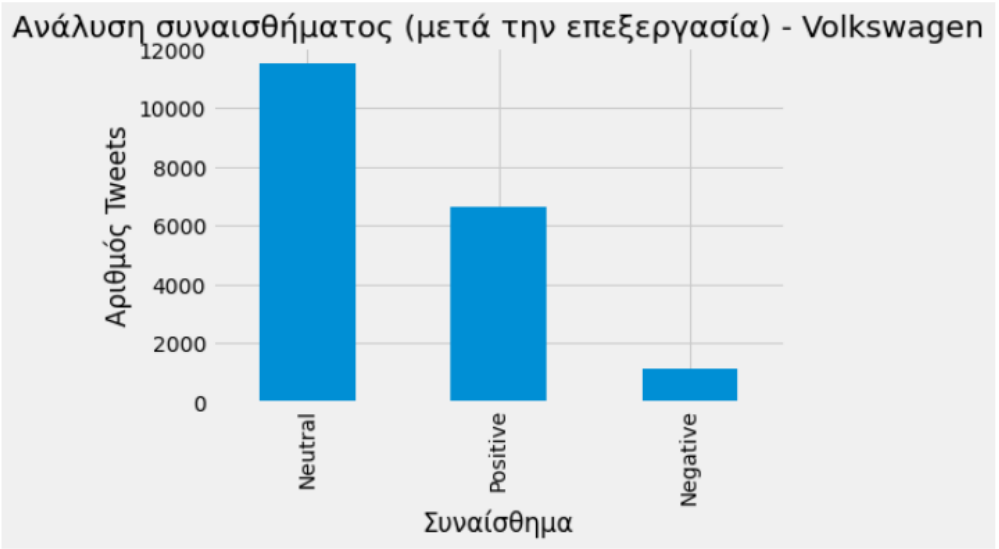
```
Neutral     9525
Positive    8355
Negative    1392
Name: Analysis_Tweet, dtype: int64
```

```
# 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
```

Ανάλυση Συναισθήματος – Ραβδογράμματα πλήθους θετικών, αρνητικών και ουδέτερων Tweets – Αυτοκινητοβιομηχανία Volkswagen (με στελέχωση)

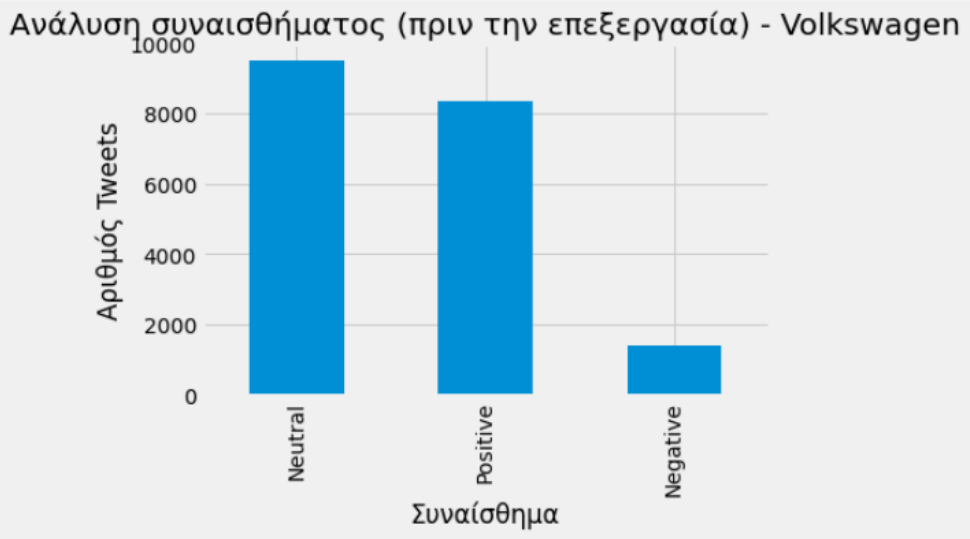


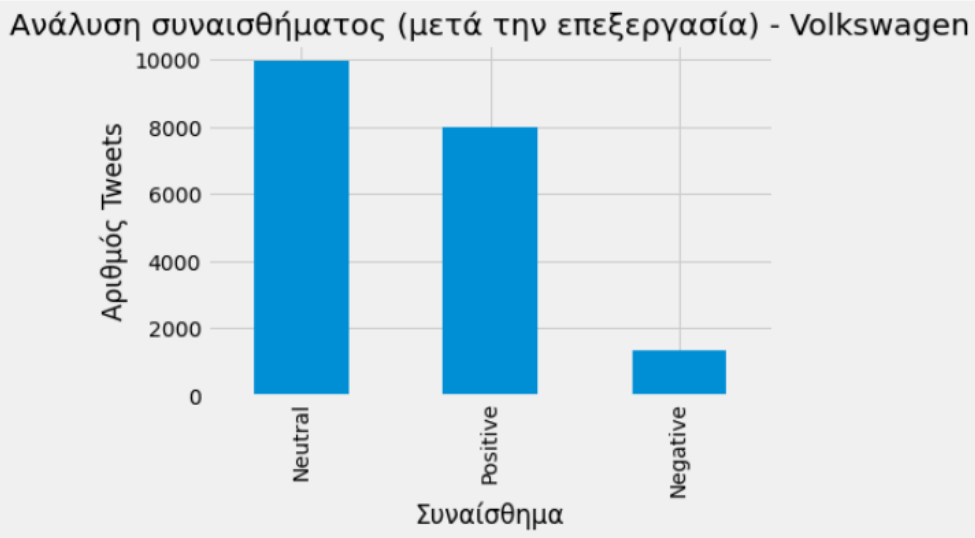


Πίνακας ποσοστών

	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

Ανάλυση Συναισθήματος – Volkswagen (χωρίς στελέχωση)






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

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

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

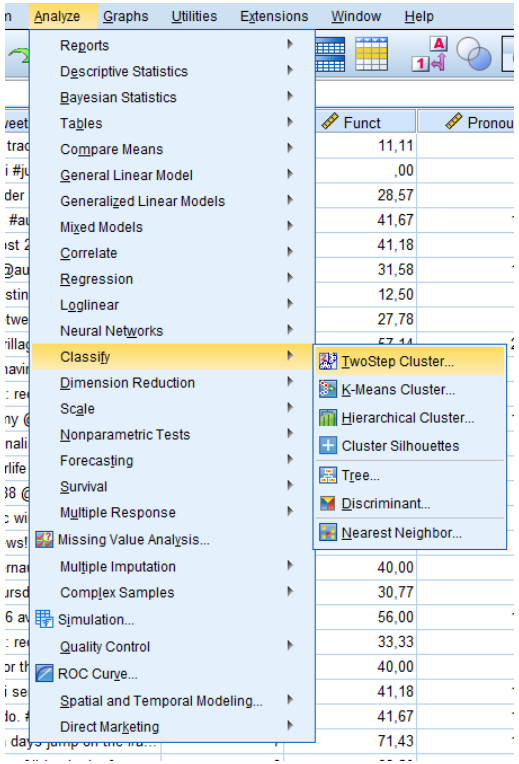
```
# Υπολογισμός θετικών, αρνητικών και ουδέτερων επεξεργασμένων tweets

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

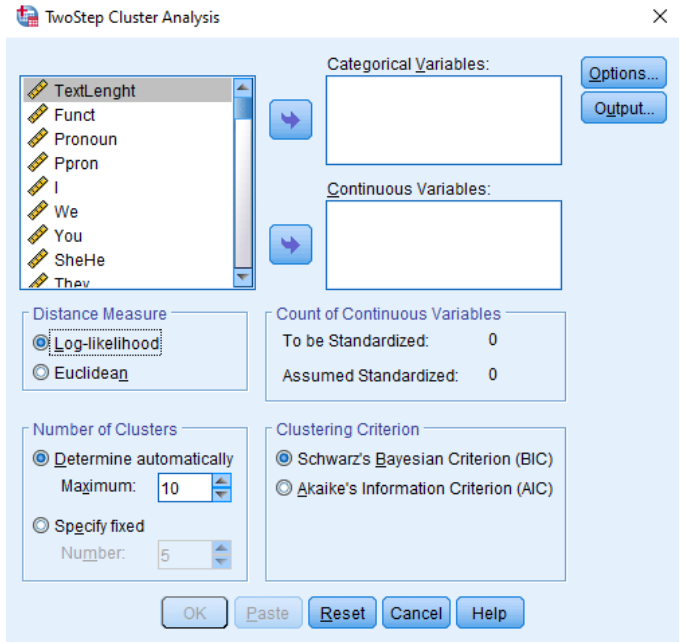
```
Neutral     9973
Positive    7975
Negative    1324
Name: Analysis_Tweet_RFSA, dtype: int64
```

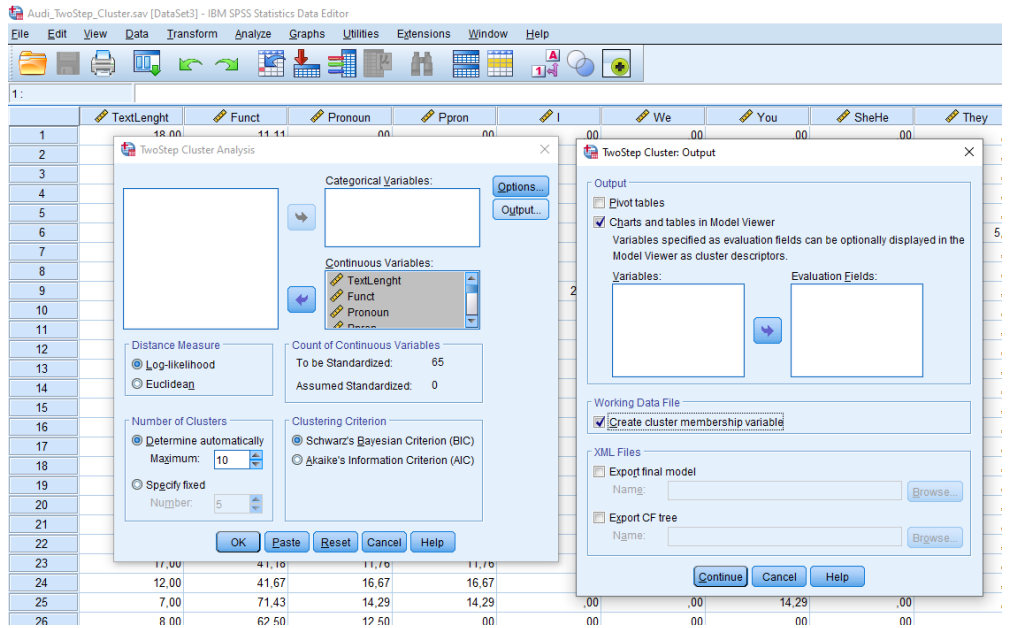
	Ποσοστό Θετικών Tweets	Ποσοστό Ουδέτερων Tweets	Ποσοστό Αρνητικών Tweets	Σύνολο
Πολικότητα (πριν την επεξεργασία)	43.4	49.4	7.2	100.0
Πολικότητα (μετά την επεξεργασία)	41.4	51.7	6.9	100.0

Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Βήματα Συσταδοποίησης Δύο Βημάτων

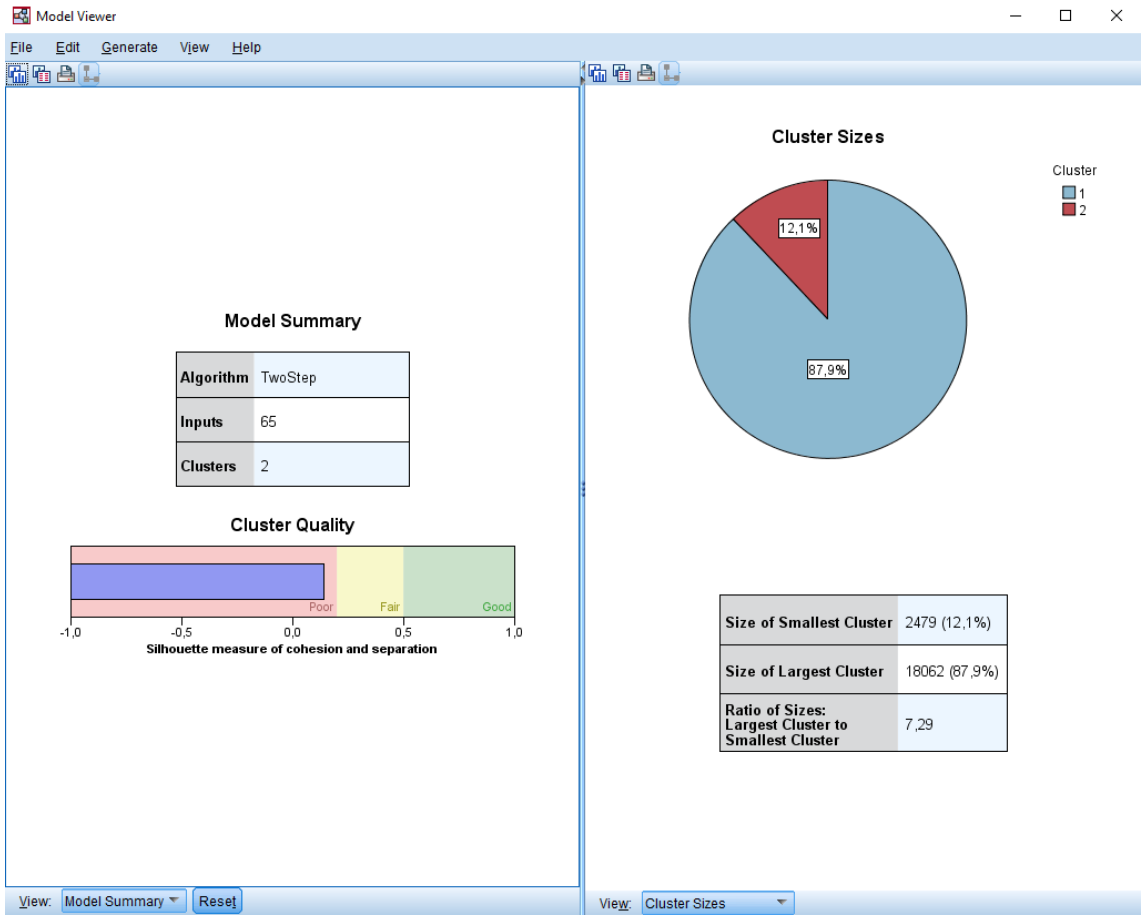


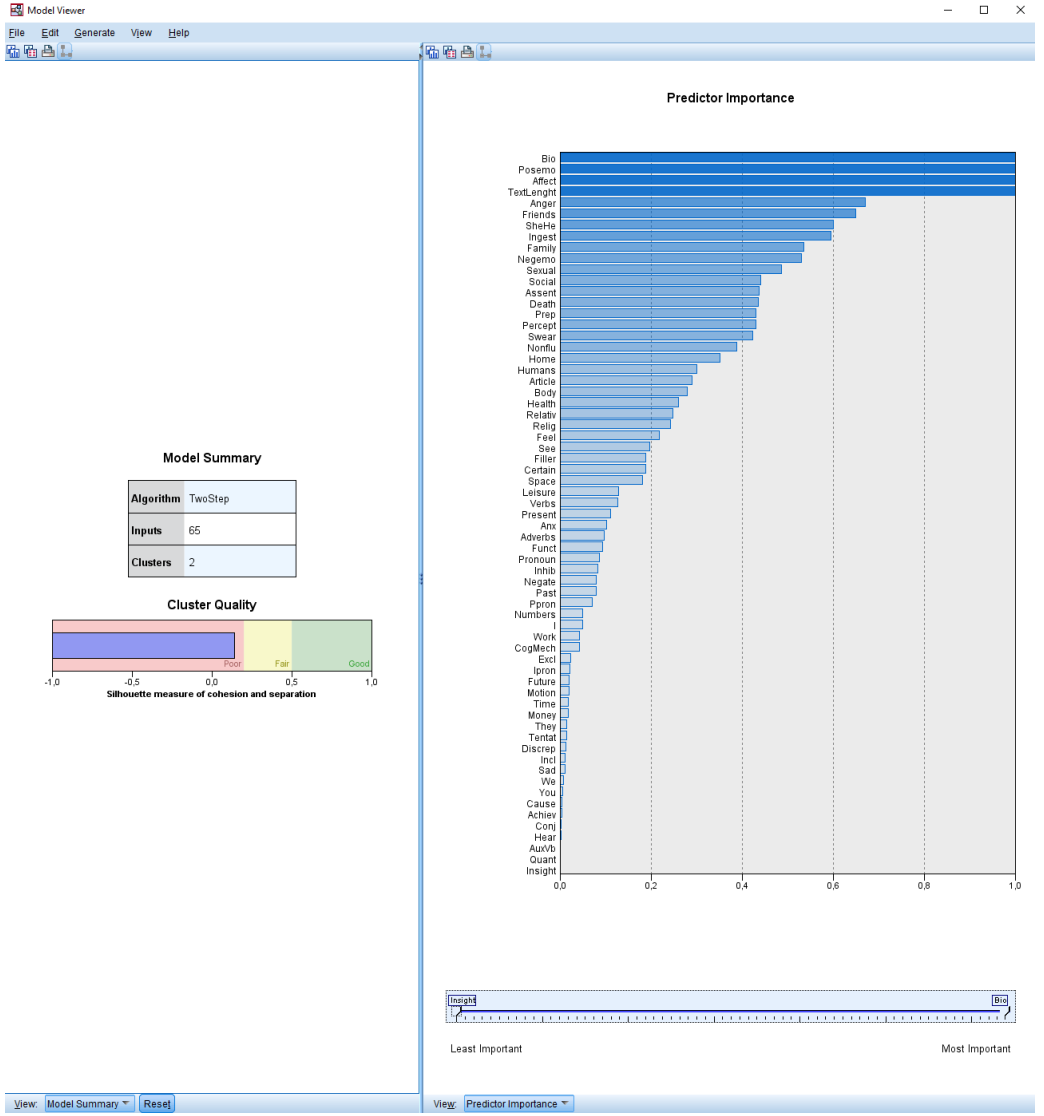




Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Βήματα Συσταδοποίησης Δύο Βημάτων





Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi

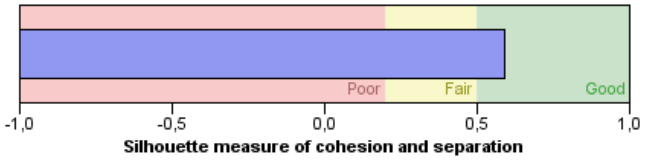
```
TWOSTEP CLUSTER
  /CONTINUOUS VARIABLES=Social Humans Negemo Anger Bio Health Sexual Home
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 10 BIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES
  /SAVE VARIABLE=TSC_3133.
```

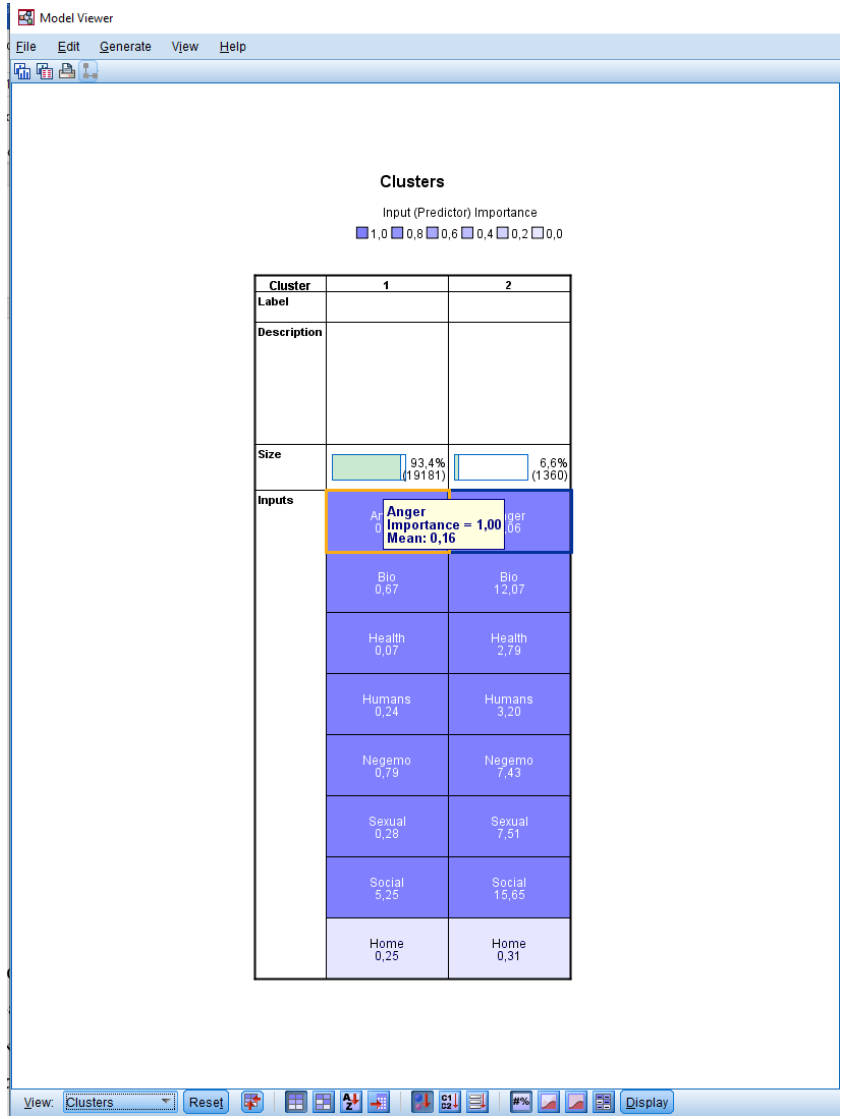
TwoStep Cluster

Model Summary

Algorithm	TwoStep
Inputs	8
Clusters	2

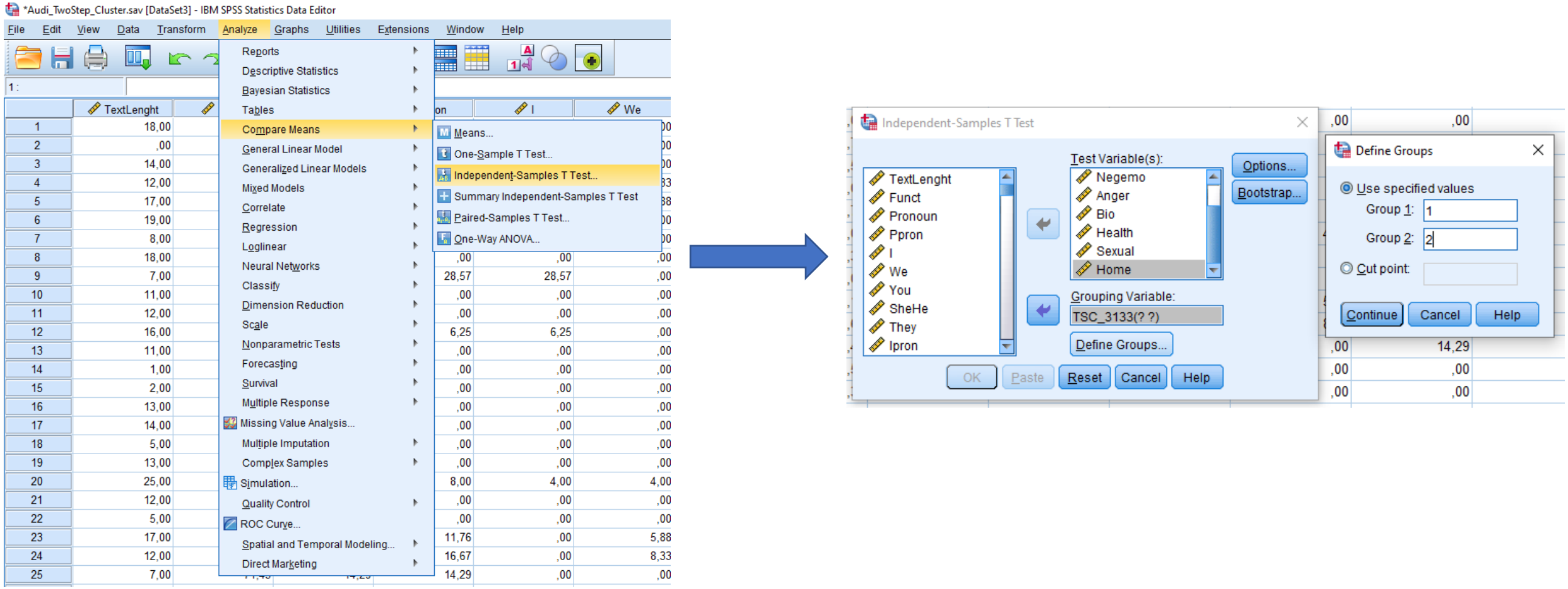
Cluster Quality





TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi

Μη συσχετισμένος έλεγχο t-test



TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi
Μη συσχετισμένος έλεγχο t-test

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

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi

Μη συσχετισμένος έλεγχο t-test

- ✓ Έλεγχος για την ισότητα ή μη των διακυμάνσεων των δυο συστάδων γίνεται αυτόματα μαζί με την υλοποίηση μη συσχετισμένου ελέγχου t-test, με χρήση του test Levene.
- ✓ Η μηδενική και εναλλακτική υπόθεση του test Levene έχουν ως εξής:
- ➢ Μηδενική Υπόθεση H0 : Οι διακυμάνσεις των δύο ομάδων είναι ίσες.
- ➢ Εναλλακτική Υπόθεση H1: Οι διακυμάνσεις των δύο ομάδων δεν είναι ίσες.
- ✓ Levene Test για όλες τις μεταβλητές→
  - → p-value = 0 < 0,05 → η μηδενική υπόθεση για την ισότητα των διακυμάνσεων απορρίπτεται → 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.

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi
Μη συσχετισμένος έλεγχο t-test

- ✓ Για την μεταβλητή **Home** → p-value = 0,490, η μηδενική υπόθεση γίνεται αποδεκτή σε επίπεδο σημαντικότητας 0,05 → **δεν υπάρχει στατιστικά σημαντική διαφορά σε επίπεδο σημαντικότητας 0,05** → Θα μπορούσαμε να την αφαιρέσουμε από το μοντέλο μας, τρέχοντας εκ νέου τον αλγόριθμο.
- ✓ Η τιμή του στατιστικού t είναι -0.690 → 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.

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi (μετά την αφαίρεση των μεταβλητών)

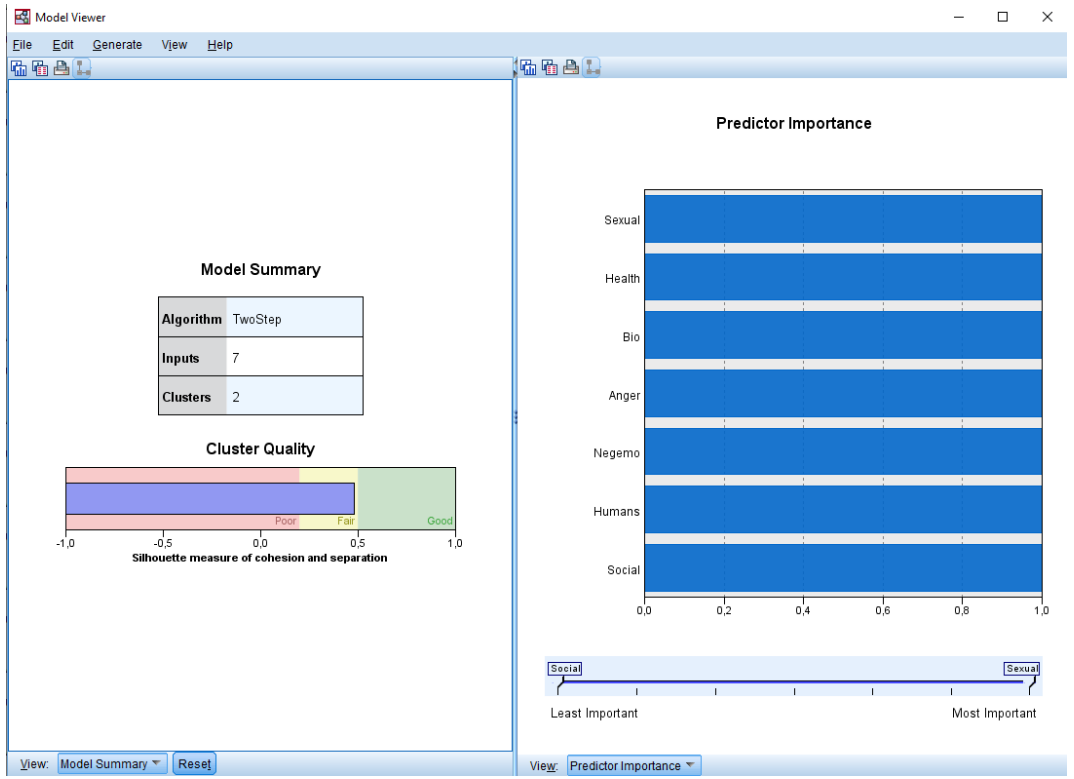
```
TWOSTEP CLUSTER
  /CONTINUOUS VARIABLES=Social Humans Negemo Anger Bio Health Sexual
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 10 BIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES
  /SAVE VARIABLE=TSC_7083.
```

TwoStep Cluster

Model Summary

Algorithm	TwoStep
Inputs	7
Clusters	2

Cluster Quality



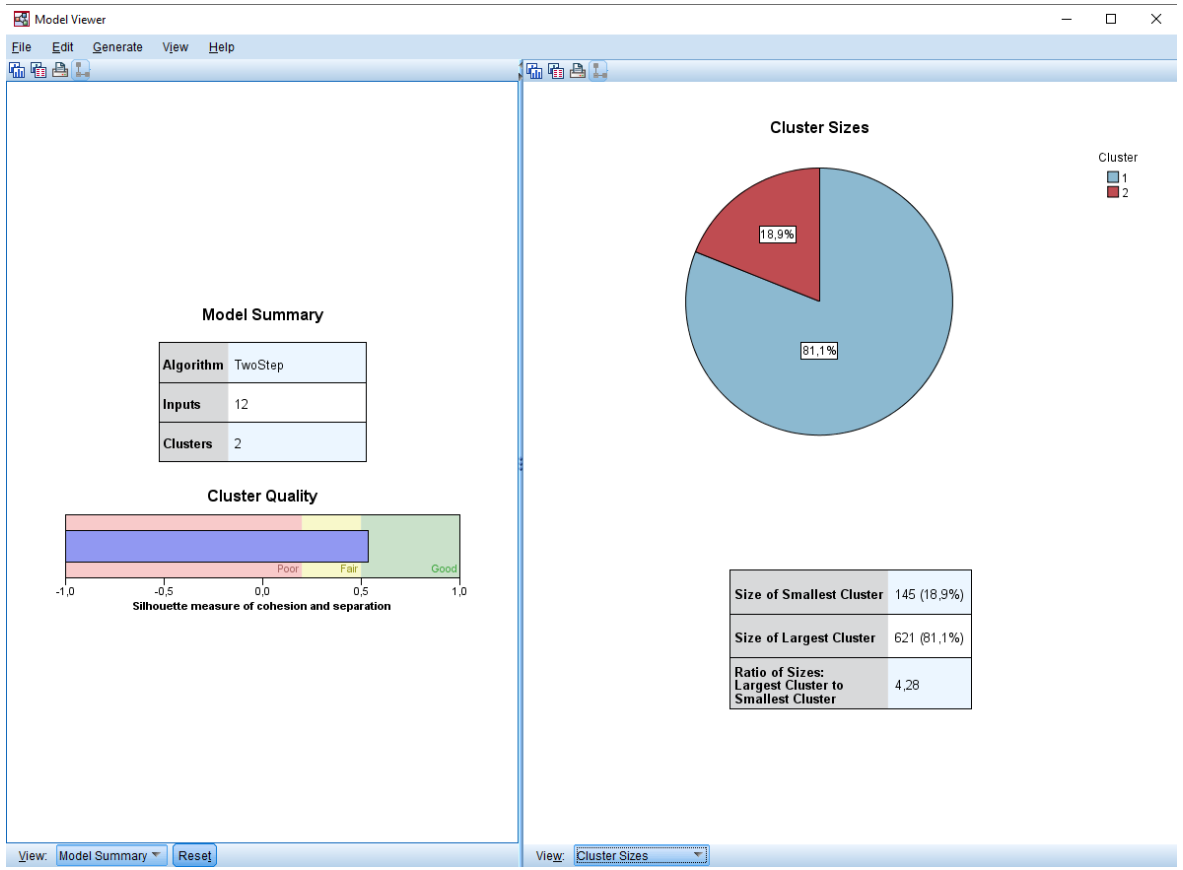
TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Audi

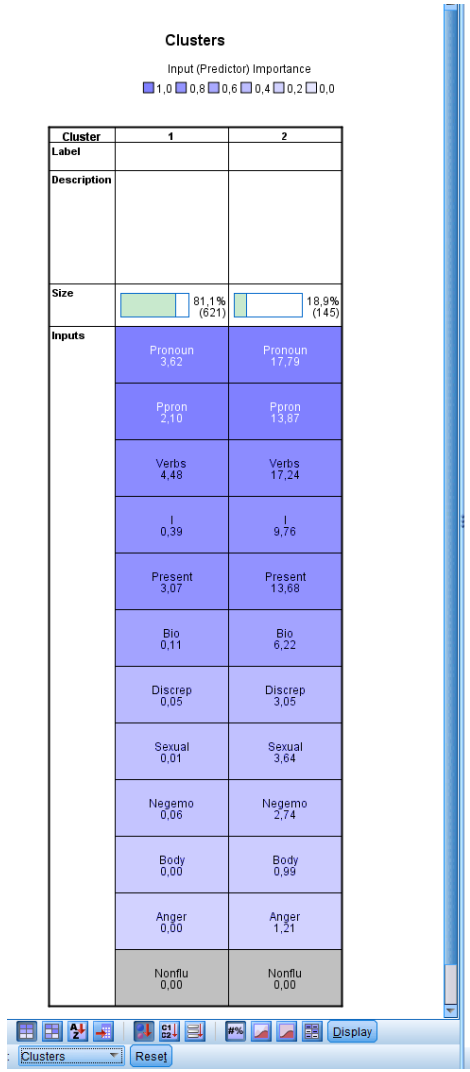
Ερμηνεία

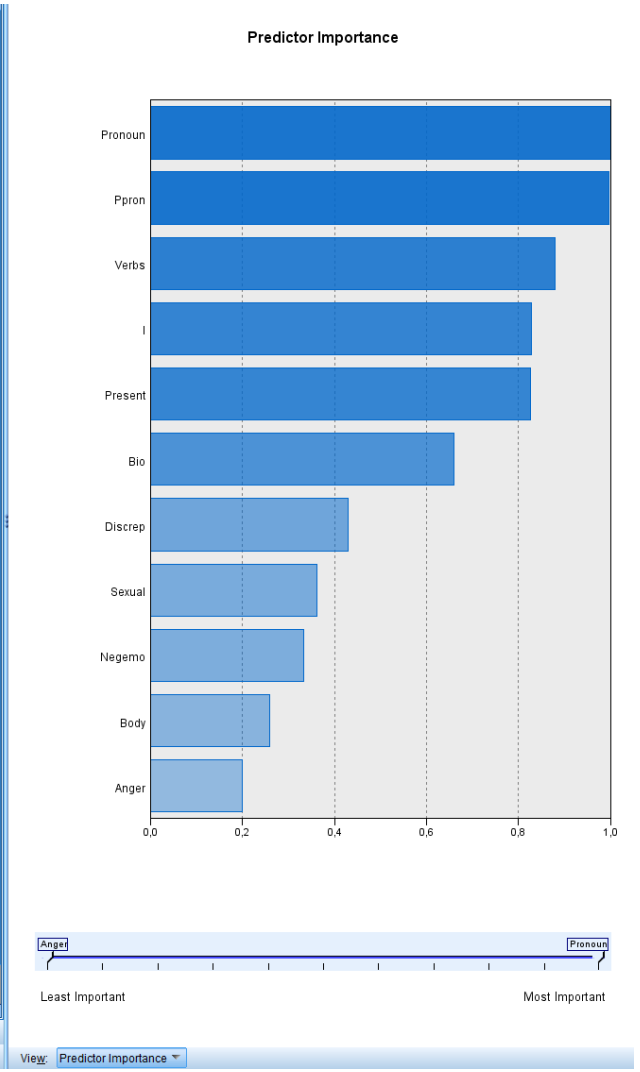
- 8 μεταβλητές συμμετείχαν στην δημιουργία των συστάδων: **Sexual, Health, Bio, Anger, Negemo, Humans, Social** και **Home.**
- Συμμετέχοντες (joiners) ή συνομιλητές σύμφωνα με την κατηγοριοποίηση της Forrester Research, καθώς πρόκειται για χρήστες που συμμετέχουν ενεργά στα κοινωνικά μέσα (μεταβλητή social), εκφράζονται έντονα χρησιμοποιώντας λέξεις που δηλώνουν θυμό (anger) και αρνητικότητα (negemo).
- Σύμφωνα με 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.

Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chevrolet







TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chevrolet

Μη συσχετισμένος έλεγχο t-test

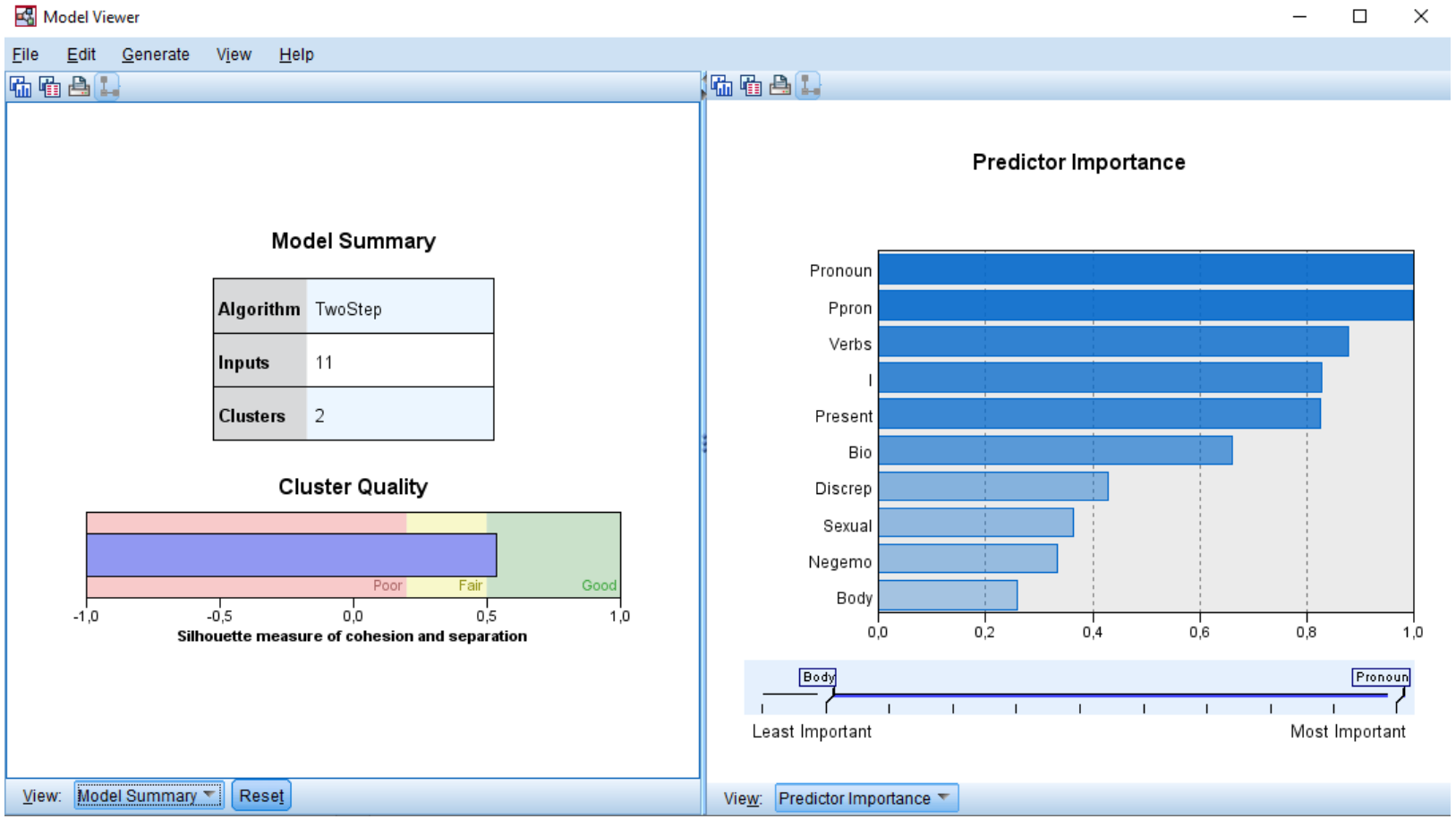
- ✓ Για κάθε μεταβλητή το p-value = 0 <0,05 → Όλες οι ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες → δεν χρειάζεται να αφαιρέσουμε κάποια από τις μεταβλητές Anger, Body, Negemo, Sexual και Discrep.
- ✓ Η μεταβλητή Nonflu και πάλι δεν εμφανίζεται στον πίνακα ελέγχου ανεξαρτησίας και κατά συνέπεια μπορεί να αφαιρεθεί.

T-Test

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chevrolet (μετά την αφαίρεση των μεταβλητών)



TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chevrolet

Ερμηνεία

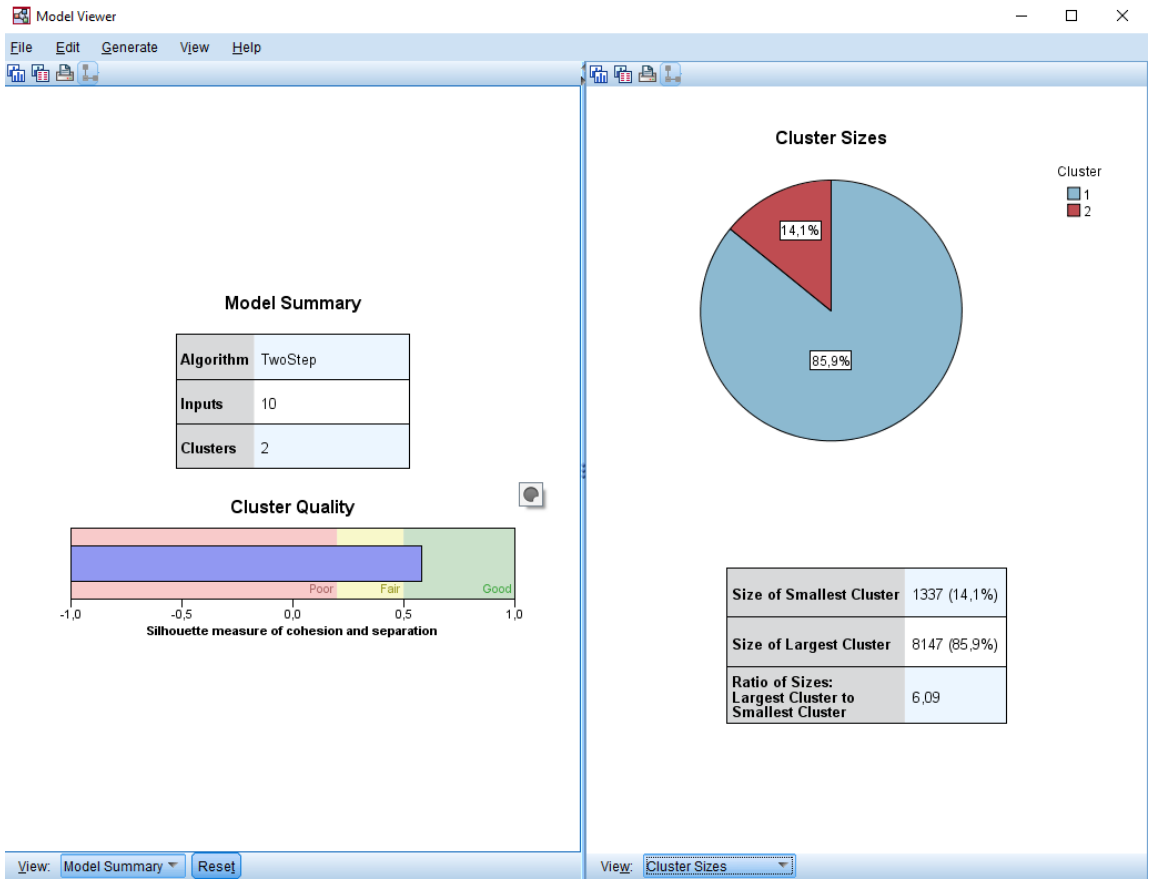
- 11 μεταβλητές συμμετείχαν στην δημιουργία των συστάδων: **Pronoun, Ppron, I, Verbs, Present, Negemo, Anger, Discrep, Bio, Body** και **Sexual**.

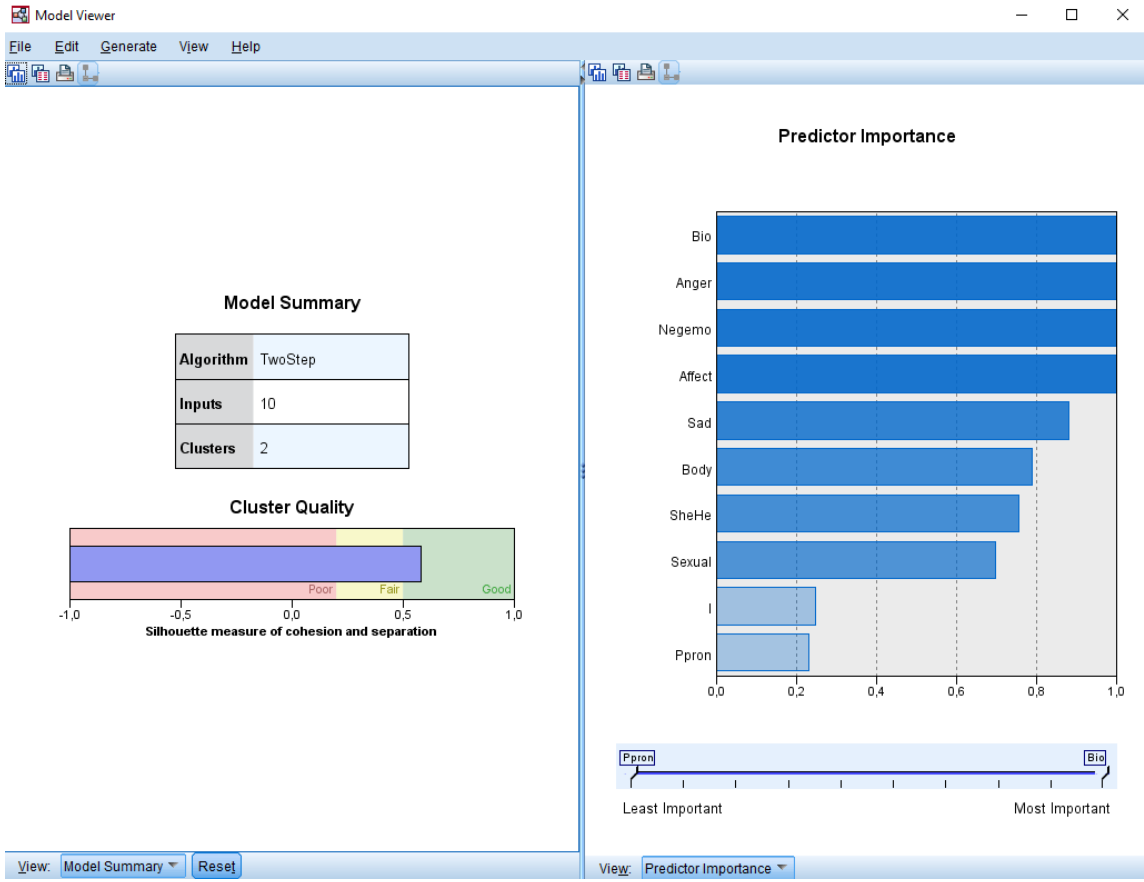
- Συνομιλητές (conversationalists) σύμφωνα με την κατηγοριοποίηση της Forrester Research, καθώς πρόκειται για χρήστες που ζουν στο παρόν, χρησιμοποιούν σε μεγάλο βαθμό τις προσωπικές αντωνυμίες εκφράζοντας και υπερασπίζοντας την άποψή τους, ενώ τα αρνητικά συναισθήματα που φαίνεται να έχουν δεν είναι 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.

- 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.

Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chrysler





TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chrysler
Μη συσχετισμένος έλεγχο t-test

- ✓ Για τις 10 μεταβλητές του τελικού μοντέλου (πριν την αφαίρεση της μεταβλητής Ppron 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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Chrysler Ερμηνεία

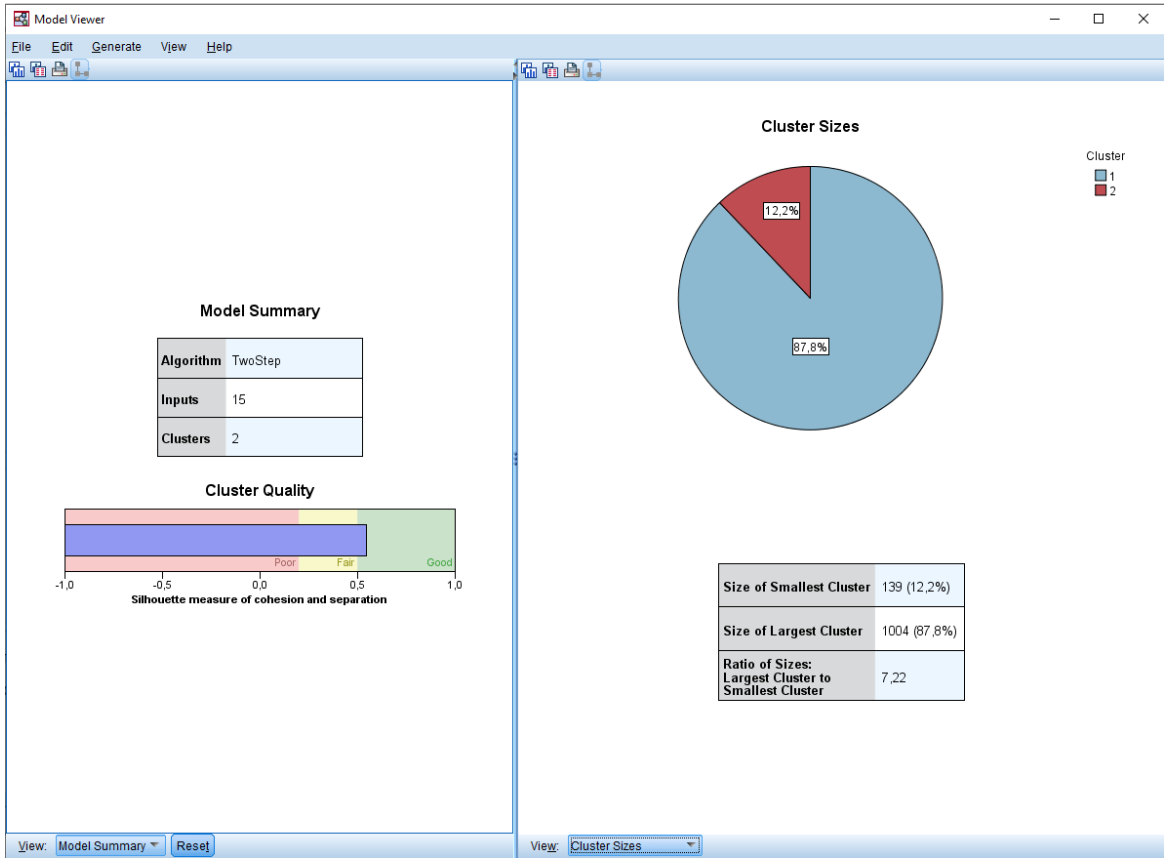
- 10 μεταβλητές συμμετείχαν στην δημιουργία των συστάδων: **I, Affect, Negemo, Bio, Anger, Sad, Body, SheHe, Sexual, Ppron**.

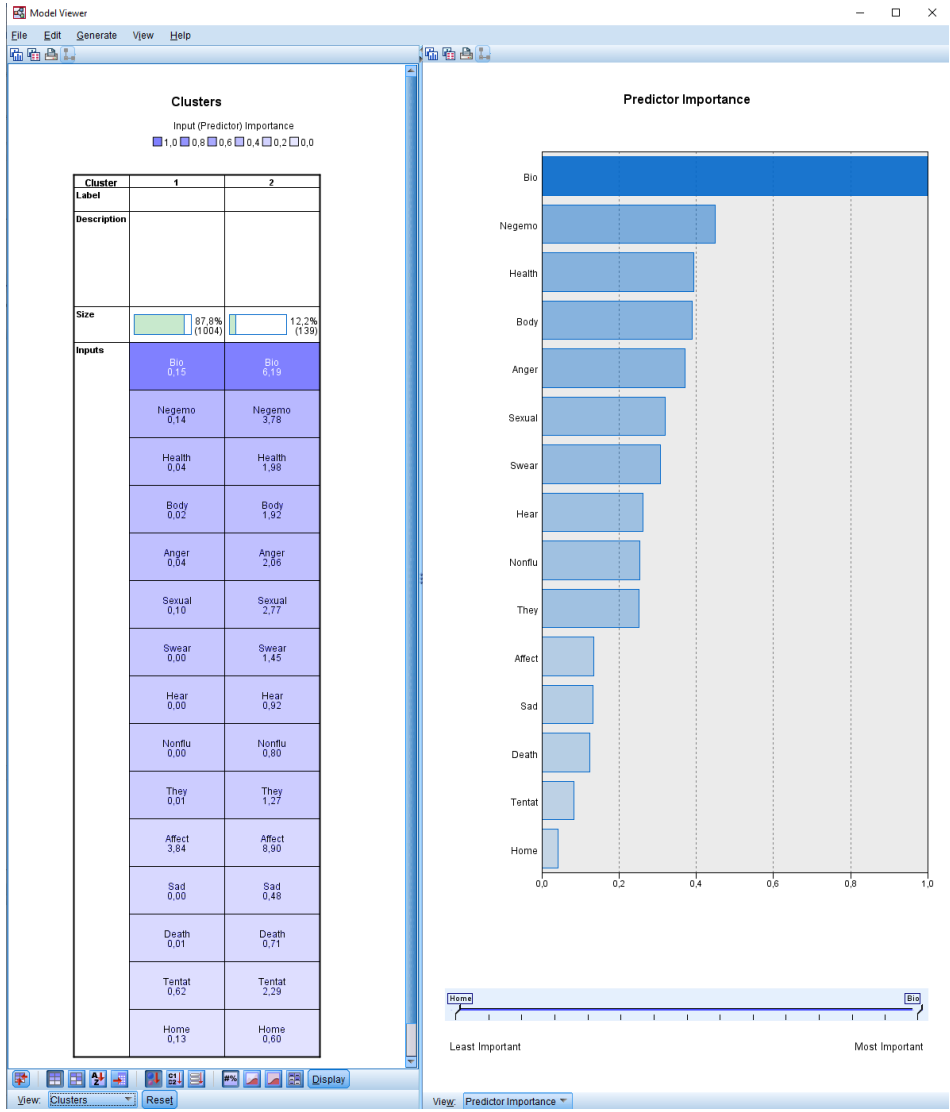
- Κριτικοί (critics) σύμφωνα με την κατηγοριοποίηση της Forrester Research. Οι κριτικοί κατά την αντίδρασή τους σε περιεχόμενο άλλων χρηστών κάνουν συχνή χρήση προσωπικών αντωνυμιών και λέξεων που εκφράζουν συναισθήματα, είτε αυτά είναι θετικά, είτε αρνητικά (negemo, sad, affect).

- 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.

Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας ΚΙΑ





TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας KIA

Μη συσχετισμένος έλεγχο t-test

- ✓ Ραρά 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.
- ✓ Μη συσχετισμένο έλεγχο t-test για τις 15 μεταβλητές του τελικού μοντέλου.
- ✓ 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 → διαφέρουν ως προς τις δύο συστάδες → δεν χρειάζεται να τις αφαιρέσουμε από το μοντέλο.
- ✓ Μεταβλητή Home → p-value = 0,062 > 0,05 → υπάρχει στατιστικά σημαντική διαφορά → αφαίρεση από το μοντέλο

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας KIA

Μη συσχετισμένος έλεγχο t-test (μετά την αφαίρεση της μεταβλητής Home)

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας KIA

Ερμηνεία

- 14 μεταβλητές συμμετείχαν στην δημιουργία των συστάδων: **They, Swear, Affect, Negemo, Anger, Sad, Tentat, Hear, Bio, Body, Health, Sexual, Death** και **Nonflu**.

- Κριτικοί ή θεατές (spectators) σύμφωνα με την κατηγοριοποίηση της Forrester Research.

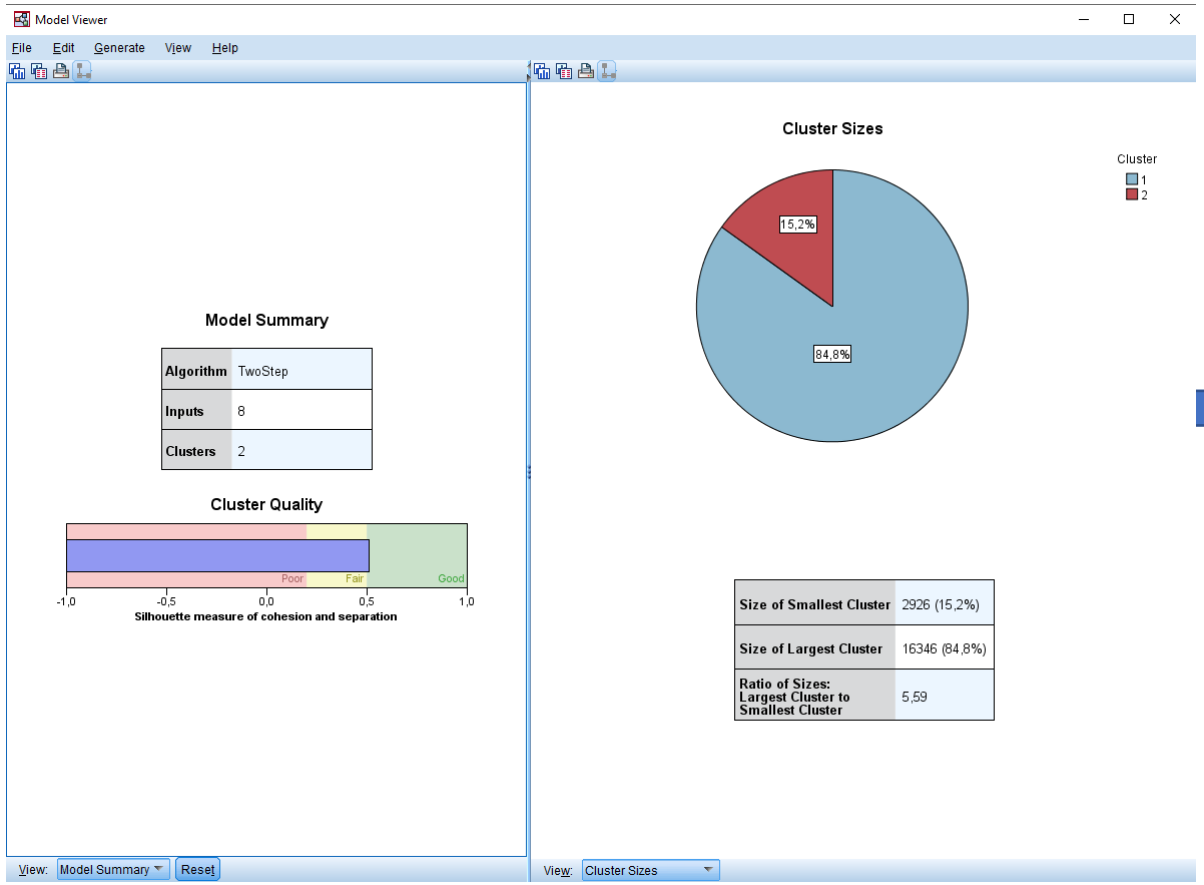
- Οι κριτικοί κατά τον σχολιασμό του χρησιμοποιούν υβριστικές λέξεις (που μπορεί να περιέχουν και σεξουαλικό περιεχόμενο) και εκφράζονται συχνά με θυμό, που αντανακλά ίσως και την ψυχολογική τους κατάσταση.

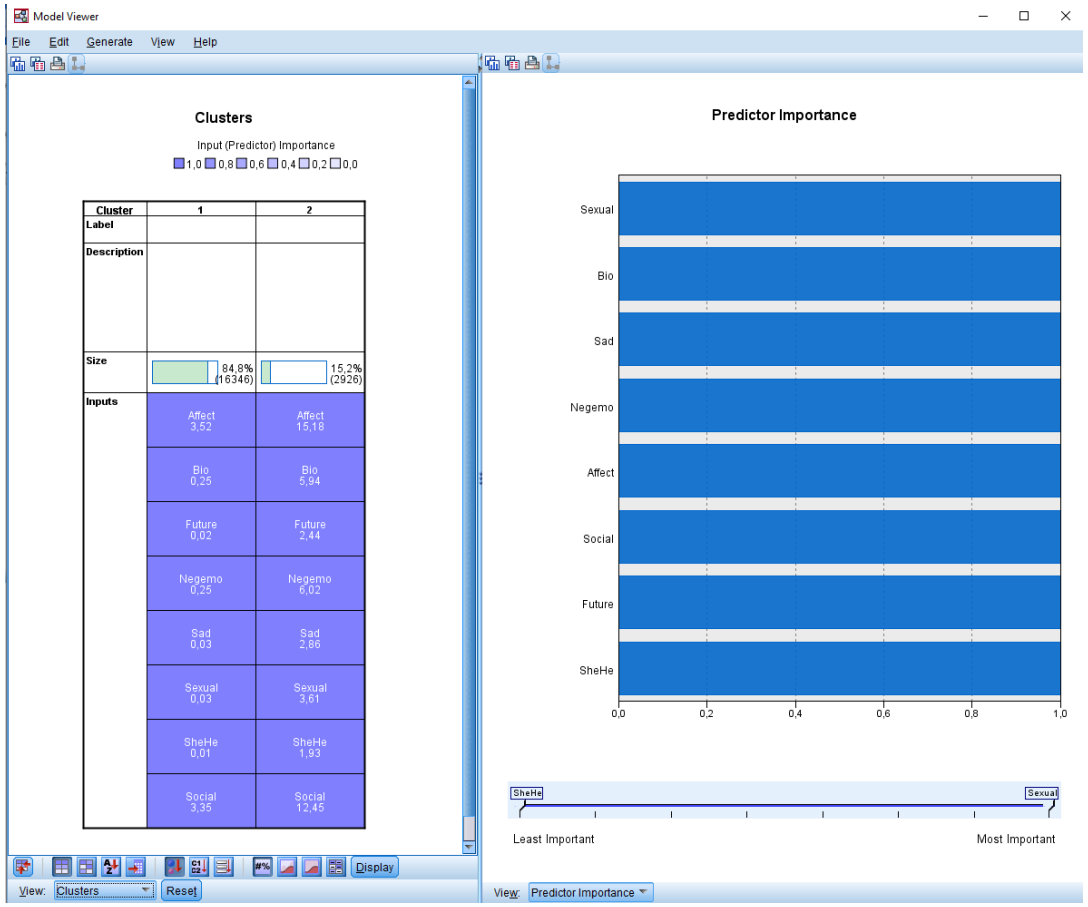
- Οι θεατές είναι άτομα που ακούν, διαβάζουν, αλλά και σχολιάζουν το περιεχόμενο που δημοσιεύουν οι άλλοι χρήστες, χρησιμοποιώντας συχνά εκφράσεις χωρίς βαθυστόχαστες σκέψη που συνοδεύονται συχνά από λέξεις της κατηγορίας Nonflu (προέρχεται από τη λέξη nonfluencies).

- 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.

Εξόρυξη προφίλ χρηστών

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Volkswagen





TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Volkswagen

Μη συσχετισμένος έλεγχο t-test

- ✓ Από τον μη συσχετισμένο έλεγχο t-test επιβεβαιώσαμε ότι όλες οι ποσοτικές μεταβλητές διαφέρουν ως προς τις δύο συστάδες

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

TwoStep Cluster – Αποτελέσματα Αυτοκινητοβιομηχανίας Volkswagen

Ερμηνεία

- 8 μεταβλητές συμμετείχαν στην δημιουργία των συστάδων: **SheHe, Future, Social, Affect, Negemo, Sad, Bio και Sexual**.

- 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, σύμφωνα με την Forrester Research.

- 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.

Συμπεράσματα – Σχόλια – Προτάσεις

❖ Ανάλυση συναισθήματος

- Audi, Chevrolet → Μεταστροφή του ποσοστού των θετικών και ουδέτερων tweets, μετά την διαδικασία της στελέχωσης

  ➢ Αιτία: 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

- Χωρίς στελέχωση

  ✓ Audi, KIA, Chevrolet → Υψηλό ποσοστό θετικών tweets → Επιβεβαίωση διαφημιστικής στρατηγικής

  ✓ Chrysler, Volkswagen → Υψηλό ποσοστό ουδέτερων tweets → Ένδειξη στασιμότητας

  →Ανάγκη εύρεσης τρόπων βελτίωσης της εικόνας τους, προσέλκυσης νέων πελατών και ενίσχυσης του αισθήματος δέσμευσης των ήδη υπαρχόντων πελατών

Συμπεράσματα – Σχόλια – Προτάσεις

❖ Εξόρυξη προφίλ χρηστών

- Σύμφωνα με την μελέτη των Tausczik και Pennebaker και την εργασία των Schwarts et al. φαίνεται να υπάρχει συσχέτιση μεταξύ των κατηγοριών των λέξεων του λεξικού LIWC με το φύλο, την ηλικία και το μοντέλο OCEAN
- Εφαρμογή αλγορίθμου συσταδοποίησης δύο βημάτων στα tweets των αυτοκινητοβιομηχανιών: Audi, Chevrolet, Chrysler, KIA και Volkswagen

  Κατηγοριοποίηση χρηστών βάση του μοντέλου OCEAN:

  ✓ Chevrolet, Chrysler, KIA → Tweets με εμφάνιση αρνητικών συναισθημάτων →
  → Νευρωτικές προσωπικότητες

  ✓ Audi, Volkswagen → Tweets με εμφάνιση κοινωνικών διεργασιών και αρνητικών συναισθημάτων → Εξωστρεφείς ή Νευρωτικές προσωπικότητες

  ➢ Δυσκολία διάκρισης μεταξύ των εξωστρεφών και νευρωτικών → Έλλειψη δυνατότητας ανίχνευσης της γλώσσας των κοινωνικών δικτύων (netspeak) από το λεξικό LIWC2007.

Συμπεράσματα – Σχόλια – Προτάσεις

- Εξόρυξη προφίλ χρηστών
  - Γνώση του τύπου προσωπικότητας μέσω των tweets των χρηστών →

→ Καθορισμός των παραμέτρων, του είδους και το ύφους των μελλοντικών διαφημίσεων

και

→ Των μηνυμάτων που αυτές θέλουν θα περάσουν στους καταναλωτές τους, ώστε να ανταποκριθούν με τον καλύτερο δυνατό τρόπο στις απαιτήσεις και τις προσδοκίες τους

Ευχαριστώ για την προσοχή σας!