ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΤΟ ΣΥΝΑΙΣΘΗΜΑ ΑΠΟ ΔΕΔΟΜΕΝΑ ΤΟΥ TWITTER

Διπλωματική Εργασία

της

Μαρία Γεωργιάδου

Θεσσαλονίκη, Φεβρουάριος 2021

ΕΞΟΡΥΞΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ ΠΟΥ ΒΑΣΙΖΟΝΤΑΙ ΣΤΟ ΣΥΝΑΙΣΘΗΜΑ ΑΠΟ ΔΕΔΟΜΕΝΑ ΤΟΥ TWITTER

Μαρία Γεωργιάδου

Πτυχίο Μαθηματικών, ΑΠΘ, 2019

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπουσα Καθηγήτρια
Γεωργία Κολωνιάρη

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26/02/2021

Ονοματεπώνυμο 1	Ονοματεπώνυμο 2	Ονοματεπώνυμο 3
Γεωργία Κολωνιάρη	Γεώργιος Ευαγγελίδης	Κωνσταντίνος Ψάννης

Μαρία Γεωργιάδου

Περίληψη

Η ανάγνωση και η γραφή tweets μπορεί να επηρεάσει τα συναισθήματα των ανθρώπων. Μπορούν να βιώσουν ταυτόχρονα συναισθήματα, όπως ευτυχία και ανησυχία, ή θλίψη και θυμό, όταν διαβάζουν μία ανάρτηση στο Twitter. Ωστόσο, η εξαγωγή συναισθημάτων που μπορεί να προκληθούν από ένα κείμενο είναι μια περίπλοκη εργασία. Για να εξάγουμε σχέσεις μεταξύ συναισθημάτων και λέξεων, προτείνεται μια μέθοδο ανάλυσης συναισθημάτων που βασίζεται στην εξόρυξη δεδομένων. Συγκεκριμένα, μέσα από σύνολα δεδομένων που περιέχουν το κείμενο των tweets και την ταξινόμηση τους σε συναισθήματα ή την πολικότητα (θετική/αρνητική) που δημιουργούν, θα αντληθούν κανόνες εξόρυξης με λέξεις αντιστοιχίζοντας στο δεξί μέρος του κανόνα σε συναισθήματα ή την πολικότητα των λέξεων αυτών. Πραγματοποιείται προεπεξεργασία των κειμενικών δεδομένων (διαίρεση των tweets σε όρους, μετατροπή όλων των χαρακτήρων σε πεζούς, αφαίρεση των αριθμών και των συμβόλων, αφαίρεση stopwords, λημματοποίηση, αποκατάληξη και λεξικογραφική ανάλυση), έτσι ώστε να γίνει καλύτερη μέτρηση των συχνών στοιχειοσυνόλων. Ο αλγόριθμος που χρησιμοποιείται είναι η Apriori και η υλοποίηση γίνεται με κώδικα στην γλώσσα προγραμματισμού Python και χρήση του προγράμματος RapidMiner.

Λέξεις κλειδιά : Συναισθηματική ανάλυση, εξόρυξη γνώμης, αλγόριθμος Apriori, Tweets.

Abstract

Reading and writing tweets can affect people's emotions. They may experience emotions such as happiness and anxiety at the same time, or sadness and anger when reading a Twitter post. However, extracting emotions that can be caused by a text is a complex task. To extract relationships between emotions and words, a method of sentiment analysis based on data mining is proposed. Specifically, through data sets that contain the text of tweets and their classification into emotions or the polarity (positive / negative) which they create, word mining rules will be created by matching the right part of the rule to emotions or the polarity of those words. Text data are pre-processed (tokenization, converting all characters to lowercase, subtracting numbers and symbols, removing stopwords, lemmatization, stemming, and pos tag) to get better measurements for common itemsets. Apriori algorithm has been used and the implementation has been done with code in the Python programming language and using the program RapidMiner.

Keywords : sentiment analysis, opinion mining, Apriori algorithm, Tweets.

Περιεχόμενα

1. Εισαγωγή 8
1.1 Εξόρυξη γνώσης από δεδομένα 8
1.2 Στόχος έρευνας 13
1.3 Συνεισφορά της έρευνας 13
1.4 Δομή της έρευνας 14
2. Βιβλιογραφική Επισκόπηση 15
2.1 Μηχανική Μάθηση 15
2.2 Λεξικά 20
2.3 Μέτρα- Βάρη 22
2.4 Κατάταξη- Δημοτικότητα 23
2.5 Άλλες Υλοποιήσεις 25
3. Μεθοδολογία 27
3.1 Αλγόριθμοι 27
3.2 Υλοποίηση 33
4. Αποτελέσματα 44
5. Συμπεράσματα 58
5.1 Σύνοψη 58
5.2 Μελλοντικέ επεκτάσεις 58
6. Παραπομπές 59

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 3.1: Προεπεξεργασία κειμένου 39

Εικόνα 3.2 Ανάλυση της υλοποίησης σε RapidMiner 42

Εικόνα 3.3 Προεπεξεργασία δεδομένων στο RapidMiner 43

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 4.1 Πλήθος συχνών στοιχειοσυνόλων των D1-D3 45
με διάφορα ελάχιστα όρια.

Πίνακας 4.2 Πλήθος συχνών στοιχειοσυνόλων των D4-D6 με 45
διάφορα ελάχιστα όρια.

Πίνακας 4.3 Πλήθος κανόνων συσχέτισης με διάφορα ελάχιστα όρια. 46

Πίνακας 4.4 Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D1-D3. 48

Πίνακας 4.5 Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D4-D6. 49

Πίνακας 4.6 Κανόνες εξόρυξης για το αρχείο D1. 50

Πίνακας 4.7 Κανόνες εξόρυξης για το αρχείο D1 με TLSP. 50

Πίνακας 4.8 Κανόνες εξόρυξης για το αρχείο D2. 51

Πίνακας 4.9 Κανόνες εξόρυξης για το αρχείο D2 με TLSP. 51

Πίνακας 4.10 Κανόνες εξόρυξης για το αρχείο D3 52
με TLSP(ADJECTIVE).

Πίνακας 4.11 Κανόνες εξόρυξης για το αρχείο D3 με TLSP(NOUN). 52

Πίνακας 4.12 Κανόνες εξόρυξης για το αρχείο D4 . 53

Πίνακας 4.13 Κανόνες εξόρυξης για το αρχείο D4 με TLSP (NOUN/ADJ). 53

Πίνακας 4.14 Κανόνες εξόρυξης για το αρχείο D5. 54

Πίνακας 4.15 Κανόνες εξόρυξης για το αρχείο D5 με TLSP. 54

Πίνακας 4.16 Κανόνες εξόρυξης για το αρχείο D6 με TLSP (NOUN). 54

Πίνακας 4.17 Κανόνες εξόρυξης για το αρχείο D6 με TLSP(ADJECTIVE/VERB). 55

Πίνακας 4.18 Ισχυρότεροι κανόνες εξόρυξης με stopwords 55
για τα αρχεία D1 – D3.

Πίνακας 4.19 Ισχυρότεροι κανόνες εξόρυξης με stopwords 56
για τα αρχεία D4 – D6.

1.ΕΙΣΑΓΩΓΗ

1.1 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Καθημερινά συλλέγεται και αποθηκεύεται ένας τεράστιος όγκος δεδομένων, τόσο δομημένα όσο και αδόμητα , στο διαδίκτυο ,σε βάσεις δεδομένων ,σε cloud τεχνολογίες καθώς και σε χειρόγραφα έντυπα. Οι μεγάλες ποσότητες δεδομένων περιέχουν χρήσιμες πληροφορίες , οι οποίες με τα χρόνια παρέμεναν ανεκμετάλλευτες.

Η επιρροή της τεχνολογίας σε όλο τον κόσμο αυξάνεται ραγδαία. Σε αυτήν τη νέα εποχή όπου όλα μεταφέρονται στην ηλεκτρονική πραγματικότητα, τα δεδομένα έχουν αναχθεί ως το νέο νόμισμα. Έχει γίνει ένα από τα πιο σημαντικά πλεονεκτήματα κάθε οργανισμού, επιχείρησης αλλα και σε κρατικό επίπεδο . Ως εκ τούτου, οι χρήσεις της εξόρυξης δεδομένων, ιδιαίτερα προερχόμενα από το διαδίκτυο, στο σημερινό κόσμο δεν μπορούν να υπονομευθούν ή να υποτιμηθούν.

Βασικά, ο κύριος σκοπός της χρήσης της εξόρυξης δεδομένων είναι ο χειρισμός τεράστιου όγκου δεδομένων είτε ύπαρξης είτε αποθήκευσης στις βάσεις δεδομένων καθορίζοντας κατάλληλες μεταβλητές που συμβάλλουν στην ποιότητα της πρόβλεψης που θα χρησιμοποιηθεί για την επίλυση του προβλήματος. Ορισμός από τους Gargano & Raggad, 1999. "Οι αναζητήσεις εξόρυξης δεδομένων για κρυφές σχέσεις, μοτίβα, συσχετίσεις και αλληλεξαρτήσεις σε μεγάλες βάσεις δεδομένων που μπορεί να παραβλέψουν οι παραδοσιακές μέθοδοι συλλογής πληροφοριών".

Εκτός αυτού, ένας άλλος συγγραφέας συμφώνησε επίσης με τη γνώμη σχετικά με τον ορισμό της εξόρυξης δεδομένων. Σύμφωνα με τον Palace, (1996) προστέθηκε στο προηγούμενο :«Η εξόρυξη δεδομένων είναι η διαδικασία εύρεσης συσχετίσεων ή μοτίβων ανάμεσα σε δεκάδες πεδία σε μεγάλες σχεσιακές βάσεις δεδομένων».

Επιπλέον, η εξόρυξη δεδομένων ορίζεται και ως μια διαδικασία συμπίεσης γνώσεων ή πληροφοριών χρησιμοποιώντας ένα κατάλληλο πλαίσιο ή ένα μοντέλο για την ανάλυση μέχρι να παραχθεί ένα αποτέλεσμα που βοηθά στην πραγματοποίηση του στόχου της μελέτης. Ο Imberman (2001) δήλωσε ότι η εξόρυξη γνώσης αποτελεί : «Ως εξαγωγή γνώσεων, ανακάλυψη πληροφοριών, συλλογή πληροφοριών, διερευνητική ανάλυση δεδομένων, αρχαιολογία δεδομένων, επεξεργασία προτύπων δεδομένων και ανάλυση λειτουργικής εξάρτησης».

Η παραπάνω δήλωση συμφωνεί και προσθέτει το πλαίσιο ή το μοντέλο που υιοθετεί σίγουρα για να αποκαλύψει την πραγματική περίσταση. Ορισμός από τους Ma et al (2000):«Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής τεχνικών τεχνητής νοημοσύνης (όπως προηγμένη μοντελοποίηση και διέγερση κανόνων) σε ένα μεγάλο σύνολο δεδομένων για τον προσδιορισμό των προτύπων στα δεδομένα».

Από την άλλη πλευρά, η εξόρυξη δεδομένων γίνεται μερικά βήματα κατά τη διάρκεια της ανάλυσης και αυτό το βήμα εξαρτάται από τη μεθοδολογία που επιλέγεται. Κάθε μεθοδολογία δεν διαφέρει πολύ από την άλλη μεθοδολογία. Μέσω των Forcht & Cochran(1999):«Η εξόρυξη δεδομένων είναι μια δια δραστική διαδικασία που περιλαμβάνει τη συγκέντρωση των δεδομένων σε μορφή ευνοϊκή για την ανάλυση. Μόλις διαμορφωθούν τα δεδομένα, πρέπει να καθαριστούν ελέγχοντας για προφανή σφάλματα ή ελαττώματα (όπως ένα στοιχείο που είναι εξαιρετικά ακραίο) και απλώς αφαιρώντας τα ».

Όπως συζητείται παραπάνω, είναι ευδιάκριτο ότι η εξόρυξη γνώσης είναι χρήσιμη και επωφελής σε πολλά κομμάτια, καθώς το μοντέλο που υλοποιείται μειώνει το κόστος και τον χρόνο. Στην συνέχεια, μετατρέπονται σε στρατηγικό πλεονέκτημα της κάθε επιχείρησης, αναλύοντας κριτικά τα αποτελέσματα που προέκυψαν.

Σύμφωνα με τους (Goebel & Gruenwald, 1999) για την αναζήτηση του μοτίβου των δεδομένων, χρησιμοποιούνται μεθοδολογίες για την αποσαφήνιση καθώς και για τον εντοπισμό της σχέσης μεταξύ μιας μεταβλητής και των υπόλοιπων μέσα στην βάση των δεδομένων, ενώ το αποτέλεσμα θα οδηγήσει στη δημιουργία μιας απόφασης ή να προβλέψει τον αντίκτυπο της δράσης. Η επιλογή των μεθοδολογιών θα πρέπει να προσδιορίζεται με τον κατάλληλο τρόπο που να ταιριάζει με τους κανόνες και τις προϋποθέσεις για τα δεδομένα που πρόκειται να αναλυθούν. Οι μεθοδολογίες περιέχουν:

- Στατιστικές Μέθοδοι: επικεντρώνονται στη δοκιμή προκαταλήψεων υποθέσεων και στην προσαρμογή μοντέλων σε δεδομένα.
- Λόγος περίπτωσης (Case-Based Reasoning): τεχνολογία που προσπαθεί να λύσει ένα δεδομένο πρόβλημα κάνοντας άμεση χρήση προηγούμενων εμπειριών και λύσεων

- Νευρωνικά δίκτυα: σχηματισμένα από μεγάλο αριθμό προσομοιωμένων νευρώνων, που συνδέονται μεταξύ τους με τρόπο παρόμοιο με τους εγκεφαλικούς νευρώνες που επιτρέπει στο δίκτυο να «μάθει».
- Δέντρα απόφασης: κάθε μη τερματικός κόμβος αντιπροσωπεύει μια δοκιμή ή απόφαση για το εξεταζόμενο στοιχείο δεδομένων και μπορεί επίσης να ερμηνευθεί ως μια ειδική μορφή ενός συνόλου κανόνων, που χαρακτηρίζεται από την ιεραρχική οργάνωση κανόνων.
- Επαγωγή κανόνα: Οι κανόνες δηλώνουν μια στατιστική συσχέτιση μεταξύ των εμφανίσεων ορισμένων χαρακτηριστικών σε ένα στοιχείο δεδομένων ή μεταξύ ορισμένων στοιχείων δεδομένων σε ένα σύνολο δεδομένων.
- Bayesian Belief Networks: γραφικές αναπαραστάσεις των κατανομών πιθανότητας που προέρχονται από μετρήσεις συνυπάρξεων στο σύνολο των στοιχείων δεδομένων.
- Γενετικοί αλγόριθμοι / Εξελικτικός προγραμματισμός: διατύπωση υποθέσεων σχετικά με τις εξαρτήσεις μεταξύ μεταβλητών, με τη μορφή κανόνων συσχέτισης ή κάποιου άλλου εσωτερικού φορμαλισμού.
- Fuzzy Sets: αποτελούν μια ισχυρή προσέγγιση για την αντιμετώπιση όχι μόνο των ελλιπών, θορυβωδών ή ανακριβών δεδομένων, αλλά μπορεί επίσης να είναι χρήσιμη στην ανάπτυξη αβέβαιων μοντέλων δεδομένων που παρέχουν πιο έξυπνη και ομαλότερη απόδοση από τα παραδοσιακά συστήματα.
- Rough Sets: τα ακατέργαστα σύνολα είναι μια μαθηματική έννοια που ασχολείται με την αβεβαιότητα στα δεδομένα και χρησιμοποιείται ως αυτόνομη λύση ή συνδυάζεται με άλλες μεθόδους όπως μεθόδους επαγωγής κανόνων, ταξινόμησης ή ομαδοποίησης

Η επιστήμη αυτή, βρίσκει εφαρμογή τόσο σε επιχειρήσεις (ασφάλιση, τραπεζική, λιανική) και σε επιστημονικές έρευνες (αστρονομία, ιατρική) όσο και σε θέματα κυβερνητικής ασφάλειας (εντοπισμός εγκληματιών και τρομοκρατών). Λόγω της ευρείας χρήσης της επιστήμης της εξόρυξης γνώσης από δεδομένα, διαπιστώνεται και ένα τεράστιο πλήθος πρωτότυπων και μεταρρυθμιστικών ερευνών στην παγκόσμια επιστημονική κοινότητα. Μερικά παραδείγματα για την εφαρμογή της στους διάφορους τομείς :

1. Ανάλυση αγοράς: Διάφορα πεδία όπως προφίλ πελατών, ανάλυση μεταξύ αγορών, μάρκετινγκ στόχων, καθορισμός του τρόπου αγοράς πελατών κ.λπ. Ο Ayres (2007) χρησιμοποίησε στο βιβλίο του, τη διαδικτυακή πλατφόρμα λιανικής Amazon.com, που λέει σε έναν πιθανό πελάτη ότι οι άνθρωποι που τους αρέσουν ένα συγκεκριμένο προϊόν ενδιαφέρονται επίσης και σε άλλα ορισμένα είδη ως παράδειγμα μάρκετινγκ μέσω της εξόρυξης δεδομένων.
2. Εταιρική Ανάλυση: Εδώ είναι χρήσιμο για διάφορους τομείς όπως ο οικονομικός προγραμματισμός και η αξιολόγηση των περιουσιακών στοιχείων, ο προγραμματισμός πόρων, ακόμη και ο ανταγωνισμός είναι σημαντικοί.
3. Ανίχνευση απάτης: Χρησιμοποιείται κυρίως στον τομέα των υπηρεσιών πιστωτικών καρτών, ακόμη και των τηλεπικοινωνιών για τον εντοπισμό απάτης. Οι Brause et al. (1999) καταφέρνουν να κατασκευάσουν ένα μηχανισμό ανίχνευσης απάτης που έχει ποσοστό εμπιστοσύνης 25% και εγκυρότητα αναφορικά με την ανίχνευση των περιπτώσεων απάτης της τάξης του 91%. Επιτυγχάνει δηλαδή να ανιχνεύσει σωστά στις 10 περιπτώσεις απάτης οπότε επιτυγχάνει εξοικονόμηση περίπου 9 εκατομμυρίων δολαρίων το χρόνο.
4. Ιατρική: Η εξόρυξη δεδομένων επιτρέπει πιο ακριβή διαγνωστικά. Έχοντας όλες τις πληροφορίες του ασθενούς, όπως ιατρικά αρχεία, φυσικές εξετάσεις και τρόπους θεραπείας, επιτρέπει την αποτελεσματικότερη θεραπεία. Με ιδιαίτερη εμβρίθεια στο βιβλίο τους οι Lavrač και Zupan (2005) για την εφαρμογή τεχνικών εξόρυξης γνώσης από δεδομένα σε ιατρικά και προγνωστικά προβλήματα.
5. Τηλεόραση και ραδιόφωνο. Υπάρχουν δίκτυα που εφαρμόζουν εξόρυξη δεδομένων σε πραγματικό χρόνο για τη μέτρηση του διαδικτυακού τους τηλεοπτικού (IPTV) και ραδιοφωνικού κοινού. Αυτά τα συστήματα συλλέγουν και αναλύουν ανώνυμες πληροφορίες από προβολές καναλιών, εκπομπές και προγραμματισμό. Ένα παράδειγμα, της συνεισφοράς της εξόρυξης γνώσης στο μέσα μαζικής ενημέρωσης, είναι η έρευνα των Zhu et al (2017), στην οποία δημιούργησαν ένα πρόγραμμα πρόβλεψης της δημοτικότητας, με σκοπό την διευκόλυνση και την δημιουργία στρατηγικού πλάνου των φορέων τηλεοπτικής μετάδοσης, αλλά και των διαφημιζόμενων.
6. Αστρονομία: Η χρήση τέτοιων εργαλείων εξόρυξης δεδομένων θα επιτρέψει στους επιστήμονες να εργαστούν με μεγάλα δείγματα δεδομένων. Για

παράδειγμα, οι Wang et al (2008) μέσα από ένα πρόσφατα σχεδιασμένο αυτόματο σύστημα για την εκτίμηση φωτομετρικών-ερυθρών αλλαγών, έγινε ένα ουσιαστικό εργαλείο για τον αυτόματο προσδιορισμό των φυσικών παραμέτρων των γαλαξιών, των κβάζαρ και των αστεριών.

Ένας τομέας που περιλαμβάνεται στην εξόρυξη δεδομένων είναι η συναισθηματική ανάλυση, η οποία αναφέρεται και ως εξόρυξη γνώμης. Η συναισθηματική ανάλυση επιτυγχάνεται μέσα από την επεξεργασία φυσικής γλώσσας και προσδιορίζεται η συναισθηματική πολικότητα που υπάρχει στα κειμενικά δεδομένα. Μέσα από ένα μη δομημένο και αδιοργάνωτο σώμα κειμένου εξάγεται η εξόρυξη γνώμης, έτσι ώστε να κατανοήσουν την άποψη και τα συναισθήματα που έχουν οι πελάτες μιας εταιρίας με την ίδια. Πολλές φορές, διαφημιστικές καμπάνιες από μεγάλες επιχειρήσεις χρησιμοποιούν και στηρίζονται στις τεχνικές της συναισθηματικής ανάλυσης, ώστε να καλύψουν τις ανάγκες τις αγοράς (Stedman, 2019).

Μέσα από την εξόρυξη γνώμης μπορούν να προκύψουν πολλά μοντέλα επικεντρώνονται στην άντληση πολικότητας των απόψεων (θετική, αρνητική, ουδέτερη), στην ανίχνευση των συναισθημάτων (χαρά, λύπη, θυμός, φόβος, ντροπή κ.λπ.) και στην πρόθεση ενδιαφέροντος (ενδιαφέρονται / δεν ενδιαφέρονται). Η συναισθηματική ανάλυση επιτυγχάνεται μέσα από διαφορετικές μεθόδους. Μια από αυτές τις μεθόδους είναι οι βαθμολόγηση με αστέρια (από 1 έως 5 αστέρια, με το 1 ως πολύ αρνητικό και το 5 πολύ θετικό) προϊόντων ή κριτικές και δημοσκοπήσεις. Η χρήση, επίσης, λεξικών που περιέχουν λέξεις ήδη κατηγοριοποιημένες σε συναισθήματα, αποτελεί μία ακόμα τεχνική εξόρυξης γνώμης μέσα από ένα σύστημα βαθμολόγησης του κειμένου με την ύπαρξη των λέξεων που περιέχει. Η εύρεση και κατανομή των σχόλιων σε θετικά και αρνητικά των χρηστών σε πλατφόρμες κοινωνικής δικτύωσης για την ανίχνευση της πολικότητας, είναι μια τεχνική που έχει κυριαρχήσει στην σημερινή εποχή με την ανάπτυξη των πλατφόρμων δικτύωσης (Wolff, 2020).

1.2 ΣΤΟΧΟΣ ΕΡΕΥΝΑΣ

Η ανακάλυψη Κανόνων Συσχέτισης είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Οι Κανόνες Συσχέτισης αφορούν την ανακάλυψη και διατύπωση σχέσεων που υπάρχουν στα δεδομένα. Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων. Στην παρούσα διπλωματική εργασία θα πραγματοποιηθεί, στο τομέα της εξόρυξης γνώσης, η δημιουργία κανόνων συσχέτισης, με δεδομένα κείμενου από το Twitter, με σκοπό την συσχέτιση μεταξύ λέξεων και συναισθημάτων. Λέξεις, που δεν έχουν συναισθηματική έννοια και φόρτιση και παρέχουν μια πληροφορία, θα ταυτιστούν με κάποιο συναίσθημα που αυτά προκαλούν στους χρήστες. Τα δεδομένα που θα χρησιμοποιήσουμε είναι μεγάλα σύνολα δεδομένων κειμένου από το Twitter, τα γνωστά σε όλους tweets, τα οποία έχουν ήδη ταξινομηθεί σε ένα συναίσθημα. Μέσα από επεξεργασία του συνόλου από το γενικό συναίσθημα για κάθε tweet,θα γίνει προσπάθεια άντλησης κανόνων συσχέτισης των συχνών λέξεων που εμφανίζονται στα κείμενα με την συναισθηματική φόρτιση που προκαλούν .

1.3 ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΕΡΕΥΝΑΣ

Η εξόρυξη γνώσης σε δεδομένα κειμένου είναι, σχετικά, ένας νέος επιστημονικός τομέας, στο πεδίο των υπολογιστών, και έχει παρουσιάσει ραγδαία αύξηση η χρήση της, λόγω του μεγάλου όγκου διαθέσιμων μη δομημένων δεδομένων. Οι κανόνες συσχέτισης που θα παραχθούν, στην εργασία αυτή, μετατρέπει, αυτόματα, τα δεδομένα σε προσβάσιμα και δίνει μια νέα χρήση σε αυτά. Οι σχέσεις μεταξύ των λέξεων θα προσδώσουν γνώσεις και πληροφορίες που κρύβονται στο περιεχόμενο και στο υπόβαθρο του κειμένου, παρουσιάζοντας, με αυτόν τον τρόπο νέες πληροφορίες, μοτίβα και τάσεις ανάμεσα στις λέξεις που χρησιμοποιούν οι χρήστες στα κοινωνικά δίκτυα. Τα, πλέον δομημένα, δεδομένα που θα παρουσιαστούν θα μπορούν να χρησιμοποιηθούν σε πολλές κατηγορίες προβλημάτων και εμπειριών. Η συναισθηματική ανάλυση και η πολικότητα των λέξεων, μέσα από τους κανόνες συσχέτισης, έχουν την δυνατότητα να προβλέψουν και να διαχειριστούν κινδύνους και εγκλήματα μέσα στα μέσα κοινωνικής δικτύωσης. Το φαινόμενο του εκφοβισμού στον κυβερνοχώρο, γνωστό στους νέους με τον όρο «Cyberbullying”, θα γίνει εύκολα ανιχνεύσιμο με σκοπό την αντιμετώπιση του. Στον επιχειρηματικό κόσμο, κάθε εταιρία θα παρέχει μια βελτιωμένη εξυπηρέτηση πελατών και εμπειρίας, αφού θα γνωρίζει τα συναισθήματα που προκαλεί το προϊόν τους σε κάθε πελάτη εξατομικευμένα. Μέσα

από την συναισθηματική ψυχοσύνθεση των πελατών-χρηστών θα δημιουργείται με μεγαλύτερη ακρίβεια η ψηφιακή διαφήμιση της επιχείρησης, διατηρώντας το απόρρητο του χρήστη, σε αντίθεση με την τωρινή συχνή χρήση των cookies.

Η συναισθηματική ανάλυση των tweets, με άντληση της πολικότητας κάθε, συχνά εμφανιζόμενης, λέξης δεν είναι μια πρωτοφανής έρευνα, πολλοί καταξιωμένοι ερευνητές έχουν ασχοληθεί με αυτό το θέμα. Όμως, όπως θα καταστεί σαφές και στην βιβλιογραφική ανασκόπηση, η πρόοδος στον τομέα αυτό, είναι σημαντική αλλα με λίγα χρόνια εξέλιξης, διότι αποτελεί έναν νέο κλάδο αναζήτησης. Οι περισσότεροι ερευνητές έχουν προάγει την συναισθηματική ανάλυση μέσα από αλγορίθμους μηχανικής μάθησης, όπως τον αλγόριθμο Naïve Bayes και την χρήση έτοιμων λεξικών, καταφέρνοντας με τα χρόνια να πετύχουν αρκετά μεγάλη ακρίβεια. Η εργασία, η οποία θα παρουσιαστεί παρακάτω με λεπτομέρειες, εξετάζει την δημιουργία κανόνων συσχέτισης με χρήση του αλγορίθμου Apriori, ένας αλγόριθμος που δεν έχει προταθεί από αρκετούς επιστήμονες. Τέλος οι λέξεις θα αντιστοιχηθούν σε συναισθήματα, φέρνοντας μια διαφορετική οπτική από την μέχρι τώρα ταξινόμηση σε θετικά, αρνητικά και ουδέτερα.

1.4 ΔΟΜΗ ΤΗΣ ΕΡΕΥΝΑΣ

Στο επόμενο μέρος, θα παρουσιαστεί μια ολοκληρωμένη βιβλιογραφική ανασκόπηση με τις μεγαλύτερες έρευνες που έχουν πραγματοποιηθεί στον κλάδο και αναλύονται τα αποτελέσματα, σύμφωνα με την βελτίωση της ακρίβειας και τους πρωτότυπες μεθοδολογίες που επέφερε ο κάθε ερευνητής στην δική του προσέγγιση. Στην συνέχεια, θα αναλυθεί η μεθοδολογία που ακολουθήθηκε για την παρούσα εργασία, τόσο η επεξήγηση του αλγορίθμου, όσο και η επεξεργασία των δεδομένων για την επιλογή, την μορφή, την διάταξη και την εισαγωγή τους. Στο τέταρτο κεφάλαιο θα γίνει η επίδειξη και η ερμηνεία των κανόνων συσχέτισης που παράχθηκαν και την ακρίβεια που υπέδειξαν. Έπειτα, θα εκτεθεί συνοπτικά η έρευνα και τα αποτελέσματα και θα προταθούν μελλοντικές επεκτάσεις. Οι αναφορές σε έρευνες, δημοσιεύματα, άρθρα και ιστότοποι που χρησιμοποιήθηκαν θα καταταχθούν στο τέλος.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

2.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Στο άρθρο τους οι Pang και Lee, (2004) , οι ερευνητές παρουσίασαν ένα μοντέλο για την ταξινόμηση των κριτικών για ταινίες ως «θετικό» ή «αρνητικό». Υπέβαλαν μία μέθοδο κατηγοριοποίησης κειμένου χρησιμοποιώντας μηχανική μάθηση και βρέθηκε η διαφορά μεταξύ υποκειμενικής ανίχνευσης και ταξινόμηση πολικότητας. Τα αποτελέσματα έδειξαν ότι χρησιμοποιώντας υποκειμενική ανίχνευση έδειξε σε μικρότερες κριτικές, στον αλγόριθμο Naive Bayes, ότι ήταν πιο αποτελεσματική σε σύγκριση με το πρωτότυπο έγγραφο χωρίς ανίχνευση υποκειμενικότητας. Επίσης, η ελάχιστη ταξινόμηση ενίσχυε την ακρίβεια.

Όπως συζητήθηκε από τους Kennedy και Inkpen, (2006), έγινε η σχεδίαση δύο μεθόδων συναισθηματικής ανάλυσης, με τις ετικέτες των κριτικών για ταινίες να είναι θετικές, αρνητικές ή ουδέτερες. Η πρώτη μέθοδος είναι να κατηγοριοποιηθούν οι κριτικές χρησιμοποιώντας τον αριθμό των αρνητικών και θετικών λέξεων, και να προχωρήσει στην κατάλληλη κατηγορία με μία μέτρηση των όρων χρησιμοποιώντας εξωτερικούς πόρους. Η δεύτερη μέθοδος είναι η χρήση του Support Vector Machine-SVM (εποπτευόμενα μοντέλα μάθησης), δηλαδή εποπτευόμενα μοντέλα με αλγορίθμους μάθησης. Τα αποτελέσματα δείχνουν ότι η μέθοδος υπολογισμού όρων πέτυχε μεγαλύτερη ακρίβεια και η ακρίβεια του SVM ήταν αρκετά υψηλή.

Μία άλλη λύση για την Ανάλυση Συναισθημάτων, αποτελεί η μέθοδος Opinion Digger, η οποία χρησιμοποιήθηκε από τους Moghaddam και Ester (2010). Αυτή η μη εποπτευόμενη μεθοδολογία μηχανικής εκμάθησης λειτουργούσε σε επίπεδο πρότασης στα δεδομένα από κριτικές προϊόντων. Η άποψη για το προϊόν και η τυπική βαθμολογία από τις κριτικές συσχετίζονται και συγκρίνονται (χρησιμοποιώντας το Amazon, Snapdeal, flipkart1 κ.λπ.). Η προτεινόμενη έρευνα αποτελείται από δύο δευτερεύουσες μεθόδους. Στην αρχή, εισήγαγαν τις πληροφορίες, που είναι κατακερματισμένες σε προτάσεις. Επαναλαμβανόμενα ουσιαστικά στις προτάσεις διαχωρίζονται ως απόψεις. Οι απόψεις (επαναλαμβανόμενα ουσιαστικά) εάν σχηματίζουν οποιοδήποτε επαναλαμβανόμενο σχέδιο-μοτίβο, αποθηκεύονται. Δεύτερον, οι απόψεις παραλληλίζονται με την κατευθυντήρια γραμμή αξιολόγησης

(όπως 4 σημαίνει «Καλό», 3 σημαίνει «Μέσος όρος» κ.λπ.) και ως εκ τούτου κατηγοριοποιούνται ως "Καλό", "Μέσο" και "Κακό". Σημαντικό πλεονέκτημα είναι η υψηλή απόδοσή του στη βαθμολογία προϊόντων στο επίπεδο διαστάσεων με απώλεια μόνο 0,49.

Οι Pak et al (2010) έφεραν την χρήση των N-gram (ακολουθία με n όρους) και των POS ετικετών (λεξικογραφική ανάλυση που κατηγοριοποιεί τους όρους με βάση το μέρος του λόγου που ανήκουν) ως χαρακτηριστικά που χρησιμοποιούνται μέσα από τον ταξινομητή Naïve Bayes για την κατηγοριοποίηση δεδομένων, σε θετικά, αρνητικά και ουδέτερα ως προς τα συναισθήματα. Συλλέχθηκαν με αυτόματο τρόπο από το Twitter. Η χρήση των Naïve Bayes για την ταξινόμηση δεν εξασφάλισε την ακρίβεια σε μεγάλα ποσοστά.

Οι Maas et al (2011) ασχολήθηκαν με την συναισθηματική ανάλυση σε επίπεδο εγγράφων και πρότειναν τη χρήση ενός συνδυασμού εποπτευόμενων και μη αλγορίθμων με πλούσιο συναισθηματικό περιεχόμενο για τεχνικές εκμάθησης σύλληψης των διανυσματικών λέξεων, οι οποίες περιλαμβάνουν συνεχή, πολυδιάστατα συναισθήματα και σχολιασμούς χωρίς συναισθήματα. Για αξιολόγηση, οι συγγραφείς χρησιμοποίησαν το σύνολο δεδομένων κριτικών ταινιών IMDB και το μοντέλο τους είχε καλύτερη απόδοση σε σύγκριση με άλλες τεχνικές εκμάθησης συλλογής διανυσμάτων.

Οι συγγραφείς Wang και Manning, (2012) χρησιμοποίησαν Naive Bayes-NB (είναι μια συλλογή αλγορίθμων ταξινόμησης που βασίζονται στο Θεώρημα του Bayes) και τους ταξινομητές Support Vector Machine-SVM (εποπτευόμενα μοντέλα μάθησης) και τη χρήση ενός τροποποιημένου SVM μαζί με αναλογίες καταμέτρησης NB. Τα αποτελέσματα επέφεραν ότι ο Naive Bayes ήταν καλύτερος από τους SVM σε σύντομα έγγραφα για αξιολογήσεις ταινιών. Σε έγγραφα με περισσότερα δεδομένα, ωστόσο, το SVM ήταν καλύτερο. Ο τροποποιημένος αλγόριθμος είχε καλά αποτελέσματα. Επίσης, η λογιστική παλινδρόμηση αντί του SVM απέδωσαν τα ίδια αποτελέσματα και η χρήση του bigram (ακολουθία με δύο στοιχεία) δίνει ένα σταθερό κέρδος στην συναισθηματική ανάλυση.

Όπως αναφέρθηκε στο άρθρο τους, οι Narayanan et al (2013) , προσπάθησαν να επιτύχουν ενίσχυση της ακρίβειας του αλγορίθμου Naive Bayes για την συναισθηματική ανάλυση. Πρότειναν μια προσέγγιση, η οποία να ορίσει έναν συγκεκριμένο αριθμό κατηγοριοποιήσεων συμβολοσειρών για να βελτιωθεί η ακρίβεια. Το αποτέλεσμα ήταν ότι ο συνδυασμός του χειρισμού άρνησης, η επιλογή χαρακτηριστικών και λέξεις n-grams (ακολουθία με n όρους) επέφερε βελτίωση στην ακρίβεια, χρησιμοποιώντας έναν ταξινομητή Naive Bayes με προεπιλογή έναν γραμμικά αυξανόμενο χρόνο εκπαίδευσης και δοκιμών. Στο σύνολο δεδομένων IMDB για κριτικές ταινιών η ακρίβεια έφτασε 88,80%.

Οι Amine et al (2014), για τον εντοπισμό των απόψεων που εκφράζουν τα δεδομένα και την διασφάλιση της υποκειμενικότητας τους, χρησιμοποίησαν δύο τεχνικές, την εποπτευόμενη μάθηση μέσα από το μοντέλο πιθανοτήτων Bayesian και ο σημασιολογικός προσανατολισμός μέσω από τον αλγόριθμο Turney. Το μοντέλο Bayesian στηρίζεται στα emoticons, ενώ ο αλγόριθμος Turney στα επίθετα που υπάρχουν μέσα στο κείμενο με προκαθορισμένα μοτίβα και μοντέλα ετικετών POS (λεξικογραφική ανάλυση στην οποία κατηγοριοποιεί τους όρους σε μέρη του λόγου). Με την αναφορά στην χρήση των επιθέτων που χρησιμοποιούνται, κατέστησαν σαφές ότι είναι αναγκαίο να δοθεί προσοχή στην αρνητικότητα, όταν μετριέται η πολικότητα. Οι έρευνες έχουν αποδείξει, ότι δεν είναι αρκετή η χρήση μόνο των emoticons, αλλά μαζί με τον σημασιολογικό προσανατολισμό των λέξεων αποφέρει καλά αποτελέσματα.

Η χρήση των Naïve Bayes για την ταξινόμηση δεν εξασφαλίζει την ακρίβεια σε μεγάλα ποσοστά. Στην έρευνα των Bizhanova et al (2014), τα bi-grams (συνδυασμός δύο λέξεων) μαζί με τα emoticons χρησιμοποιήθηκαν ως χαρακτηριστικά για την αυτόματη ταξινόμηση μηνυμάτων του Twitter, όμως δεν απέφεραν καλύτερη απόδοση.

Από την ανάλυση συναισθημάτων σε δεδομένα από το Twitter δεν θα έλειπε η ταξινόμηση με επεξεργασία των διαφόρων hashtags στα μηνύματα, που χρησιμοποιούν συχνά οι χρήστες του, μέσα από δύο στάδια ταξινομητή Support Vector Machine (εποπτευόμενα μοντέλα μάθησης). Οι Kharde et al (2016), εξοικειώθηκαν σε συμβολικές τεχνικές και επεξήγησαν το διάνυσμα των χαρακτηριστικών. Αυτό το

διάνυσμα είχε 8 χαρακτηριστικά, όπως τις ετικέτες POS (λεξικογραφική ανάλυση στην οποία κατηγοριοποιεί τους όρους σε μέρη του λόγου), λέξεις κλειδιά, αρνήσεις, συναισθήματα, απόλυτη τιμή των θετικών λέξεων και αρνητικών, hashtags και emoticons. Η ανάλυση πραγματοποιήθηκε σε επίπεδο πρότασης.

Οι Tripathy et al (2016) επεξεργάστηκαν τέσσερις ταξινομητές. Την μέγιστη Εντροπία (ME), Naive Bayes (NB), Support Vector Machine (SVM) και Stochastic Gradient Descent (SGD) για συναισθηματική ανάλυση στο σύνολο δεδομένων για ταινίες από το IMDB. Έκαναν χρήση των μέτρων precision, recall, f-measure και ακρίβεια για την αξιολόγηση τους. Τα αποτελέσματα έδειξαν ότι χρησιμοποιώντας έναν συνδυασμό από unigram, bigram, trigram (ακολουθία με έναν, δύο ή τρεις όρους αντίστοιχα) επέφερε καλύτερη ακρίβεια σε όλους τους ταξινομητές που χρησιμοποιήθηκαν.

Οι Bhanap et al (2017), ταξινόμησαν τις γνώμες των tweets σε θετικές , αρνητικές και ουδέτερες με την χρήση αλγορίθμων μηχανικής μάθησης.

Οι HarishRao et al. (2017), πρότειναν το VADER, ένα εργαλείο ανάλυσης συναισθημάτων, για την ανάλυση ενός προϊόντος μέσα από διαδικτυακά σχόλια. Οι ερευνητές απέδειξαν ότι το εργαλείο αυτό, παρουσιάζει καλύτερη απόδοση από άλλες μη εποπτευόμενες μεθόδους. Σε κείμενα-tweets εξασφαλίζει ακρίβεια 84,4% με 99%.

Οι Bimantara et al (2019) παρουσίασαν στην έρευνα τους, μία συναισθηματική ανάλυση για τον διαδικτυακό εκφοβισμό από σχόλια των χρηστών του Instagram. Το Instagram είναι από τα δημοφιλέστερα κοινωνικά μέσα με βάση την φωτογραφία στον κόσμο. Σε αυτό, κάθε χρήστης μπορεί να δηλώνει την προτίμηση «Μου αρέσει» και να παράγει σχόλια στις αναρτήσεις των υπόλοιπων χρηστών. Στην έρευνα τους ταξινόμησαν τα σχόλια σε σημάδια εκφοβισμού ή όχι ,ώστε να αναπτυχθεί στο μέλλον ένας αλγόριθμος για ευκολότερο εντοπισμό τέτοιων σχολίων. Η επιλογή του μέσου αυτού δεν ήταν τυχαία, αφού το 42% των περιπτώσεων που δέχεται διαδικτυακό εκφοβισμό πραγματοποιήθηκε στο Instagram. Ο αριθμός των δεδομένων που λήφθηκαν είναι 455 και η διαδικασία συλλογής δεδομένων γίνεται χρησιμοποιώντας την βιβλιοθήκη API του Instagram. Όλα τα δεδομένα θα μετατραπούν σε πεζά

γράμματα και σύμβολα, σημεία στίξης, αριθμοί, ονόματα και διευθύνσεις URL θα βγουν από το πείραμα. Ο αλγόριθμος που χρησιμοποιήθηκε για την ταξινόμηση είναι ο Naïve Bayes Classifier. Στην συνέχεια, κάθε σχόλιο θα περάσει από προεπεξεργασία(Case Folding, Cleansing, Tokenizing, Word Replacer, Stop Word Removal) και κάποια μελλοντικά εξαγόμενα στάδια με TF-IDF μέθοδο. Για αξιολόγηση και έλεγχο χρησιμοποιείται η μέθοδο K-FOLD Cross Validation. Τέλος το πείραμα χωρίζεται σε 2 μέρη, στο ένα χρησιμοποιείται Stemming(μείωση των παραλλαγών των πραγματικών λέξεων που προέρχονται από τις ίδιες βασικές λέξεις) και στο άλλο όχι. Τα αποτελέσματα κατηγοριοποιούνται σε θετικά, δηλαδή αυτά που δεν περιέχουν εκφοβισμό και σε αρνητικά, σε εκείνα τα οποία περιέχουν εκφοβισμό. Το συμπέρασμα ήταν 171 αρνητικά και 284 θετικά . Το καλύτερο αποτέλεσμα παρουσίασε ακρίβεια 83.53% από το πείραμα με την Stemming επεξεργασία.

Συναισθηματική ανάλυση σε σχόλια από μέσα κοινωνικής δικτύωσης, ερεύνησαν ο Celik και ο Osmanoglou (2020), έτσι ώστε να εξάγουν συμπεράσματα για τις σκέψεις και τα συναισθήματα των χρηστών για διάφορες εταιρίες. Τα δεδομένα παράχθηκαν από το Facebook , σχόλια και κριτικές σε σελίδες, με αριθμό συνολικά 5000 δεδομένα. Τα πειράματα πραγματοποιήθηκαν με χρήση της γλώσσας προγραμματισμού Jupyter Notebook- Python. Random Forest, Logistic Regression, Multinomial Naive Bayes, Support Vector classifier algorithms χρησιμοποιήθηκα στα δεδομένα. Τα σχόλια και οι κριτικές έχουν μετατραπεί σε αριθμητικά δεδομένα με τις μεθόδους CountVectorizer και TF-IDF. Τα δεδομένα ελήφθησαν από τα κοινωνικά δίκτυα μέσω της γλώσσας Ruby και είχαν όλα μια συγκεκριμένη μορφή. Όλα γραμμένα με μικρά γράμματα στην τουρκική γλώσσα, τα οποία μετατράπηκαν σε λατινικά. Ενώ οι αριθμοί ειδικοί χαρακτήρες, emojis και περιττές λέξεις αποκόπηκαν. Το σύνολο των 5000 δεδομένων χωρίστηκε τυχαία, στο 75% σε δεδομένα εκπαίδευσης και στο 25% σε δεδομένα ελέγχου. Τα αποτελέσματα κατηγοριοποιήθηκαν σε 3 κατηγορίες με βάση τ συναίσθημα που εκφράζουν, σε θετικά, αρνητικά και ουδέτερα. Τα δεδομένα εκπαίδευσης έβγαλαν ως συμπέρασμα 1004 θετικά, 1000 αρνητικά και 1000 ουδέτερα. Το ποσοστό επιτυχίας κυμαίνεται από 42 έως 91%, αρκετά μικρό λόγω μη τήρηση κανόνων γραμματικής. Το εύρος τιμών ακρίβειας είναι αρκετά ευρύ. Το καλύτερο αποτέλεσμα ήταν ακρίβεια 0.7 με χρήση λογιστικής παλινδρόμησης , ενώ το μικρότερο 0.56 με χρήση KNN αλγορίθμου. Τέλος η σύγκριση TF-IDF και CountVectorizer + TF-IDF δεν παρουσιάζει σημαντικές διαφορές.

2.2 ΛΕΞΙΚΑ

Οι Rajman και Besancon, (1997) μέσα από δύο παραδείγματα παρουσίασαν τεχνικές εξόρυξης κειμένου, δημιουργία συσχετισμών, πρότυπες εξαγωγές εγγράφων και με επεξεργασία φυσικής γλώσσας τεχνικές. Στην εξόρυξη των συσχετισμών, είχαν εξάγει κανόνες συσχέτισης από μια συλλογή ευρετηρίου από έγγραφα. Έγινε ανάλυση του συσχετισμού μεταξύ λέξεων-κλειδιών και σε μια συλλογή από έγγραφα ευρετηρίου, ώστε να εξάγουν τις πληροφορίες που ζητούσαν. Αρκετοί ερευνητές, όπως οι Ahonen et al (1997 και 1998)και οι Mannila et al (1998), που εφάρμοσαν τις παραπάνω τεχνικές εξόρυξης για να σχηματίσουν κανόνες από κείμενα, δηλαδή η εξόρυξη κανόνων χρησιμοποιείται για την ανάλυση της γλώσσας, γιατί διατηρεί την διαδοχική δομή των όρων σε ένα έγγραφο.

Οι Zhuang et al (2006) πρότειναν η χρήση μιας προσέγγισης πολλαπλών γνώσεων για την αυτόματη παραγωγή μιας ομαδοποίησης των κατηγοριών με βάση τα χαρακτηριστικά σε ένα σύστημα ανάλυσης κριτικής μιας ταινίας. Στην προσέγγισή αυτή συνδυάζονται η στατιστική ανάλυση, WordNet (λεξικολογική βάση δεδομένων με τις σημασιολογικές σχέσεις των λέξεων) και η γνώση ταινιών. Τα αποτελέσματα δείχνουν ότι η προσέγγισή ήταν αποτελεσματική.

Ο Denecke (2008), αντιμετώπισε την ανάλυση κειμένου σε συναισθήματα με ένα πολύγλωσσο σύστημα. Εκμεταλλεύτηκαν τους λεξικούς πόρους στα Αγγλικά από το SentiWordNet (ένας λεξικός πόρος στον οποίο κάθε όρος περιγράφεται με βάση πόσο αντικειμενικός είναι με τις κατηγορίες θετικός, αρνητικός ή ουδέτερος) . Με την συνδρομή ενός λογισμικού μετάφρασης, μετέφρασαν αρχικά διαφορετικές γλώσσες στα Αγγλικά. Έπειτα, τα κατηγοριοποιήσαν σε «θετικά» ή «αρνητικά» αναζητώντας λέξεις με συναισθήματα, όπως επίθετα. Ο συγγραφέας χρησιμοποίησε τις γερμανικές κριτικές ταινιών του Amazon και παραλλήλισε την δουλειά του σε έναν ταξινομητή στατιστικής πολικότητας με βάση n-grams (ακολουθία με n όρους). Με τα αποτελέσματα αποφάνθηκε ότι η προσέγγισή του ήταν καλή για πολύγλωσση συναισθηματική ανάλυση.

Η ταξινόμηση συναισθημάτων από διαδικτυακές κριτικές πελατών για κάποιο προϊόν με την χρήση της συνάφειας των λεξικών με την δομή της πρότασης (Lexical Contextual Sentence Structure) ήταν μια πρωτοπόρα τεχνική από τον Khan (2011). Ουσιαστικά, ήταν μια σημασιολογική ή βασισμένη σε κανόνες (Λεξικό πολικότητας) προσέγγιση της ανάλυσης των κριτικών από πελάτες. Αρχικά, σε πρώτο στάδιο, οι προτάσεις κατακερματίστηκαν με την είσοδο τους και την αποθήκευσαν με την χρήση της μεθόδου «POS» (λεξικογραφική ανάλυση στην οποία κατηγοριοποιεί τους όρους σε μέρη του λόγου). Στην συνέχεια, υπολογίστηκε η πολικότητα της κάθε πρότασης με κύρια διάκριση το πλαίσιο και την δομή τους. Τα ουσιαστικά χαρακτηρίστηκαν ως κύριο πηγή μέτρησης της πολικότητας. Το πρόγραμμα SentiWordNet (ένας λεξικός πόρος στον οποίο κάθε όρος περιγράφεται με βάση πόσο αντικειμενικός είναι με τις κατηγορίες θετικός, αρνητικός ή ουδέτερος) χρησιμοποιήθηκε για τον χαρακτηρισμό των προτάσεων , ως θετικές ή αρνητικές, από την σημασιολογική βαθμολογία των διαθέσιμων λέξεων. Προσεγγίζει ακρίβεια κοντά στο 86%. Ως θετικό της προσέγγισης αυτής είναι το γεγονός ότι είναι ένα ανεξάρτητο πεδίο ορισμού, όμως παρουσιάζει πλήρη εξάρτηση από το SentiWordNet.

Η χρήση του Twitter ως εργαλείο εξαγωγής δεδομένων δεν είναι τυχαία, αφού περιέχει κείμενα με ανώτατο όριο τους 140 χαρακτήρες. Η συναισθηματική ανάλυση σε tweets, έχει συζητηθεί από την ερευνητική κοινότητα και έχει αποφέρει αποτελέσματα με διάφορες πτυχές συναισθημάτων, επιπέδων, προσεγγίσεων, μεθοδολογιών, εφαρμογών και τεχνικών για την εξόρυξη απόψεων με βάση λεξικά ή μετρήσεις αξιολογήσεων. Σε αυτά τα πλαίσια, προτάθηκε η χρήση ενός υπάρχοντος προγράμματος, του Weka,(2012) ή ενός λογισμικού που θα κατηγοριοποιούσε τα κείμενα σε θετικά ή αρνητικά σε σύγκριση με κάποιες ήδη αποθηκευμένες λέξεις και τα emoticons θα λαμβάνονταν υπόψη ως χαρακτηριστικά.

Σε μια άλλη έρευνα η Kaur και η Baghla (2018) πραγματοποίησαν συναισθηματική ανάλυση σε αγγλικά tweets. Στην πλατφόρμα κοινωνικής δικτύωσης Twitter οι χρήστες της εκφράζουν και μοιράζονται μέσω γραπτού λόγου σκέψεις και απόψεις για γεγονότα, προϊόντα σε διάφορα θέματα. Η ανάλυση, αυτή την φορά, γίνεται με ταξινόμηση και με σύγκριση των λέξεων σε λεξικά, στα οποία υπάρχει ήδη προκαθορισμένο το συναισθηματικό κίνητρο. Χρησιμοποιείται η γλώσσα

προγραμματισμού Java και Net beans για την διεπαφή. Ως βάση δεδομένων χρησιμοποιείται η γλώσσα SQL. Αρχικά δημιουργείται ένα λεξικό με θετικές και αρνητικές λέξεις. Συλλέγονται τα tweets από την πλατφόρμα Twitter. Ακολουθεί, ένα από τα σημαντικότερα κομμάτια, η προεπεξεργασία των δεδομένων. Στο κομμάτι αυτό, γίνεται διόρθωση των ορθογραφικών λαθών, κατάταξη των λέξεων σε λίστες θετικές-αρνητικές, αφαίρεση των περιττών λέξεων, URL, ονόματα και διπλότυπους χαρακτήρες, και αντικατάσταση των συντομογραφιών. Στην συνέχεια υπολογίζεται το συναισθηματικό σκορ κάθε tweet από την σύγκριση με τα λεξικά. Η κατηγοριοποίηση του πειράματος κατατάσσεται σε θετικά, αρνητικά και ουδέτερα. Το πείραμα χωρίζεται σε 4 τυχαία σύνολα δεδομένων. Το πρώτο σύνολο με 74 tweets, είχε ως αποτέλεσμα 22 θετικά, 20 αρνητικά και 32 ουδέτερα, και πέτυχε ακρίβεια 42%, Παρόμοια αποτελέσματα απέδωσαν και τα υπόλοιπα 3 σύνολα, με ακρίβεια 64%, 57% και 55% αντιστοίχως. Η αποδοτικότητα του αλγορίθμου κυμαίνεται στο 86%.

2.3 ΜΕΤΡΑ- ΒΑΡΗ

Στην ανάλυση συναισθημάτων τα επίθετα και τα επίρρημα φέρνουν μεγαλύτερη εγκυρότητα, παρά η χρήση μόνο των επιθέτων. Έχοντας, με αυτόν τον τρόπο, μια γλωσσική προσέγγιση της ανάλυσης συναισθημάτων στο επίπεδο εγγράφου, που προτάθηκε από τους Benamara et al (2006). Αυτό το ερευνητικό έργο ξεκίνησε με τη μέτρηση της έντασης του βαθμού των επιρρημάτων (χρησιμοποιώντας γλωσσικούς ταξινομητές) και συνδυασμούς με επιρρήματα και επίθετα (χρησιμοποιώντας μεθόδους βαθμολογίας). Ο στόχος αυτών των μεθόδων δεν είναι παρά να προστεθεί ένα σχετικό βάρος (σε μια μεταβλητή, με μια κλίμακα από 0 έως 1) της βαθμολογίας του επιρρήματος σε σχέση με τη βαθμολογία του επίθετου. Αυτός ο τρόπος, επιδιώκει να προσδιορίσει ποιο βάρος είναι αποτελεσματικότερο στις ανθρώπινες απόψεις. Πειραματίστηκαν σε περίπου 200 έγγραφα από νέους πόρους που αποδεικνύει ότι η ανάλυση αποφέρει καλύτερα συμπεράσματα όταν οι απόψεις αποτελούνται από 35% περισσότερα επιρρήματα από ότι επίθετα. Παράγεται η συσχέτιση Pearson (συσχέτιση μεταξύ ανθρώπινου συναισθήματος και αλγορίθμων ανάλυσης συναισθημάτων) περίπου 0,47 (κυμαίνεται μεταξύ -1 και 1).

Οι Liu et al (2006) αξιολογούν την αποτελεσματικότητα των συστημάτων στάθμισης για την εξαγωγή λέξεων-κλειδιών για την ομαδοποίηση των γονιδίων, πάνω στην

ανάλυση συναισθημάτων σε κείμενο. Το αποτέλεσμα, που παράγεται από τις σταθμισμένες λέξεις-κλειδιά με TF-IDF (στατιστικό μέτρο που αξιολογεί πόσο σχετική είναι μια λέξη με ένα έγγραφο σε μια συλλογή εγγράφων) ξεπερνούσαν αυτές που παράγονται από κανονικοποιημένες σταθμισμένες λέξεις-κλειδιά με z-score (Η βαθμολογία Z είναι μια αριθμητική μέτρηση που περιγράφει τη σχέση μιας τιμής με το μέσο όρο μιας ομάδας τιμών).

Οι συγγραφείς, Chen και Wu (2008), πρότειναν μια τεχνική εξόρυξης κειμένου που ανακαλύπτει κανόνες συσχέτισης από έγγραφα για έναν συγκεκριμένο χρήστη. Προέρχεται από τις γνώσεις του χρήστη και από το επίπεδο γνώσης των κειμένων και εκμεταλλεύεται αυτές τις γνώσεις με τη μορφή κανόνων σύνδεσης. Επιπλέον, εφαρμόζεται TF-IDF (στατιστικό μέτρο που αξιολογεί πόσο σχετική είναι μια λέξη με ένα έγγραφο σε μια συλλογή εγγράφων) για επιλογή σημαντικών και ουσιαστικών φράσεων από κάθε έγγραφο.

Στην έρευνα που πραγματοποίησαν οι Abbasi et al(2008), προτάθηκε ένα πλαίσιο εργασίας για ταξινόμηση κριτικών για ταινίες από διαδικτυακές ανοιχτές συνομιλίες σε πολλές γλώσσες (Αγγλικά και Αραβικά). Οι συγγραφείς χρησιμοποίησαν γενετικό αλγόριθμο με εντροπία (EWGA) για βελτίωση της απόδοσης. Για αξιολόγηση μιας ταινίας χρησιμοποιήσαν ως σύνολο δεδομένων τις κριτικές και τα αποτελέσματα έδειξαν ότι η χρήση του EWGA με εποπτευόμενα μοντέλα μάθησης έφερε υψηλότερη απόδοση σε σύγκριση με άλλες μεθόδους επιλογής χαρακτηριστικών, με ακρίβεια άνω του 91%.

2.4 ΚΑΤΑΤΑΞΗ- ΔΗΜΟΤΙΚΟΤΗΤΑ

Ο Moghaddam και ο Ester (2011) ανέφεραν για πρώτη φορά την αλληλεξαρτημένη λανθάνουσα κατανομή Dirichlet, η οποία βαθμολογεί το προϊόν σε επίπεδο διαστάσεων μέσα από ένα γραφικό μοντέλο. Ένα καθοριστικό κριτήριο για την αξιολόγηση των προϊόντων είναι ο αριθμός των αστεριών που αποδίδει ένας χρήστης. Στην έρευνα τους, θεώρησαν ότι υπάρχει αλληλεξάρτηση ανάμεσα στα ουσιαστικά,

που θεωρούνται κύρια πηγή κατηγοριοποίησης, και την βαθμολόγηση με αστέρια. Το μοντέλο παρουσιάζει μια ομαδοποίηση με τα ουσιαστικά ως κεφαλή των συμπλεγμάτων και τα αστέρια ως διανομές. Τα στοιχεία, που εμφανίζονται στα διακριτά δεδομένα, αναπαρίσταται ως ένα πεπερασμένο σύνολο με λανθάνουσες μεταβλητές. Η ακρίβεια ακούμπησε το 73%. Το μειονέκτημα της προσπάθειας είναι ότι μπορεί να παρουσιαστούν λάθη στην αναπαράσταση των δεδομένων και στην γραφική αναπαράσταση, τα οποία να οδηγήσουν σε μη αναμενόμενη έξοδο.

Το μοντέλο με τις περισσότερες κοινές αναφορές ήταν αυτό του Albornoz (2011) και της ομάδας του, για την συναισθηματική ανάλυση των αξιολογήσεων για προϊόντα. Η αξιολόγηση του προϊόντος γίνεται σε παγκόσμιο επίπεδο και λαμβάνει τις απόψεις ολόκληρες, χωρίς κατακερματισμό. Η μέθοδος ακολουθεί 4 στάδια.

1. Επισήμανση των σημαντικών χαρακτηριστικών
2. Προσδιορισμό των προτάσεων που περιέχουν τα χαρακτηριστικά.
3. Υπολογισμός πολικότητας και δυναμικότητας
4. Κατάταξη της κριτικής σε κατηγορίες.

Η ακρίβεια πρόβλεψης που παράγεται είναι 71% (3 κατηγορίες) και 46,9% (5 κατηγορίες).

Οι Chen et al. (2013), μέσα από την εξαγωγή tweets, που έχουν σχολιαστεί, για ένα συγκεκριμένο θέμα συζήτησης, προσπάθησαν να αποδώσουν ένα μοντέλο κατάταξης από το Twitter για ανάλυση συναισθημάτων. Έλαβαν υπόψη τις αναφορές , τους συνδέσμους, τις πληροφορίες του συγγραφέα και τον αριθμό των ακολούθων του. Οι ερευνητές, την επόμενη χρονιά, εισήγαγαν ένα σύστημα φιλτραρίσματος, το CrowdE, το οποίο δημιουργεί φίλτρα πάνω στα tweets με πλήθος ετικετών.

Οι Anwar et al (2015) επικεντρώθηκαν στην εύρεση των συναισθημάτων σε καθορισμένες τοποθεσίες μέσα από δεδομένα από χρήστες του Twitter, επιτρέποντας τις εταιρίες να προσεγγίζουν τις περιοχές με αρνητικό δείκτη δημοτικότητας.

2.5 ΑΛΛΕΣ ΥΛΟΠΟΙΗΣΕΙΣ

Προγενέστερες έρευνες σχετικά με την εξόρυξη κανόνων συσχέτισης από το κείμενο έχουν αναλύσει τη χρήση των λέξεων-κλειδιών που έχουν εκχωρηθεί χειροκίνητα. Όπου, χρησιμοποίησαν λέξεις-κλειδιά, ως συνδετικούς κρίκους για την δημιουργία κανόνων συσχέτισης, όπως παρουσίασαν οι Feldman και Dagan (1995) και οι Feldman και Hirsh (1996). Τα μειονεκτήματα των προσεγγίσεων που εκχωρούν τις λέξεις-κλειδιά χειροκίνητα είναι ότι:

(1) είναι αρκετά χρονοβόρα διαδικασία

(2) οι λέξεις-κλειδιά είναι σταθερές (δηλαδή, δεν αλλάζουν με την πάροδο του χρόνου ή διαφέρουν ανάλογα με έναν συγκεκριμένο χρήστη)

(3) οι λέξεις-κλειδιά υπόκεινται σε ασυμφωνία

(4) οι πηγές κειμένου περιορίζονται μόνο σε εκείνους που περιέχουν τις λέξεις-κλειδιά.

Το σύστημα OPINE, που προτάθηκε από τον Propescu (2005), είναι μια μη εποπτευόμενη εξαγωγή πληροφοριών μέσω διαδικτυακού συστήματος, το οποίο εξάγει χαρακτηριστικά για κάποιο προϊόν από απόψεις και κριτικές προϊόντων. Ελέγχει την μελλοντική δυνατότητα, την άποψη του κοινού σχετικά με το χαρακτηριστικό του προϊόντος, καθορίζει την πολικότητα των απόψεων και στη συνέχεια κατατάσσει το προϊόν ανάλογα . Για την μελλοντική κατάταξη, εξάγονται τα ουσιαστικά από τις απόψεις και τις κριτικές του προϊόντος . Υψηλότερα σκορ από τη συχνότητα κατωφλίου διατηρούνται αλλιώς απορρίπτονται. Οι ερευνητές χρησιμοποιούν χειροκίνητη δημιουργία κανόνων για την εξαγωγή δεδομένων. Η προώθηση του OPINE είναι ο τομέας της ανεξαρτησίας. Αλλά αποτυγχάνει για την χρήση στην πραγματική ζωή, καθώς το σύστημα OPINE δεν είναι εύκολα διαθέσιμο.

Οι Norvag et al (2006) περιέγραψαν την εξόρυξη κανόνων συσχέτισης σε χρονικές συλλογές κειμένων και την επεξεργασία της χρονικής διαδικασίας μέσα από διάφορα βήματα.

Οι ερευνητές Lotfy et al (2016) παρουσίασαν έναν νέο αλγόριθμο για την συναισθηματική ανάλυση από έναν ιστότοπο συζητήσεων στον οποίο ένας οργανισμός

εμπιστεύεται και εξετάζει τις κριτικές του με μια προσπάθεια δημιουργίας βάσης δεδομένων. Το πλεονέκτημα ήταν ότι δημιούργησε μια δομή για τις εξαγόμενες απόψεις. Με αυτόν τον τρόπο, καθορίζονται τα πεδία των χαρακτηριστικών που θα ελεγχθούν στην βάση δεδομένων. Στην δημιουργημένη βάση τα μη δομημένα δεδομένα μετατρέπονται σε ημι-δομημένα ή σε δομημένα, και αποθηκεύουν λεπτομέρειες από τα tweets, όπως μήνυμα, ταυτότητα όνομα χρήστη, χρόνος και βαθμολογία συναισθήματος. Η διαδικασία αυτή, έγινε με δεδομένα από το Twitter με σύνδεση στο Alchemy API και την χρήση της μεθόδου κλήσης REST.

Οι Asur et al (2018) διεξήγαγαν έρευνα για την αντιμετώπιση των προβλημάτων στα στάδια της ανακάλυψης, συλλογής και επεξεργασίας των δεδομένων από κοινωνικά μέσα και δημιουργήθηκε δομημένη βιβλιογραφία για την ανίχνευση των λύσεων. Δημιουργώντας, έτσι, ένα προσοδοφόρο έδαφος για μελλοντική χρήση των δεδομένων από κοινωνικές πλατφόρμες, αναλύοντας τα σε συναισθήματα και μελετώντας τον τρόπο διάδοσης των θετικών και αρνητικών απόψεων.

3. ΜΕΘΟΔΟΛΟΓΙΑ

3.1 ΑΛΓΟΡΙΘΜΟΙ

Ένας από τους πιο σημαντικούς κλάδους της εξόρυξης γνώσης από δεδομένα είναι η εξόρυξη κειμένου. Ο τομέας της Ανακάλυψης Γνώσης μέσα από Κείμενο (Knowledge Discovery in Text KDT) καθώς και ο τομέας της εξόρυξης κειμένου (Text Mining) είναι δυο ταχέως αναπτυσσόμενοι τομείς κυρίως λόγω της μεγάλης ανάγκης για ανάλυση τεράστιων ποσοτήτων δεδομένων κειμένων, τα οποία παράγονται καθημερινά και σε μεγάλο βαθμό οφείλεται στην τεράστια ανάπτυξη του διαδικτύου στην εποχή αυτή (Βαζιργιάννης, Χαλκίδη, 2003).

Η ανακάλυψη γνώσης σε κείμενο (KDT) καθώς και η εξόρυξη κειμένου (TM) έχει προσεγγίσει αρκετούς ερευνητές, αλλά αποτελεί ένα σχετικά νέο πεδίο έρευνας με στόχο την λύση του προβλήματος της υπερ-πληθώρας πληροφοριών χρησιμοποιώντας κάποια εργαλεία και αλγόριθμους, όπως την εξόρυξη δεδομένων, την διαχείριση των πληροφοριών, την επεξεργασία φυσικής γλώσσας και πολλά άλλα.

Χρησιμοποιούμε συνήθως τον όρο KDT για να δηλώσουμε την διαδικασία της μετατροπής των μη δομημένων δεδομένων, προερχόμενα από κείμενο σε υψηλής στάθμης πληροφοριών και γνώσεων, ενώ ο όρος TM αναφέρεται κυρίως για κάποιο βήμα της διαδικασίας να αντληθεί γνώση από το κείμενο, το οποίο έχει να κάνει με την δημιουργία προτύπων από κειμενικά δεδομένα. Αν επεκτείνουμε τον ορισμό για την ανακάλυψη γνώσης σε δεδομένα των Fayyad και Piatetsky – Shapiro μπορούμε να δώσουμε τον εξής απλό ορισμό :«Η ανακάλυψη γνώσης σε Κείμενο (KTD) είναι μια μη τετριμμένη διαδικασία ανακάλυψης έγκυρων, καινούργιων, δυνητικά χρήσιμων και τελικά κατανοητών προτύπων σε δεδομένα κειμένου».

Η Εξόρυξη Γνώσης από Κείμενα μπορεί να ορισθεί ως μία διαδικασία εξαγωγής νέας πληροφορίας μέσω της οποίας ο χρήστης αλληλοεπιδρά σε μία συλλογή κειμένων χρησιμοποιώντας ένα σύνολο εργαλείων ανάλυσης (Sehral, 2008). Η εξόρυξη κειμένου αναγνωρίζεται ως μια αυτόματη διαδικασία χρησιμοποιώντας την επεξεργασία φυσικής γλώσσας, ώστε να αποσπαστεί η πληροφορία από το κείμενο. Οι μηχανές-υπολογιστές μπορούν πλέον να κατανοήσουν τα δεδομένα που έχουν μετατραπεί σε

πληροφορίες, ταξινομώντας τα κείμενα ανά συναίσθημα, θέμα και πρόθεση με αυτήν την αυτοματοποιημένη διεργασία. Με δομημένη μορφή, μετά την εξαγωγή, παρουσιάζονται οι πληροφορίες, για περαιτέρω ανάλυση ή μετάδοση τους με πίνακες HTML σε σύμπλεγμα, χάρτες μυαλού, γραφήματα κ.λπ. Η εξόρυξη κειμένου χρησιμοποιεί μια ποικιλία μεθοδολογιών για την επεξεργασία του κειμένου, όπως είναι η επεξεργασία φυσικής γλώσσας (NLP) .

Η Επεξεργασία του φυσικού λόγου είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής, της τεχνητής νοημοσύνης και της γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των ανθρώπινων (φυσικών) γλωσσών (Βλαχάβας et al. , 2006). Οι μηχανές,δηλαδή, «κατανοούν» το κείμενο μέσα από την επεξεργασία της φυσικής γλώσσας, όπως με το ίδιο τρόπο ένας άνθρωπος έχει την ικανότητα να εμπεδώσει μια φυσική γλώσσα, για παράδειγμα τα αγγλικά, ισπανικά κλπ. Επίσης, πέρα από την επεξεργασία της φυσικής γλώσσας, περιλαμβάνεται και η δημιουργία της, έτσι ώστε να μπορεί η μηχανή να παράξει με την σειρά της κείμενο, σαν τον ανθρώπινο διάλογο.

Βασικές διαδικασίες ανάλυσης κειμένου που χρησιμοποιούνται είναι :

- Η μείωση των διαστάσεων με τον προσδιορισμό της ριζικής προέλευσης των λέξεων
- Η ανάκτηση πληροφοριών
- Επεξεργασία φυσικής γλώσσας, όπως μέρος της επισήμανσης ομιλίας, συντακτική ανάλυση και άλλους τύπους γλωσσικής ανάλυσης.
- Η αναγνώριση ονομαστικής οντότητας πχ αναγνώριση ονομάτων ή συντομογραφιών
- Η αποσαφήνιση, η χρήση ενδείξεων με βάση τα συμφραζόμενα
- Αναγνώριση οντοτήτων που προσδιορίζονται με μοτίβο πχ τηλεφωνικοί αριθμοί ή ποσότητες
- Συγκέντρωση εγγράφων, αναγνώριση συνόλων παρόμοιων εγγράφων κειμένου.
- Coreference : προσδιορισμός των ουσιαστικών φράσεων και άλλων όρων που αναφέρονται στο ίδιο αντικείμενο
- Η αναγνώριση των ενώσεων μεταξύ οντοτήτων
- Η ανάλυση συναισθημάτων

- Η ποσοτική ανάλυση κειμένου

Μια από τις σημαντικότερες διεργασίες της εξόρυξης γνώσης από δεδομένα είναι η εξαγωγή κανόνων συσχέτισης (association rules). Προκαλούν σημαντικό ενδιαφέρον στην επιστημονική κοινότητα, καθώς με συνοπτικό και κατανοητό τρόπο εκφράζουν τις χρήσιμες πληροφορίες που έχει ανάγκη ο χρήστης.

Οι κανόνες συσχέτισης φανερώνουν, ουσιαστικά, κρυμμένες συνδέσεις μεταξύ κάποιων γνωρισμάτων ή προϊόντων ή εμπειριών. Η μορφή που παραθέτουν την σχέση είναι $A \rightarrow B$, με το A και το B να αποτελεί ένα στοιχείο από το σύνολο γνωρισμάτων I μέσα στα δεδομένα και $A \subseteq I$, $B \subseteq I$, $A \cap B = \emptyset$,όπως ορίζεται ο κανόνας συσχέτισης σύμφωνα με το βιβλίο « Introduction to Data Mining», των Tan, Steinbach, Kumar (2006).

Μέσα σε μία πληθώρα δεδομένων οι κανόνες συσχέτισης ψάχνουν τις συνδέσεις που μπορούν να αντλήσουν. Δημιουργούνται κανόνες, στους οποίους η εμφάνιση κάποιου στοιχείου προβλέπεται και στηρίζεται από κάποιο άλλο στοιχείο του συνόλου συναλλαγών. Επεξηγηματικά, όταν εμφανίζεται ένα στοιχείο, το αναμενόμενο είναι να εμφανιστεί μαζί και κάποιο άλλο συγκεκριμένο στοιχείο, αλλά δεν υποστηρίζεται ότι η ύπαρξη ενός στοιχείου αποτελεί την αιτία εμφάνισης κάποιου άλλου. Ένα από τα χαρακτηριστικά παραδείγματα αποτελεί οι κανόνες συσχέτισης μεταξύ δύο προϊόντων. Αν ένα προϊόν A , για παράδειγμα, αγοράζεται με την ίδια συχνότητα και κάποια συγκεκριμένη ημέρα με το προϊόν B τότε η επιχείρηση μπορεί να εξάγει συμπεράσματα για το μάρκετινγκ που θα χρησιμοποιήσει στην προώθηση των δύο προϊόντων. Η χρήση των κανόνων συσχέτισης με τα προϊόντα, όπως περιεγράφηκε, προέκυψε από μία έρευνα που υλοποίησε η Heath (1992) για λογαριασμό μιας μεγάλης βιομηχανίας. Η Heath και η ομάδα της έλεγξε όλες τις αποδείξεις- συναλλαγές που έγιναν στην συγκεκριμένη βιομηχανία, ώστε να βρει συσχετίσεις για την αγορά των προϊόντων τους. Μετα από ενδελεχή μελέτη προέκυψε ότι οι άνδρες με ηλικία 30-40 ετών, την Παρασκευή και ώρα 4-7 μ.μ. που έκαναν ψώνια και αγόρασαν πάνες μωρού, το πιθανότερο είναι το καλάθι του να περιείχε και μπύρες.

Παρακάτω παρατίθενται κάποιοι ορισμοί για την αξιολόγηση και την εύρεση των κανόνων συσχέτισης, όπως αναφέρονται από τους Tan, Steinbach, Kumar (2006), στο βιβλίο « Introduction to Data Mining».

Στοιχειοσύνολο : «Ένα σύνολο από διακριτά στοιχεία ορίζεται ως $I=\{i_1, i_2, \ldots, i_k\}$, ενώ ένα στοιχειοσύνολο ορίζεται ένα υποσύνολο του I».

Συναλλαγές : « Ένα σύνολο από συναλλαγές παρουσιάζεται ως $T=\{t_1, t_2, \ldots, t_N\}$, όπου κάθε t_i είναι ένα στοιχειοσύνολο».

Για την αξιολόγηση των κανόνων συσχέτισης και την εύρεση του βαθμού ενδιαφέροντος χρησιμοποιούμε την υποστήριξη (support) και την εμπιστοσύνη (confidence).

Υποστήριξη : «Υποστήριξη κανόνα-support (s) είναι το ποσοστό των συναλλαγών που περιέχουν το X και το Y ($X \cup Y$) με τύπο υπολογισμού $\sigma(X \cup Y)/|T|$, όπου $\sigma(X \cup Y)$ η συχνότητα εμφάνισης του στοιχειοσυνόλου που περιέχει το X και το Y και $|T|$ ο αριθμός των δοσοληψιών».

Εμπιστοσύνη : «Εμπιστοσύνη – confidence (c) αναφέρει πόσες από τις συναλλαγές σε ποσοστό που περιέχουν και το X και το Y ($X \cup Y$) με τύπο $\sigma(X \cup Y)/\sigma(X)$».

Η εξαγωγή ενός κανόνα συσχέτισης προϋποθέτει την ικανοποίηση κάποιων ελάχιστων κριτηρίων για την υποστήριξη και την εμπιστοσύνη.

Ελάχιστη υποστήριξη: «Ο κανόνας πρέπει να έχει υποστήριξη μεγαλύτερη από το όριο, που ονομάζεται ελάχιστη υποστήριξη (minsup)».

Ελάχιστη εμπιστοσύνη: «Ο κανόνας πρέπει να έχει εμπιστοσύνη μεγαλύτερη από το όριο, που ονομάζεται ελάχιστη εμπιστοσύνη(minconf)».

Συχνά στοιχειοσύνολα: «Τα στοιχειοσύνολα(itemsets) που έχουν υποστήριξη μεγαλύτερη από το minsup καλούνται συχνά ή μεγάλα».

Η εξαγωγή κανόνων συσχέτισης περιορίζεται σε 2 προβλήματα , την ανακάλυψη των συχνών στοιχειοσυνόλων στα δεδομένα και, έπειτα, την δημιουργία του κανόνα που ικανοποιεί το ελάχιστο όριο εμπιστοσύνης. Με αυτόν τον τρόπο, γίνεται και η εκτίμηση για το αν ο κανόνας, που δημιουργήσαμε, έχει τα κριτήρια που ζητάει ο χρήστης.

Το πρώτο πρόβλημα, είναι και αυτό που έχει απασχολήσει αρκετά και τους ερευνητές, αφού ενέχει αρκετή δυσκολία. Σε αρχικό στάδιο, εντοπίζονται όλα τα δυνατά συχνά στοιχειοσύνολα και τοποθετούνται σε ένα πλέγμα. Με την βοήθεια του αλγορίθμου Brute- force υπολογίζεται, για κάθε στοιχειοσύνολο το οποίο αποτελεί και ένα υποψήφιο συχνό στοιχειοσύνολο , η εμπιστοσύνη του σύμφωνα με την βάση δεδομένων που τα διατρέχει. Η υλοποίηση του αλγορίθμου αυτού, αποτελεί μια αρκετά χρονοβόρα και υπολογιστικά ακριβή διαδικασία. Στο δεύτερο πρόβλημα, η λύση του είναι πιο απλή. Έστω L όλα τα συχνά στοιχειοσύνολα που βρέθηκαν στο πρώτο βήμα, ανακαλύπτονται όλα τα υποσύνολα τους, αποκλείοντας αυτά που είναι ίσα με το κενό. Αν κάθε υποσύνολο χαρακτηριστεί ως a, τότε αντλείται ο κανόνας a => (L-a), αν ο λόγος sup(l)/sup(a), της εμπιστοσύνης του κανόνα είναι τουλάχιστον όσο και το όριο ελαχίστης εμπιστοσύνης (Κοτσιανίτης, Κανελλόπουλος, 2006)

Ένας άλλος αλγόριθμος για την υλοποίηση του πρώτου βήματος, την εύρεση συχνών στοιχειοσυνόλων είναι η Αρχή Apriori, η οποία αναφέρει «Αν ένα στοιχειοσύνολο είναι συχνό τότε όλα τα υποσύνολα του είναι συχνά». . Η μαθηματική έκφραση της αρχής είναι : ότι για κάθε X, Y ισχύει $(X \subseteq Y) \Rightarrow S(X) \geq S(Y)$. Ισχύει, επίσης, και το ισοδύναμο της Αρχής Apriori, δηλαδή ότι «Αν ένα στοιχειοσύνολο είναι μη συχνό, τότε όλα τα υπερσύνολα του είναι μη συχνά».

Ο αλγόριθμος Apriori ήταν ένας από τους πρώτους αλγόριθμους που χρησιμοποιήθηκε για εξόρυξη γνώσης και δημιουργία κανόνων συσχέτισης. Αργότερα, οι Agarwal και Srikant (1994), τον βελτίωσαν και έγινε γνωστός ως Αρχή Apriori. Ο κώδικας που δημοσίευσαν επεξηγεί την προσέγγιση της Αρχής, κάνοντας συνεχόμενες επαναλήψεις, έτσι ώστε να ολοκληρωθεί η εύρεση των συχνών αντικειμένων στα δεδομένα με διάφορους συνδυασμούς ανάμεσα τους. Παρακάτω, ακολουθεί ο κώδικας όπως παρουσιάστηκε από τους Agarwal και Srikant :

Apriori(T,e)

$L1 \leftarrow$ {large1 itemsets}

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \{c = a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a, \{S \subseteq C \mid |S| = k\ \ 1\} \subseteq L_{k-1}\}$

For transactions $t \in T$

$D_t \leftarrow \{c \in C_k \mid c \subseteq t\}$

For candidates $c \subseteq D_t$

count $|c| \leftarrow$ count $[c] + 1$

$L_k \leftarrow \{c \in C_k \mid$ count $[c] >= \varepsilon\}$

$k \leftarrow k+1$

return $\cup L_k$

Επεξηγηματικά ο αλγόριθμος Apriori ακολουθεί τα εξής βήματα, με το ελάχιστο όριο υποστήριξης να δίνεται στο πρόβλημα ή να ο καθορίζει ο χρήσης :

1. Στην πρώτη επανάληψη, κάθε στοιχείο αντιμετωπίζεται ως ένας υποψήφιος για σετ με μοναδικό αντικείμενο. Ο αλγόριθμος υπολογίζει τις εμφανίσεις κάθε στοιχείου.
2. Με προκαθορισμένο το ελάχιστο όριο υποστήριξης (min sup) , ελέγχονται όλα τα σετ με ένα αντικείμενο αν πληρούν το κριτήριο min sup και απορρίπτονται όσα δεν το ικανοποιούν. Μόνο οι υποψήφιοι που είχαν εμπιστοσύνη ίση ή μεγαλύτερη από το όριο, περνάνε στην επόμενη επανάληψή.
3. Στην δεύτερη επανάληψη , ανακαλύπτονται τα συχνά στοιχειοσύνολα με κατανομή σετ των δύο αντικειμένων, κάθε σετ δημιουργείται με την σύνδεση όλων των αντικείμενων ανα δύο μεταξύ τους.
4. Και πάλι γίνεται έλεγχος αν οι υποψήφιοι των σετ με 2 στοιχεία ικανοποιούν το min sup , διαφορετικά απορρίπτονται
5. Στην επομένη επανάληψη θα δημιουργηθούν ομάδες με 3 στοιχεία, όπου αποτελούνται από ένα συχνό στοιχειοσύνολο με 2 στοιχεία και ένα συχνό στοιχειοσύνολο με 1 στοιχείο διαφορετικό από τα άλλα 2. Εντοπίζονται τα μη συχνά και δεν συνεχίζουν στο επόμενο βήμα. Η σκέψη που ακολουθεί ο

αλγόριθμος είναι ότι εάν όλα τ υποσύνολα με δύο στοιχεία είναι συχνά τότε και τα υπερσύνολα θα είναι συχνά, αλλιώς απομακρύνονται. Αντίθετα αν δεν ήταν συχνό ένα σετ τότε και το υπερσύνολο του δεν θα μπορούσε να είναι συχνό.

6. Το επόμενο στάδιο ξεκινάει με την εύρεση των σετ με 4 στοιχεία κ.ο.κ. Ο αλγόριθμος διακόπτεται όταν επιτυγχάνεται το πιο συχνό σύνολο αντικειμένων.

3.2 ΥΛΟΠΟΙΗΣΗ

Python

Η υλοποίηση της έρευνας, που θα αναπτυχθεί στην εργασία αυτή, θα γίνει με ένα κομμάτι κώδικα με γλώσσα προγραμματισμού την Python. Η Python είναι μια ερμηνευμένη, αντικειμενοστρεφής, γλώσσα προγραμματισμού υψηλού επιπέδου με δυναμική σημασιολογία, όπως αναφέρεται και στην επίσημη ιστοσελίδα της γλώσσας (www.python.org). Κυκλοφόρησε για πρώτη φορά στο κοινό το 1991 και ο ιδρυτής της είναι ο Guido van Rossum. Τον Δεκέμβριο του 2020 ανακηρύχθηκε ως μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού στην τρίτη θέση (Δείκτης TIOBE, 2020). Μια μεγάλη τυπική βιβλιοθήκη συνοδεύει την γλώσσα, όπου βρίσκει εφαρμογή σε αρκετούς τομείς εργασιών, αλλά μπορεί να γίνει επεκτάσιμη μέσω άλλων χιλιάδων βιβλιοθηκών που μπορούν οι χρήστες να κατεβάσουν μέσα από το αποθετήριο (Python Package Index - PyPI) για λογισμικό τρίτων, το οποίο παρέχει πάνω από 200.000 πακέτα και βρίσκεται σε λειτουργία από τον Νοέμβριο του 2019 (Debill, Erik, 2019).

Η Python αποτελεί ένα μεγάλο εργαλείο για την εξόρυξη γνώσης και είναι ευρέως διαδεδομένη και συνυφασμένη με την επιστήμη των δεδομένων. Οι ερευνητές έχουν στην διάθεση τους, πληθώρα βιβλιοθηκών για αρκετές υλοποιήσεις και επεξεργασίες δεδομένων. Προσεγγίζει με ευκολία και ταχύτητα την ανακάλυψη μοτίβων και συσχετίσεων με βάση την αξιολόγηση της απόδοσης, καθώς και την μείωση των δεδομένων, δύο προβλήματα που είναι καθοριστικά στην υλοποίηση της ανάλυσης συσχετίσεων σε κειμενικά δεδομένα, το θέμα δηλαδή που πραγματεύεται η υλοποίηση που θα πραγματοποιηθεί. Μεγάλος αριθμός, επίσης γνωστών εταιριών μεταχειρίζεται την Python όπως :

- Google (Norvig,2012)

- Instagram (Instagram Engineering, 2019)
- Facebook (Recordon, 2009)
- Spotify (Van der Meer, 2018)
- YouTube (Do, 2019)
- Nasa (Shafer, Daniel, 2003)
- CERN (CERN Publication, 2006)
- Yahoo (Hassan, 2019)
- Amazon (Costa, 2020)

Δεδομένα

Τα δεδομένα που θα χρησιμοποιηθούν είναι αρχεία της μορφής csv. Το σύνολο των αρχείων που θα γίνει υλοποίηση είναι έξι και έχουν προέλευση από τον ιστότοπο "Kaggle", ο οποίος περιέχει αρκετά σύνολα δεδομένων για χρήση στην επιστήμη της εξόρυξης γνώσεων και απόψεων. Όλα τα σύνολα περιέχουν tweets στην αγγλική γλώσσα.

Το πρώτο αρχείο είναι το "emotion stimulus" , το οποίο θα αναφέρεται ως D1 αρχείο. Περιέχει 2.415 εγγραφές και τα κείμενα των tweets είναι κατηγοριοποιηθεί σε συναισθήματα. Τα συναισθήματα που περιέχει είναι :

1. Happy
2. Sad
3. Anger
4. Fear
5. Shame

Το δεύτερο αρχείο ονομάζεται "text emotion" και είναι το μεγαλύτερο σύνολο δεδομένων, αφού περιέχει 39.812 tweets. Θα αναφέρεται στην παρούσα εργασία ως D2. Τα tweets έχουν ανατεθεί σε :

1. Sad
2. Worry
3. Happy

4. Surprise
5. Love
6. Relief
7. Fun

Το τρίτο σύνολο δεδομένων είναι το “customer support”, με 19.633 εγγραφές. Θα χαρακτηρίζεται ως D3. Τα συναισθήματα που κατανέμει τα κειμενικά δεδομένα είναι:

1. Fear
2. Anger
3. Surprise
4. Love
5. Joy
6. Sad

Τα επόμενα τρία σύνολα δεδομένων ονομάζονται “Coachella”, “extra data” και “twitter argument”, με D4, D5, D6 χαρακτηρισμό αντίστοιχα. Τα tweet που περιέχουν κατανέμονται με βάση την πολικότητα τους, δηλαδή είτε θετικά είτε αρνητικά(Positive/Negative). Το πλήθος των εγγραφών τους είναι 4.173, 12.502, 27.465 αντίστοιχα.

Ένα παράδειγμα για τα σύνολα δεδομένων που παρουσιάζονται στα αρχεία CSV είναι:

[happy, "I 'm really cheerful in the GROUP NOW!”]

Στην συνέχεια, μόλις τα δεδομένα εισαχθούν στον κώδικα, πρέπει να παρουσιαστούν με την κατάλληλη μορφή, για το λόγω αυτό έγινε χρήση της βιβλιοθήκης Pandas από την Python. Λειτουργεί με δεδομένα, δημιουργώντας μια μορφή πίνακα (DataFrame), έτσι ώστε να γίνεται η ανάγνωση και η εγγραφή τους γρήγορα και αποδοτικά. Αναδιαμόρφωση, ενσωματωμένη ευρετηρίαση, τεμαχισμός βάσει ετικετών, συγχώνευση για τα δεδομένα, είναι μερικές λειτουργίες που προσφέρονται στο πακέτο Pandas.

Στον κώδικα που δημιουργήθηκε, γίνεται ανάγνωση και ενσωμάτωση των tweets σε ένα DataFrame με την μορφή πίνακα, μέσω της εντολής read_csv . Στην πρώτη στήλη του πίνακα εμφανίζεται το συναίσθημα που κατηγοριοποιείται το κείμενο του tweet. Στην δεύτερη στήλη, υπάρχουν όλα τα tweets με τον κείμενο τους σε φυσική γλώσσα. Ο πίνακας έχει τόσες γραμμές όσες και τα tweets, κάθε tweet αντιστοιχεί και σε μία γραμμή. Σε περίπτωση που αρχείο δεδομένων περιέχει παραπάνω από τις δύο στήλες που περιεγράφηκαν παραπάνω, τότε οι στήλες αυτές αφαιρούνται με την εντολή drop_columns.

Προεπεξεργασία Δεδομένων

Τα κειμενικά δεδομένα, που πλέον βρίσκονται στην δεύτερη στήλη, είναι ένα συνεχόμενο κείμενο από λέξεις, όπως γράφτηκε από τους χρήστες του Twitter . Με την μορφή αυτή δεν είναι εύκολη η ανάλυση των πληροφοριών τους και γι' αυτό είναι απαραίτητη η επεξεργασία του κειμένου.
Θα ακολουθήσουν τα βήματα που έγιναν στο στάδιο της επεξεργασίας κειμένου :

1. Διαίρεση σε όρους (Tokenization)

   Η Διαίρεση σε όρους είναι ο κατακερματισμός μιας συμβολοσειράς σε λέξεις. Οι όροι μπορούν να είναι οποιοσδήποτε χαρακτήρας, όπως λέξη , αριθμός, ολόκληρη πρόταση, σημεία στίξης ή σύμβολα (Manning, Raghavan, Schütze, 2008). Στον συγκεκριμένο κώδικα τα Tweets, που αποτελούν το κείμενο, κόπηκαν σε λέξεις. Πλέον τα δεδομένα παρουσιάζονται σε λίστες, κάθε λίστα είναι ένα tweet το οποίο περιέχει τις λέξεις του tweet με κόμμα ανάμεσα τους. Το στάδιο της Λημματοποίησης υλοποιήθηκε με την βοήθεια της βιβλιοθήκης Python " NTLK", η οποία είναι ένα εργαλείο για την επεξεργασία φυσικής γλώσσας και εκτελέστηκε η μέθοδος *.tokenize()*.

   Π.χ. [happy, ",I, ',m, really, cheerful, in, the, GROUPS, NOW, !, ",]

2. Μετατροπή σε πεζούς χαρακτήρες

   Κάποιοι χαρακτήρες ή και λέξεις μπορεί να ήταν με κεφαλαία γράμματα, για να επιτευχθεί η συναισθηματική ανάλυση των λέξεων, μετατράπηκαν όλοι οι

χαρακτήρες σε πεζά γράμματα και με λατινικούς χαρακτήρες. Η μετατροπή ολοκληρώθηκε με την μέθοδο συμβολοσειράς της python *.lower()*.

Π.χ. [happy, ",I, ',m, really, cheerful, in, the, groups, now, !, ”,]

3. Αφαίρεση αριθμών και συμβόλων

Μέσα στα tweets κάποιοι χρήστες είναι επιτρεπτό να κάνουν χρήση αριθμών ή άλλων συμβόλων. Τα δεδομένα αυτά δεν ήταν χρήσιμα στην συγκεκριμένη έρευνα και δεν παρείχαν κάποια πληροφορία για τα συναισθήματα που δημιουργούσαν τα κείμενα και αφαιρέθηκαν. Η ανίχνευση και αφαίρεση των συμβόλων έγινε με την χρήση της μεθόδου *.isalpha()*, που περιλαμβάνει η γλώσσα προγραμματισμού.

Π.χ. [happy, I, m, really, cheerful, in, the, groups, now]

4. Αφαίρεση των Stopwords

Ένας όρος με μεγάλη συχνότητα εμφάνισης, που όμως δεν προσδίδει κάποια ιδιαίτερη σημασία στο κείμενο , θεωρείται stopword. Ένα stopword, μπορεί να είναι ένα άρθρο, ένας σύνδεσμος ή πρόθεση. Μπορούν να επιφέρουν μείωση της απόδοσης του αλγορίθμου σε περίπτωση που παραμείνουν στο ευρετήριο των λέξεων, λειτουργώντας ως θόρυβος. Με την αφαίρεση, αυτών των συχνών όρων, θα επέλθει και μείωση τω δεδομένων (Manning, Raghavan, Schütze, 2008). Η εντολή που χρησιμοποιήθηκε για την αφαίρεση των stopwords στον κώδικα της Python είναι *stopwords.words('english')*, που παρέχεται από την βιβλιοθήκη NLTK, στην γλώσσα των αγγλικών που βρίσκονται τα tweets που ελέγχουμε.

Π.χ. [happy, really, cheerful, groups, now]

5. Λημματοποίηση (Lemmatization)

Η Λημματοποίηση αποτελεί μια διαδικασία που ομαδοποιεί τα παράγωγα και τις καταλήξεις ενός μέρος του λόγου, ανάγοντας τους όρους στην αρχική ρίζα-λήμματα των λέξεων. Αυτό συμβαίνει γιατί μπορεί κάποια λέξη με ίδιο

νοηματικό περιεχόμενο να εμφανίζεται με πολλές διαφορετικές παράγωγες λέξεις και να μην μπορεί να γίνει αντιστοίχιση μεταξύ τους (Jaaben, 2018). Μέσα από την βιβλιοθήκη NLTK, και πάλι, ομαδοποιούμε όλα τα λήμματα σε μία λίστα με την εντολή *lemma = WordNetLemmatizer()* με βάση δεδομένων το λεξικό WordNet για να μπορεί να αναγνωρίσει την ρίζα των λέξεων. Στην συνέχεια από τις λέξεις που έχουμε στα δεδομένα αντιστοιχίζονται, ελέγχοντας για τις καταλήξεις, “n”, “s”, ως οι πιο συχνές στην αγγλική γλώσσα, με τις εντολές

lemma.lemmatize(word, pos = "n")

lemma.lemmatize(word, pos = "s")

Π.χ. [happy, really, cheerful, group, now]

6. Αποκατάληξη (Stemming)

Η Αποκατάληξη έχει ως στόχο να αναγνωρίσει τις ρίζες των λέξεων, ανεξάρτητα από την κλήση ή την δημιουργία παραγώγων, αφαιρώντας τις καταλήξεις από τις λέξεις. Με την αναγωγή τους σε μια κοινή ρίζα, μειώνονται τα δεδομένα και επέρχεται μεγαλύτερη απόδοση (Jaaben, 2018). Έχουν αναπτυχθεί αρκετοί αλγόριθμοι για την διαδικασία αυτή, ένας από τους πιο γνωστούς είναι ο Porter και ο Snowball, που χρησιμοποιήθηκαν και στην επεξεργασία των tweets. Ο Porter είναι ένας από τους παλαιότερους αλγορίθμους για την Αποκατάληξη και στηρίζεται στην ιδέα ότι τα επίθετα στην αγγλική γλώσσα παρουσιάζουν μικρές καταλήξεις π.χ. -ly, -e, -ed κλπ. Ο αλγόριθμος Snowball μπορεί να χαρτογραφήσει σε πολλές γλώσσες, είναι αρκετά πιο “επιθετικός” σαν προσέγγιση και κόβει καταλήψεις σε επίθετα και ρήματα π.χ. troubling και troubled γίνονται troubl, το οποίο πολλές φορές μπορεί να μην αντιστοιχεί σε λέξη αλλά μόνο στην ρίζα. Αφού εισαχθούν από την NLTK οι δύο αλγόριθμοι , στην συνέχεια δίνεται η εντολή να υλοποιηθεί στα δεδομένα η Αποκατάληξη με *PorterStemmer.stem(word)* και *SnowballStemmer.stem(word).*

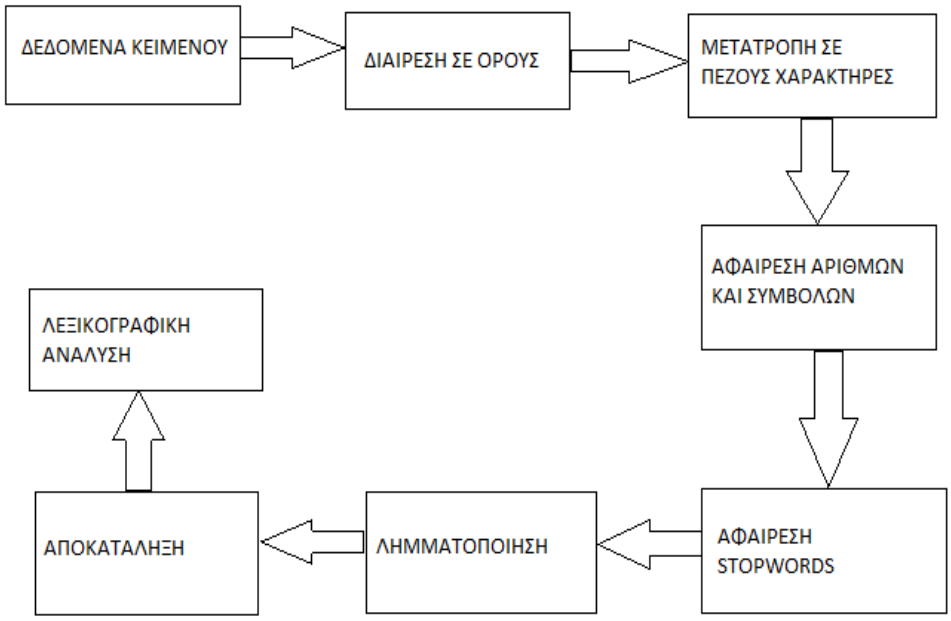
Π.χ. [happy, real, cheerful, groups, now]

7. Λεξικογραφική Ανάλυση (Pos Tagging)

Η λεξικογραφική ανάλυση αναγνωρίζει και κατηγοριοποιεί σε τι μέρος του λόγου ανήκει η κάθε λέξη από το ευρετήριο των δεδομένων που παρέχεται, δηλαδή αν είναι ουσιαστικό, ρήμα, σύνδεσμος, επίθετα κ.λπ.. Με αυτήν την κατηγοριοποίηση των λέξεων, κάθε ερευνητής μπορεί να αποκόψει και να χρησιμοποιήσει μόνο τα μέρη του λόγου που είναι πιο χρήσιμα στα νοηματικό περιεχόμενο π.χ. τα ουσιαστικά. Με την μέθοδο *pos_tag(word)* της NLTK, ολοκληρώνεται η ομαδοποίηση στα μέρη του λόγου και με την *.bigrams(postag)* γίνεται η επιλογή μεμονωμένων ομάδων.

Π.χ. [(happy, Adjective), (real, Adjective), (cheerful, Adjective), (groups, Noun), (now, Adverb)]

Η παραπάνω προεπεξεργασία κειμένου απεικονίζεται στην Εικόνα 3.1:



Εικόνα 3.1: Προεπεξεργασία κειμένου

Η σειρά που θα ολοκληρωθεί η προεπεξεργασία δεν είναι δεδομένη, διότι θα πραγματοποιηθούν πολλά πειράματα για την εύρεση της καλύτερης εμπιστοσύνης και

υποστήριξης των κανόνων συσχέτισης με διαφορετικές διατάξεις ή και ελλείψεις σταδίων.

Αλγόριθμος Apriori και Κανόνες Συσχέτισης

Μετά την προεπεξεργασία των δεδομένων, θα γίνει εισαγωγή των λέξεων κάθε tweet σε μορφή λίστας στον αλγόριθμο Apriori,έτσι ώστε να αντληθούν οι κανόνες συσχέτισης. Τα ελάχιστα όρια της εμπιστοσύνης και υποστήριξης ορίζονται από τον χρήστη στην εντολή της apriori ως min_support και min_confidence . Η εκτέλεση του αλγορίθμου στην γλώσσα προγραμματισμού Python θα γίνει με την βοήθεια της βιβλιοθήκης «efficient Apriori». Οι εντολές που πρέπει να εισαχθούν στον κώδικα είναι:

from efficient_apriori import apriori
transactions = list_of_tweets
rules = apriori(transactions, min_support=1, min_confidence=1)
print(rules)

Με αυτόν τον τρόπο, εκτυπώνονται στην κονσόλα των αποτελεσμάτων όσοι κανόνες υπακούν στα ελάχιστα όρια που έχει θέσει ο χρήστης. Όμως μπορεί να είναι οποιοσδήποτε κανόνας, που να μην ικανοποιεί την συγκεκριμένη έρευνα. Το επόμενο βήμα που χρειάζεται είναι να γίνει διαλογή των κανόνων που περιέχουν στο δεξιό μέρος κάποιο από τα συναισθήματα που έχουν κατηγοριοποιηθεί τα tweets στα δεδομένα. Αυτό γίνεται με την εντολή:

rules_rhs = filter(lambda rule: rule.rhs == "sentiment", rules)

RapidMiner

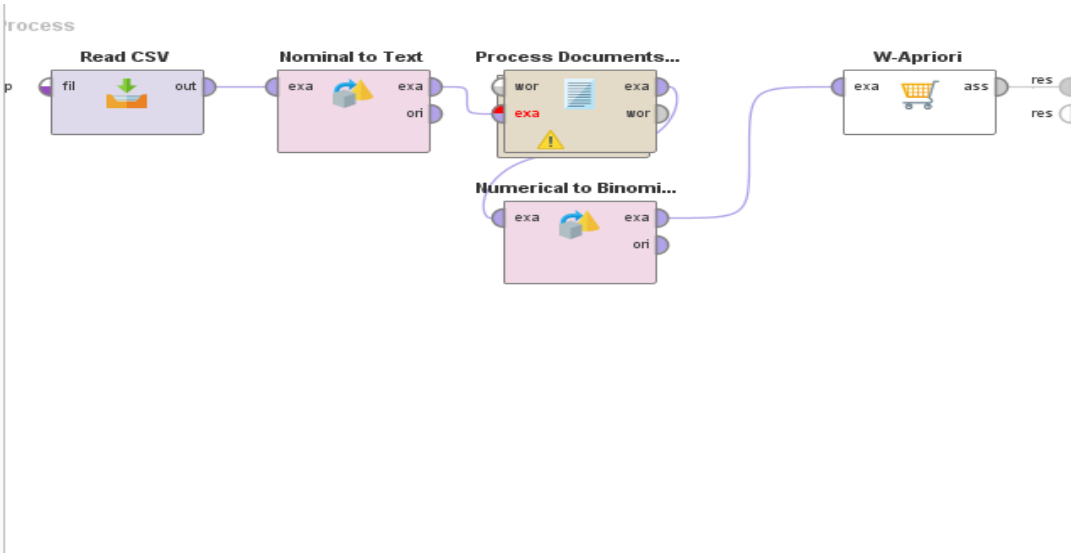
Εκτός από κώδικα στην Python, υλοποίηση θα γίνει και σε ένα πρόγραμμα ανοιχτού κώδικα, το RapidMiner. Το RapidMiner είναι ένα δωρεάν πρόγραμμα, το οποίο χρησιμοποιείται για την ανάλυση και την επεξεργασία δεδομένων, κυρίων κειμενικών

δεδομένων. Διατίθεται ως αυτόνομη εφαρμογή και περιέχει πάνω από 1500 τελεστές για την μετατροπή και την ανάλυση των δεδομένων. Κάθε τελεστής έχει μια συγκεκριμένη εργασία και η σειρά που προστίθενται στην υλοποίηση δεν είναι ανεξάρτητη του αποτελέσματος. Υποστηρίζει αρκετούς τύπους αρχείων και παρουσιάζει τα αποτελέσματα σε διάφορες μορφές π.χ. πίνακες και διαγράμματα. (Dursun, 2012)

Στην εργασία αυτή ακολουθήθηκε η ίδια μορφή υλοποίησης που έγινε και με την Python. Οι τελεστές που χρησιμοποιήθηκαν είναι :

- Η ανάλυση ξεκινάει εισάγοντας τα αρχεία με τα δεδομένα που έχουν αναφερθεί παραπάνω. Ο τελεστής που διαβάζει τα αρχεία τύπου CSV είναι ο “read CSV”, ο οποίος βρίσκεται στον φάκελο “Data access” των τελεστών και διαβάζει τα δεδομένα του αρχείου και τα χωρίζει σε στήλες ανάλογα με τον χαρακτήρα που υπάρχει στις γραμμές για την διαλογή. Εδώ επιλέγεται στις παραμέτρους να εισάγει τις στήλες με τον χαρακτήρα του κόμματος και να τις αντιστοιχίσει στις μεταβλητές αρχείου που είναι η επιλογή “text” στο “data set metadata information”.
- Στην συνέχεια συνδέεται ο τελεστής “Nominal to text”, αλλάζοντας τον τύπο των επιλεγμένων ονομαστικών τιμών σε κείμενο και χαρτογραφεί τις τιμές στις αντίστοιχες τιμές των συμβολοσειρών. Αν υπάρχουν στο αρχείο ονομαστικές τύπου στήλες, δηλαδή συμβολοσειρές ή κατηγορηματικές τιμές, τότε όλα μετατρέπονται σε τιμές κειμένου. Σε περίπτωση που δεν υπάρχουν, δεν αλλάζει τα δεδομένα.
- Ακολουθεί ο τελεστής “Process Documents from Data”, είναι το modulo που μέσα σε αυτό θα γίνει όλη η προεπεξεργασία των λέξεων. Επιλέγεται το “create word vector”, “keep text” και στο prune η επιλογή “absolute”.
- Έπειτα είναι ο “Numerical to binominal”, ο οποίος μετατρέπει τα δεδομένα από αριθμητικά σε διωνυμικά τύπου δεδομένα. Με αυτόν τον τρόπο εμφανίζει σε έναν πίνακα αν βρίσκεται η λέξη στο συγκεκριμένο tweet, αν υπάρχει η λέξη τότε μπαίνει η τιμή true, αν όχι τότε μπαίνει η τιμή false.
- Τέλος εισάγεται ο τελεστής “w-Apriori”, που μετράει το support και confidence κάθε στοιχειοσυνόλου και παρουσιάζει στα αποτελέσματα τους κανόνες

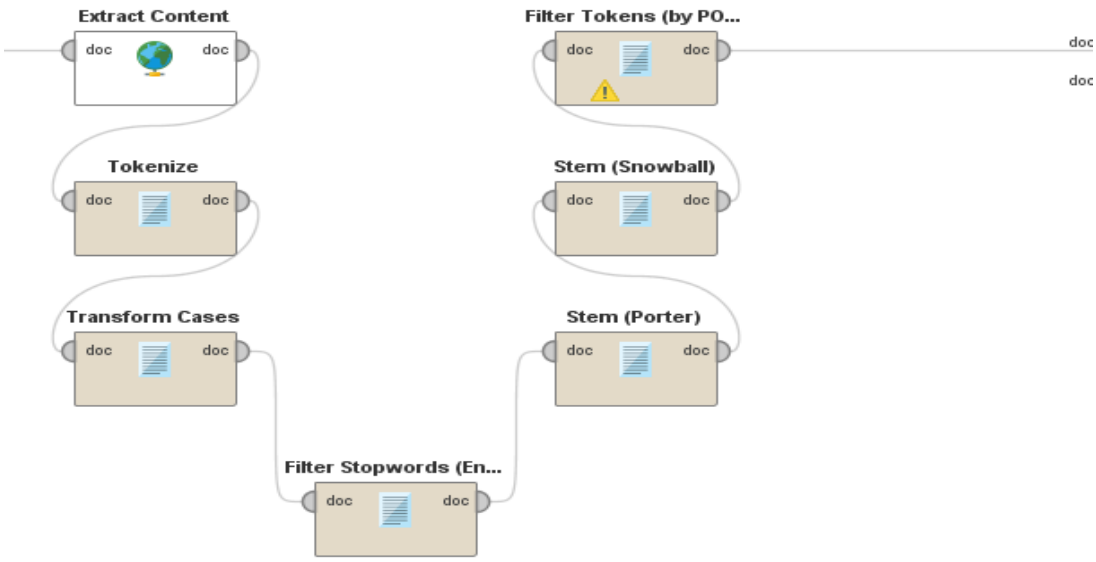
εξόρυξης. Στην παράμετρο “C” δηλώνει ο χρήστης την ελάχιστη εμπιστοσύνη που επιθυμεί να έχει ο κανόνας και στην “D” την ελάχιστη υποστήριξη.



Εικόνα 3.2 Ανάλυση της υλοποίησης σε RapidMiner

Μέσα στην τελεστή “Process Documents from Data” πραγματοποιείται η προεπεξεργασία των δεδομένων κειμένων, ως εξής :

- Πρώτος μπαίνει ο “extract content”, ο οποίος εξάγει τα δεδομένα κειμένου έτσι ώστε να τα επεξεργαστεί ο χρήστης.
- Ακολουθεί ο “Tokenize”, που κόβει τα tweets σε λέξεις
- Ο “Transform cases” μετατρέποντας όλους τους χαρακτήρες σε πεζά γράμματα.
- Στην συνέχεια είναι ο τελεστής “Stopwords” σε γλώσσα αγγλικά, αφού στην αγγλική γλώσσα είναι γραμμένα τα tweets.
- Ο “Stem” με Porter και Snowball, πραγματοποιώντας την Αποκατάληξη των όρων.
- Και τέλος ο τελεστής “Filter tokens(by POS Tags)” με την λεξικογραφική ανάλυση.



Εικόνα 3.3 Προεπεξεργασία δεδομένων στο RapidMiner

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Εκτελέστηκαν οι υλοποιήσεις με τον κώδικα στην γλώσσα προγραμματισμού Python και η υλοποίηση στο πρόγραμμα RapidMiner για κάθε ένα από τα έξι αρχεία. Σε καθένα από τα σύνολα δεδομένων εκτελέστηκαν οι υλοποιήσεις με πολλές διαφορετικές τόσο στην σειρά αλλα και στην ύπαρξη τεχνικών και τελεστών προεπεξεργασίας αντίστοιχα, στα πλαίσια του πειράματος δημιουργίας κανόνων εξόρυξης με τα καλύτερα δυνατά αποτελέσματα στην εμπιστοσύνη και την υποστήριξη.

Τα αποτελέσματα που προέκυψαν από τις δύο υλοποιήσεις, Python και RapidMiner, ήταν ίδια τόσο στο πλήθος των κανόνων, όσο και στους κανόνες που αντλήθηκαν τα ελάχιστα όρια της υποστήριξης και της εμπιστοσύνης. Διαφορές απειροελάχιστες μπορεί να εμφανίστηκαν, αλλα αυτό οφείλεται από την γλωσσολογική επεξεργασία που υπέστησαν τα σύνολα δεδομένων.

Οι τεχνικές προεπεξεργασίας για χάρη συντομίας θα αναφέρονται ως:
TL : διαίρεση του κειμένου σε όρους και μετατροπή των χαρακτήρων σε πεζούς
TLS: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς και αφαίρεση των Stopwords
TLSSTEM: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και αποκατάληξη
TLSL: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και λημματοποίηση
TLSPT: διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζούς, αφαίρεση των Stopwords και λεξικογραφική ανάλυση
Οποιαδήποτε άλλη τεχνική θα αναφέρεται αναλυτικά

Το πλήθος των συχνών στοιχειοσυνόλων με την TLS στα D1-D6 σύνολα δεδομένων με διάφορα όρια είναι :

	D1	D2	D3
min_sup=20% min_conf=20%	401	642	457
min_sup=30% min_conf=30%	119	67	182
min_sup=30% min_conf=37%	97	53	113
min_sup=30% min_conf=40%	82	22	102
min_sup=30% min_conf=45%	51	9	84

Πίνακας 4.1 Πλήθος συχνών στοιχειοσυνόλων των D1-D3 με διάφορα ελάχιστα όρια

	D4	D5	D6
min_sup=30% min_conf=30%	556	811	749
min_sup=69% min_conf=69%	267	313	283
min_sup=80% min_conf=86%	89	115	102
min_sup=82% min_conf=89%	74	96	85

Πίνακας 4.2 Πλήθος συχνών στοιχειοσυνόλων των D4-D6 με διάφορα ελάχιστα όρια

Στον Πίνακα 4.1 αποφαίνεται ότι για το σύνολο δεδομένων D1 έχουμε επαρκή αριθμό σε συχνά στοιχειοσύνολα με 30% η υποστήριξη και στο 40% η εμπιστοσύνη, με επιδίωξη τα συχνά στοιχειοσύνολα να βρίσκονται στο εύρος 50-100. Στο D2 παρουσιάζονται πολύ λιγότερα σε αριθμό συχνά στοιχειοσύνολα από τα D1 και D3, εξαιρούμενο από τα πολύ χαμηλά ελάχιστα όρια του 20% στην υποστήριξη και εμπιστοσύνη. Αυτό το αποτέλεσμα δεν κρίνεται αναμενόμενο, διότι το αρχείο D2 περιέχει τον μεγαλύτερο πλήθος tweet , που ανέρχεται στο 39.812 κειμενικά δεδομένα.

Το όριο που φαίνεται να αποδίδει για τον ζητούμενο αριθμό στο D2 είναι το 30% υποστήριξη και 37% εμπιστοσύνη με 53 συχνά στοιχειοσύνολα. Ενώ στο σύνολο δεδομένων D3 υπάρχει σε όλα τα όρια μεγαλύτερο πλήθος συχνών στοιχειοσυνόλων από τα D1,D2 και το επιθυμητό όριο είναι 30%/45% υποστήριξης και εμπιστοσύνης αντίστοιχα. Το όριο με τα καλύτερα αποτελέσματα ως προς το πλήθος των συχνών στοιχειοσυνόλων στα αρχεία D1,D2 και D3 είναι το 30%/37% υποστήριξη και εμπιστοσύνη.

Στον Πίνακα 4.2 παρουσιάζονται τα πλήθη των συχνών στοιχειοσυνόλων για τα σύνολα D4, D5 και D6, τα οποία ταξινομούν τα κείμενα ανάλογα με την πολικότητα τους σε αρνητικά και θετικά, με μεγαλύτερα ποσοστά στα ελάχιστα όρια. Στο D4 με 80% υποστήριξη και 86% εμπιστοσύνη αποδίδονται 89 συχνά στοιχειοσύνολα και με 82% υποστήριξη και 89% εμπιστοσύνη αποδίδονται 74%. Στα αρχεία D5 και D6 έχουν 96 και 85 αντίστοιχα πλήθος συχνών στοιχειοσυνόλων στο όριο 82%/89% υποστήριξη και εμπιστοσύνη, κάνοντας το όριο αυτό το πιο αποδοτικό για το επιθυμητό πλήθος και για τα 3 σύνολα δεδομένων.

Το πλήθος των κανόνων στην υλοποίηση με επεξεργασία μόνο με διαίρεση του κειμένου σε όρους, ήταν αρκετά μεγάλο, λόγω ότι χρησιμοποιήθηκαν για τους κανόνες συσχέτισης λέξεις χωρίς ιδιαίτερο ενδιαφέρον, όπως είναι τα άρθρα και οι αντωνυμίες. Με την TLS, προέκυψαν τα εξής αποτελέσματα :

	D1	D2	D3	D4	D5	D6
min_sup=10% min_conf=10%	538	11.745	940	584	449	1.288
min_sup=20% min_conf=20%	199	385	185	370	403	519
min_sup=30% min_conf=30%	46	33	72	209	331	298
min_sup=30% min_conf=37%	42	14	44	188	239	221
min_sup=30%	35	8	33	176	204	195

min_conf=40%						
min_sup=30% min_conf=45%	9	3	19	142	180	157
min_sup=69% min_conf=69%	1	0	0	102	147	122
min_sup=80% min_conf=86%	0	0	0	67	79	71
min_sup=85% min_conf=85%	0	0	0	44	68	51
min_sup=82% min_conf=89%	0	0	0	31	49	43

Πίνακας 4.3 Πλήθος κανόνων συσχέτισης με διάφορα ελάχιστα όρια

Όπως φαίνεται από τον Πίνακα 4.3 στα αρχεία D1 ως D3, τα οποία κατηγοριοποιήθηκαν τα αντίστοιχα tweets τους σε 5 συναισθήματα και πάνω προκύπτουν κανόνες με μικρότερη υποστήριξη. Αυτό οφείλεται κυρίως στον διαχωρισμό των λέξεων σε περισσότερες κατηγορίες, όπου κάθε κατηγορία παίρνει και ένα μερίδιο του ποσοστού εμφάνισης. Είναι δύσκολο μία λέξη να δηλώνει ένα ξεκάθαρο συναίσθημα, διότι μπορεί να εμφανίζεται σε διάφορα είτε θετικά είτε αρνητικά συναισθήματα. Στην αντίθετη περίπτωση, όμως, βρίσκονται τα αποτελέσματα των αρχείων D4 ως D6, τα οποία κατηγοριοποιήθηκαν με βάση την πολικότητα των tweets. Εδώ ο διαχωρισμός είναι πιο εμφανής, δεδομένου ότι υπάρχουν δύο μόνο επιλογές και με αυτόν τον τρόπο επιτυγχάνονται καλύτερα αποτελέσματα υποστήριξης και εμπιστοσύνης στους κανόνες συσχέτισης που δημιουργήθηκαν, χωρίς όμως ένα μπορεί να γίνει διευκρίνηση του εκάστοτε συναισθήματος που προκαλεί το tweet.

Συγκεκριμένα το αρχείο D1 είναι το μόνο σύνολο δεδομένων που περιέχει έναν κανόνα στο όριο υποστήριξης και εμπιστοσύνης 69%/69% από τα D1-D3, καθώς και στους κανόνες με όριο μέχρι 30% υποστήριξη και μέχρι 45% εμπιστοσύνη εμφανίζει αποδοτικά αποτελέσματα στο πλήθος των κανόνων συσχέτισης, αν υπολογιστεί ότι είναι το μικρότερο σύνολο δεδομένων με μόλις 2.415 tweets. Στην αντίθετη περίπτωση βρίσκεται το αρχείο D2, το οποίο ενώ περιέχει τον μεγαλύτερο αριθμό κειμενικών

δεδομένων δεν παρουσιάζει καλά αποτελέσματα σχεδόν σε όλους τους κανόνες, με εξαίρεση τα όρια μέχρι 20% στην υποστήριξη και εμπιστοσύνη, όπου βρίσκεται το μεγαλύτερο μέρος των κανόνων συσχέτισης του αρχείου αυτού. Το D3 έχει αποτελέσματα με μεγάλη σταθερότητα στο πλήθος των κανόνων μέχρι τα όρια 30% υποστήριξης και 45% εμπιστοσύνης, με περιεχόμενα 19.633 κείμενα από το Twitter. Το επιθυμητό όριο, που περιέχει επαρκή πλήθος των κανόνων συσχέτισης στα D1-D3, για την υποστήριξη των πειραμάτων, είναι το 30%/37% υποστήριξη/εμπιστοσύνη. Στα αρχεία D4-D6 κάθε αρχείο έχει πολύ καλά σταθερά και λογικά αποτελέσματα για το πλήθος των κανόνων συσχέτισης που παρουσιάζουν με καλύτερο όριο το 85%/85% στην υποστήριξη και εμπιστοσύνη με ένα πλήθος περίπου στους 50 κανόνες για το κάθε σύνολο.

Παρακάτω θα υπολογιστούν σύμφωνα με τα καλύτερα όρια από τους κανόνες συσχέτισης στα αρχεία , το πλήθος των κανόνων συσχέτισης ανάμεσα στις υλοποίησης με διαφορετικές τεχνικές προεπεξεργασίας. Για τα αρχεία D1-D3 θα υπολογιστούν με όριο 30% υποστήριξη και 37% εμπιστοσύνη και για τα D4-D6 με 85% υποστήριξη και εμπιστοσύνη :

	D1	D2	D3
TL	89	71	96
TLS	42	14	44
TLSS	3	4	11
TLSL	0	2	8
TLSP(NOUN)	0	2	17
TLSP(ADJECTIVE)	0	1	15
TLSP(VERB)	0	0	10

Πίνακας 4.4 Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D1-D3

Από τον Πίνακα 4.4, το αρχείο D1 που δημιουργούνται 89 κανόνες στην υλοποίηση TLκαι 42 κανόνες στην TLS, σχεδόν σε όλες τις υπόλοιπες τα πλήθος των κοινών κανόνων είναι μηδενικό, αυτό συμβαίνει διότι τα αποτελέσματα στις υπόλοιπες τεχνικές πέφτουν και αδυνατούν να δημιουργήσουν τόσο ισχυρούς κανόνες ή και αν χρησιμοποιήσουν δεν είναι κοινοί με αυτούς στις άλλες δύο. Το ίδιο συμβαίνει και στο

σύνολο D2 με 1-2 κοινούς κανόνες στις υλοποιήσεις πέραν των TL και TLS. Στο D3 όμως εμφανίζονται διαφορές, γιατί αντλούνται επαρκής αριθμός ισχυρών κανόνων σε όλες τις υλοποιήσεις και ιδιαίτερα στην λεξικογραφική ανάλυση ουσιαστικών, επιθέτων και ρημάτων, που τα προηγούμενα δύο αρχεία δεν επέδειξαν.

	D4	D5	D6
TL	105	122	114
TLS	44	68	51
TLSS	1	33	1
TLSL	0	0	0
TLSP(NOUN)	3	2	18
TLSP(ADJECTIVE)	3	3	6
TLSP(VERB)	12	2	3

Πίνακας 4.5 Πλήθος κανόνων συσχέτισης με διάφορες τεχνικές προεπεξεργασίας στα D4-D6

Στον Πίνακα 4.5, το αρχείο D4 έχει 12 κανόνες συσχέτισης κοινούς στην τεχνική TLSP(VERB) με τους 44 της TLS και τους 105 της TL, ενώ στις υπόλοιπες υλοποιήσεις τα αποτελέσματα κυμαίνονται από 0 έως 3. Το σύνολο δεδομένων D2 εμφανίζει 33 κοινούς κανόνες συσχέτισης με TLSS, σε αντίθεση με τα αρχεία D1 και D3. Το D3 έχει έναν αυξημένο αριθμό στην υλοποίηση TLSP(NOUN) με 18 κανόνες κοινούς και 6 στην TLSP(ADJECTIVE).

Από τους Πίνακες 4.4 και 4.5 φαίνεται ότι η υλοποίηση TLS αποδίδει καλύτερα και στα 6 αρχεία γιατί παρουσιάζει αυξημένο το πλήθος των κανόνων , η TLSP στα D3, D3 και D6, και στο D5 η τεχνική TLSS.

Για κάθε αρχείο, D1 έως D6, θα παρουσιαστούν παρακάτω οι πέντε αποδοτικότεροι κανόνες. Ο κανόνας που δημιουργείται αποτελείται από το αριστερό μέρος, μία λέξη από το κείμενο, και στο δεξιό μέρος, ένα συναίσθημα ή την πολικότητα που παρουσιάζει η λέξη αυτή. Οι στήλες στον πίνακα αποτελεσμάτων του κάθε συνόλου δεδομένων δείχνουν την προεπεξεργασία που πραγματοποιήθηκε σε κάθε εκτέλεση. Οι προεπεξεργασίες των δεδομένων που δεν θα αναφέρονται, επέδειξαν αποτελέσματα μη

ικανοποιητικά για την συγκεκριμένη έρευνα. Οι γραμμές στον πίνακα αντιστοιχούν στον κανόνα με την εμπιστοσύνη και την υποστήριξη του για κάθε στήλη. Τα ποσοστά των κανόνων θα δείχνουν πρώτα την υποστήριξη και μετά την εμπιστοσύνη του κανόνα με την μορφή *υποστήριξη%/εμπιστοσύνη%*

Για το αρχείο D1 τα αποτελέσματα είναι :

		TL	TLS	TLSS
1	STATE → HAPPY	59%/75%	69%/82%	55%/71%
2	VOICE → SHAME	37%/42%	40%/45%	35%/40%
3	WAR → FEAR	36%/43%	39%/44%	33%/40%
4	TIME → SAD	32%/47%	34%/48%	30%/30%
5	FACE → SAD	29%/43%	45%/51%	28%/33%

Πίνακας 4.6 Κανόνες εξόρυξης για το αρχείο D1

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	STATE → HAPPY	59%/75%	69%/82%	29%/31%
2	ACCURATE →ANGER	30%/37%	31%/37%	22%/26%
3	TELL →FEAR	28%/34%	25%/31%	22%/27%

Πίνακας 4.7 Κανόνες εξόρυξης για το αρχείο D1 με TLSP

Ο κανόνας STATE → HAPPY και στις τρεις μεθόδους ήταν πρώτος σε υποστήριξη και εμπιστοσύνη, και παρουσιάζει σημαντικά αποτελέσματα αφού στην επεξεργασία με διαίρεση σε όρους, μετατροπή σε πεζά γράμματα και αφαίρεση των stopwords εμφανίζει υποστήριξη 69% και εμπιστοσύνη 82%. Ως δεύτερος κανόνας συσχέτισης στην πρώτη και στην τρίτη μέθοδο βγήκε ο VOICE → SHAME, ενώ στην δεύτερη ο FACE → SAD, ο οποίος στις άλλες δύο επεξεργασίες δεν επέδειξε καλά ποσοστά. Τρίτος κανόνας σε όλες τις επεξεργασίες είναι ο WAR → FEAR και τέταρτος ο TIME → SAD. Ενώ στον Πίνακα 4.7 που προστίθεται η υλοποίηση TLSP με την λεξικογραφική ανάλυση ο κανόνας STATE → HAPPY παύει πλέον να είναι ισχυρός με 29% υποστήριξη και 31% εμπιστοσύνη, ποσοστά πολύ χαμηλότερα από αυτά που επέδειξε στις άλλες τρείς τεχνικές στον Πίνακα 4.6. Το ίδιο συμβαίνει και στους

κανόνες ACCURATE →ANGER και TELL →FEAR όπου τα αποτελέσματα τους δεν αποδίδουν όταν προστίθεται λεξικογραφική ανάλυση.

Για το D2 παρουσιάζονται οι εξής κανόνες :

		TL	TLS	TLSS
1	STAR →HAPPY	54%/56%	49%/50%	50%/52%
2	WISH→WORRY	39%/47%	41%/48%	31%/39%
3	GREAT → WORRY	29%/39%	37%/39%	29%/37%
4	SUN → SAD	33%/37%	37%/40%	29%/36%
5	FRIEND → LOVE	26%/32%	34%/37%	26%/29%

Πίνακας 4.8 Κανόνες εξόρυξης για το αρχείο D2

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	STAR → HAPPY	54%/56%	49%/50%	21%/26%
2	ALTERNATIVE→SAD	31%/38%	28%/32%	30%/37%
3	CARE → LOVE	27%/30%	24%/29%	19%/22%

Πίνακας 4.9 Κανόνες εξόρυξης για το αρχείο D2 με TLSP

Στο αρχείο D2 ο κανόνας με την μεγαλύτερη υποστήριξη σε κάθε μέθοδο στον Πίνακα 4.8 είναι ο STAR →HAPPY και είναι ο μόνος που εμφάνισε ποσοστά μεγαλύτερα του 50%. Ο δεύτερος κανόνας συσχέτισης είναι ο WISH→WORRY με 41% υποστήριξη στην TLS. Στην επεξεργασία μόνο με διαίρεση σε όρους και μετατροπή των χαρακτήρων σε πεζά ο τρίτος κανόνας με βάση την υποστήριξη ήταν ο GREAT → WORRY, ενώ όταν προστέθηκε η αφαίρεση των stopwords, τρίτος ήταν ο SUN → SAD. Πέμπτος κανόνας σε όλες τις επεξεργασίες βγήκε ο FRIEND → LOVE, με αποτέλεσμα 29% όταν υλοποιήθηκε μαζί με την αποκατάληξη. Στον κανόνα STAR →HAPPY όταν προστέθηκε η λεξικογραφική ανάλυση τα ποσοστά έπεσαν κατακόρυφα στο 21% η υποστήριξη και 26% εμπιστοσύνη, ενώ ο κανόνας ALTERNATIVE→SAD έδειξε βελτίωση και έγινε πιο ισχυρός με λεξικογραφική ανάλυση με επίθετα παρά με την TLS, όπως φαίνεται στον Πίνακα 4.9.

Από το αρχείο D3 αντλήθηκαν οι εξής κανόνες :

		TL	TLS	TLSS	TLSP (ADJECTIVE)
1	LITTLE→ LOVE	63%/67%	62%/65%	41%/43%	64%/69%
2	LAST → ANGER	54%/59%	54%/58%	37%/42%	58%/61%
3	HANDSOME→ JOY	50%/55%	47%/50%	37%/40%	51%/57%
4	FAMILIAR→ SUPRISE	48%/51%	47%/51%	35%/39%	48%/52%
5	NEXT → FEAR	45%/50%	44%/46%	32%/36%	46%/50%

Πίνακας 4.10 Κανόνες εξόρυξης για το αρχείο D3 με TLSP (ADJECTIVE)

Το D3 αρχείο απέδωσε κανόνες με αρκετά μεγάλα ποσοστά υποστήριξης και εμπιστοσύνης, σε σχέση με το D1 και το D2, τα οποία επίσης κατηγοριοποιούν τα κειμενικά δεδομένα σε συναισθήματα. Ο κανόνας συσχέτισης με την μεγαλύτερη υποστήριξη σε κάθε επεξεργασία είναι ο LITTLE→ LOVE. Ακολουθεί ο LAST → ANGER, με 61% στην επεξεργασία διαίρεση σε όρους, μετατροπή σε πεζά, αφαίρεση stopwords και λεξικογραφική ανάλυση ταξινομημένο μόνο σε επίθετα. Ο HANDSOME→ JOY ήταν ο τρίτος κανόνας όπως φαίνεται και στον παραπάνω πίνακα. Οι κανόνες FAMILIAR→ SUPRISE και NEXT → FEAR, αντλήθηκαν με ποσοστά 51% και 50% στην εμπιστοσύνη, παρουσιάζοντας μικρές αλλαγές στην υποστήριξη ανάμεσα στις μεθόδους. Για όλους τους κανόνες παρατηρείται ότι στην τεχνική TLSS τα ποσοστά μειώθηκαν από τις υπόλοιπες τρείς.

Αν στο αρχείο D3 αντί για επίθετα, η λεξικογραφική ανάλυση γίνει με ουσιαστικά τότε παρουσιάζονται οι εξής κανόνες :

		TL	TLS	TLSS	TLSP (NOUN)
1	SPEED→ FEAR	49%/52%	50%/53%	37%/40%	50%/53%
2	TEAM → SUPRISE	48%/50%	48%/51%	34%/38%	48%/50%
3	POSTER → HAPPY	43%/46%	48%/50%	30%.35%	48%/49%
4	TV → SAD	39%/40%	41%/45%	28%/30%	38%/42%
5	GOLD →JOY	35%/37%	36%/39%	22%/24%	36%/38%

Πίνακας 4.11 Κανόνες εξόρυξης για το αρχείο D3με TLSP (NOUN)

Οι κανόνες που δημιουργήθηκαν έχουν μικρότερα ποσοστά από αυτά που παρουσιάστηκαν στον Πίνακα 4.11 με τα αποτελέσματα των επιθέτων. Οι κανόνες βγαίνουν σε κάθε επεξεργασία με την σειρά που βρίσκονται στον πίνακα από την γραμμή 1 έως 5 και δεν παρουσιάζουν ιδιαίτερες μεταβολές στα ποσοστά υποστήριξης και εμπιστοσύνης ανάμεσα στις μεθόδους. Ενώ και πάλι παρατηρείται το φαινόμενο της μείωσης των ποσοστών στην υλοποίηση TLSS.

Και παρακάτω βρίσκονται τα αποτελέσματα για τα αρχεία D4, D5, D6

		TL	TLS	TLSS	TLSP(VERB)
1	THINK →POSITIVE	93%/97%	95%/98%	79%/81%	97%/99%
2	HAVE → NEGATIVE	88%/92%	89%/95%	76%.79%	92%/96%
3	CAN → NEGATIVE	85%/90%	86%/90%	71%/74%	89%/91%
4	GO → POSITIVE	85%89%	88%/91%	67%/73%	86%/90%
5	GET → NEGATIVE	86%/87%	86%/86%	66%/70%	86%/89%

Πίνακας 4.12 Κανόνες εξόρυξης για το αρχείο D4

		TL	TLS	TLSP (NOUN /ADJ)
1	BAG→POSITIVE	82%/83%	83%/83%	66%/68%
2	TREE→NEGATIVE	81%/83%	84%/85%	61%/64%
3	CIVIL→POSITIVE	75%/77%	75%/78%	59%/63%

Πίνακας 4.13 Κανόνες εξόρυξης για το αρχείο D4 με TLSP (NOUN/ADJ)

Το D4 αρχείο κατατάσσει κάθε tweet σύμφωνα με την πολικότητα, δηλαδή είτε θετικό είτε αρνητικό. Από τον Πίνακα 4.12 φαίνεται ότι τα αποτελέσματα των ισχυρότερων κανόνων του συνόλου δεδομένων είναι άνω των 85% η υποστήριξη και η εμπιστοσύνη. Ο κανόνας THINK →POSITIVE είναι ο ισχυρότερος που εμφανίζεται στο αρχείο σε κάθε τεχνική προεπεξεργασίας του κειμένου, με 97% υποστήριξη στην τελευταία μέθοδο του πίνακα, στην οποία εκτός από διαίρεση με τόκους, μετατροπή σε πεζά και αφαίρεση των stopwords, έγινε και λεξικογραφική ανάλυση απομονώνοντας

τα ρήματα του κειμένου . Το ίδιο δεν συμβαίνει, όμως, όταν γίνεται λεξικογραφική ανάλυση με ουσιαστικά και επίθετα, αφού στην περίπτωση αυτή τα ποσοστά ων κανόνων μειώνονται σημαντικά, όπως φαίνεται στον Πίνακα 4.13

Για το αρχείο D5:

		TL	TLS	TLSS
1	CENTER →POSITIVE	90%/93%	93%/94%	87%/90%
2	TIME→ POSITIVE	89%/92%	90%/92%	89%/90%
3	LINE → NEGATIVE	90%/91%	93%/96%	88%/85%
4	SALE →NEGATIVE	87%/91%	88%/90%	86%/90%
5	TICKET → NEGATIVE	81%/87%	85%/89%	86%/88%

Πίνακας 4.14 Κανόνες εξόρυξης για το αρχείο D5

		TL	TLS	TLSP (NOUN /ADJ/VERB)
1	CENTER →POSITIVE	90%/93%	93%/94%	73%/77%
2	CURIOUS →POSITIVE	80%/83%	82%/86%	69%/74%
3	BAKE →NEGATIVE	78%/80%	80%/81%	66%/68%

Πίνακας 4.15 Κανόνες εξόρυξης για το αρχείο D5 με TLSP

Στο αρχείο D5 τα ποσοστά υποστήριξης είναι χαμηλότερα σε σχέση με το D3 αλλα εξίσου σημαντικά, αφού είναι 80% και άνω. Ο ισχυρότερος κανόνας στην τεχνική διαίρεση του κειμένου σε όρους και μετατροπή των χαρακτήρων σε πεζά, είναι ό CENTER →POSITIVE, αν προστεθεί και η αφαίρεση των stopwords ισχυρότερος αποδεικνύεται ο LINE → NEGATIVE. Με προσθήκη στην επεξεργασία την αποκατάληξη τα ποσοστά μειώνονται και ο κανόνας TIME→ POSITIVE είναι πλέον ο ισχυρότερος που αντλήθηκε από το κείμενο. Η τεχνική TLSP (NOUN/ADJ/VERB) δεν αποδίδει τα μέγιστα σε σύγκριση με τις άλλες μεθόδους όπως αποδεικνύεται από τον Πίνακα 4.15.

Για το αρχείο D6:

		TL	TLS	TLSS	TLSP (NOUN)
1	WATCH →POSITIVE	95%/98%	96%/98%	78%/83%	98%/99%
2	ORANGE → POSITIVE	92%/94%	92%/95%	77%/80%	94%/95%
3	SERIES→ NEGATIVE	89%/92%	93%/95%	72%/79%	89%/90%
4	BOOK→POSITIVE	85%/88%	84%/89%	65%/68%	85%/90%

5	FOOD→ NEGATIVE	81%/83%	83%/87%	62%/63%	85%/88%

Πίνακας 4.16 Κανόνες εξόρυξης για το αρχείο D6 με TLSP (NOUN)

		TL	TLS	TLSP (ADJ/VERB)
1	EASY→POSITIVE	83%/88%	84%/88%	70%/72%
2	SING →POSITIVE	80%/84%	83%/85%	67%/70%
3	CLEAN →NEGATIVE	76%/78%	81%/83%	87%/88%

Πίνακας 4.17 Κανόνες εξόρυξης για το αρχείο D6 με TLSP(ADJECTIVE/VERB)

Στον Πίνακα 4.16 αναλύονται οι κανόνες εξόρυξης που αντλήθηκαν από τον αρχείο D6. Μεγαλύτερη υποστήριξη και εμπιστοσύνη επέφερε ο WATCH →POSITIVE και εξίσου καλά αποτελέσματα είχε ο ORANGE → POSITIVE, ο οποίος έχει αρκετό ενδιαφέρον διότι η λέξη orange που είναι ένα χρώμα ή ένα φρούτο έχει έντονη θετική συναισθηματική πολικότητα. Σε αντίθεση με την λέξη food, που σημαίνει στα ελληνικά φαγητό, έχει κατά 85% αρνητική πολικότητα. Όταν προστέθηκε η μέθοδος TLSS, στην οποία στο τέλος γίνεται αποκατάληξη τα ποσοστά μειώθηκαν, καθώς και με την τεχνική TLSP απομονώνοντας τα επίθετα και τα ρήματα, με εξαίρεση τον κανόνα CLEAN →NEGATIVE, ο οποίος έγινε ισχυρότερος με την τεχνική αυτή.

Σε κάθε αρχείο πραγματοποιήθηκαν με τις δύο υλοποιήσεις όλες οι τεχνικές επεξεργασίας με διαφορετική σύνθεση και σειρά. Στα αρχεία D1,D2 και D5 η προεπεξεργασία της αποκατάληξης βελτίωσε τα αποτελέσματα, ενώ στα αρχεία D3,D4 και D6 η λεξικογραφική ανάλυση. Σε κάθε περίπτωση όμως η τεχνική διαίρεσης σε όρους, μετατροπή χαρακτήρων σε πεζά και έπειτα η αφαίρεση των stopwords βελτίωσε την υποστήριξη σε όλα τα αρχεία και δημιουργήθηκαν ισχυροί κανόνες.
Οι 5 ισχυρότεροι κανόνες, με αυτήν την προεπεξεργασία, με ελάχιστο όριο υποστήριξης και εμπιστοσύνης το 30%, στα αρχεία D1 – D3 είναι :

	Υποστήριξη/εμπιστοσύνη
STATE → HAPPY	69%/82%
LITTLE→ LOVE	62%/62%
LAST → ANGER	54%/58%
SPEED→ FEAR	50%/53%
STAR →HAPPY	49%/50%

Πίνακας 4.18 Ισχυρότεροι κανόνες εξόρυξης με αφαίρεση stopwords για τα αρχεία D1 – D3

Ενώ στα αρχεία D4 – D6 με ελάχιστο όριο υποστήριξης και εμπιστοσύνης το 85% είναι:

	Υποστήριξη/εμπιστοσύνη
WATCH →POSITIVE	96%/98%
THINK →POSITIVE	95%/98%
LINE → NEGATIVE	93%/96%
SERIES→ NEGATIVE	93%/95%
CENTER →POSITIVE	93%/94%

Πίνακας 4.19 Ισχυρότεροι κανόνες εξόρυξης με stopwords για τα αρχεία D4 – D6

Κάποια συμπεράσματα που μπορεί να εντοπιστούν στις υλοποιήσεις είναι:

- Τα σύνολα δεδομένων D1, D2 και D3 παρουσιάζουν κανόνες με ποσοστά υποστήριξης 35%-40%, ενώ τα D4, D5 και D6 περίπου στο 85%. Αυτό συμβαίνει διότι τα πρώτα τρία αρχεία δημιουργούν κανόνες συσχέτισης και έχουν ταξινομηθεί τα κειμενικά τους δεδομένα με βάση τα συναισθήματα (κατά μέσο όρο 5), σε αντίθεση με τα υπόλοιπα τρία που ταξινομούν τα tweets με βάση την πολικότητα, δηλαδή θετικά και αρνητικά. Ωστόσο τα ποσοστα 35%-40% είναι αρκετά μικρά και μπορούν να θεωρηθούν ακόμα και τυχαία.
- Στα αρχεία D1-D3 τα όρια υποστήριξης και εμπιστοσύνης, που είναι αποδοτικότεροι, είναι το 30% στην υποστήριξη και το 37% στην εμπιστοσύνη και στα σύνολα D4-D6 το 85% στην υποστήριξη και στην εμπιστοσύνη.

- Η τεχνική προεπεξεργασίας που αποδίδει τα μέγιστα και βελτιώνει τα ποσοστά των κανόνων σε όλα τα κείμενα είναι η διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζά και αφαίρεση των Stopwords.
- Το D2 είναι το αρχείο με τα περισσότερα tweets, ανέρχονται στα 39.812, όμως περιέχει τους λιγότερα ισχυρούς κανόνες και τα λιγότερα συχνά στοιχειοσύνολα. Είναι ένα φαινόμενο που παρατηρείται διότι στο αρχείο περιέχονται μεγάλος αριθμός στον πλήθος λέξεις και συναισθήματα, και είναι δύσκολο να δημιουργηθούν ισχυροί κανόνες αφού κάθε λέξη μπορεί να περιέχεται σε κείμενα που αντιστοιχούν σε διαφορετικό συναίσθημα. Ταυτόχρονα, ως αποτέλεσμα των παραπάνω, παρατηρείται μεγάλος αριθμός κανόνων με μικρό ποσοστό υποστήριξης.
- Στα αρχεία D3,D4 και D6 η τεχνική προεπεξεργασίας που αποδίδει εξίσου καλά, ίσως και καλύτερα, αποτελέσματα από το Stopwords, είναι η προσθήκη της λεξικογραφικής ανάλυσης με απομόνωση των ουσιαστικών, των επιθέτων ή των ρημάτων. Τα τρία αυτά αρχεία έχουν ως κοινό ότι αποτελούνται από κείμενα που προέρχονται από την ίδια θεματολογία. Δηλαδή στο σύνολο D4 όλα τα tweets προέρχονται από κείμενα που έχουν αναρτήσει οι χρήστες του twitter για το μουσικό φεστιβάλ Coachella, στο αρχείο D3 αναφέρονται για την τεχνική υποστήριξη εταιριών στους πελάτες και στο σύνολο δεδομένων D6 τα κείμενα προέρχονται από διαφωνίες που έχουν παρατηρηθεί ανάμεσα στους χρήστες μέσα στο Twitter. Στα D1,D2 και D5 όταν προστίθεται η λεξικογραφική ανάλυση τα ποσοστά υποστήριξης μειώνονται κατακόρυφα, με πολλές φορές διαφορά στο 15% στην υποστήριξη και στην εμπιστοσύνη, διότι τα tweets που περιέχουν δεν αναφέρονται σε ένα κοινό θέμα, αλλά είναι ασύνδετα ως προς το περιεχόμενο.
- Το αρχείο D5 δείχνει μεγαλύτερη απόδοση στην τεχνική TLSS, βελτιώνοντας αισθητά τα αποτελέσματα. Στα αρχεία D1 και D2 η μέθοδος TLSS αποδίδει μια σταθεροποίηση στα ποσοστά της υποστήριξης ανάμεσα στους κανόνες συσχέτισης. Επίσης παρατηρείται ότι παρόμοια αποτελέσματα παρουσιάζουν η TLS και η TLSS στα αρχεία αυτά, παρά τις μεγάλες διαφορές που έχουν μεταξύ τους οι δύο υλοποιήσεις.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 ΣΥΝΟΨΗ

Η ανάλυση συναισθημάτων είναι ένα πεδίο που αναπτύσσεται με ταχείς ρυθμούς και έχει ένα διευρυμένο πεδίο εφαρμογής στην τεχνολογία. Αρα αυξάνεται η ανάγκη για την δημιουργία κανόνων συσχέτισης με το συναίσθημα που προσεγγίζουν οι λέξεις σε ένα κείμενο από δημόσια tweets. Η υλοποίηση που πραγματοποιήθηκε στην εργασία αυτή, είναι κώδικας στην γλώσσα προγραμματισμού Python και η χρήση του προγράμματος RapidMiner. Τα δεδομένα που χρησιμοποιήθηκαν είναι 6 σύνολα δεδομένων, αποτελούμενα από το κείμενο των tweets και το συναίσθημα ή την πολικότητα που κατηγοριοποιείται κάθε κείμενο. Σε αυτά τα δεδομένα πραγματοποιείται προεπεξεργασία των κειμένων με διάφορες τεχνικές, όπως διαίρεση του κειμένου σε όρους, αφαίρεση των Stopwords, λεξικογραφική ανάλυση κ.α.. Έπειτα με την χρήση της Αρχής Apriori μέσα από το μέτρο της υποστήριξης εξάγονται οι κανόνες συσχέτισης, αποτελούμενοι στο δεξί μέρος μία λέξη του κειμένου και στο αριστερό ένα συναίσθημα ή την πολικότητα της λέξης. Η καλύτερη μέθοδος προεπεξεργασίας αποδείχτηκε η διαίρεση του κειμένου σε όρους, μετατροπή των χαρακτήρων σε πεζά και αφαίρεση των Stopwords, όπου εμφάνισε τα καλύτερα αποτελέσματα για όλα τα σύνολα δεδομένων. Η απόδοση των υλοποιήσεων στα αρχεία που κατηγοριοποιούν τα δεδομένα σε συναισθήματα ήταν περίπου 35%-40% οι ισχυρότεροι κανόνες και στα αρχεία με ταξινόμηση των κειμένων με βάση την πολικότητα τους, περίπου 90%. Στα αρχεία με tweets με θεματολογία, η προσθήκη της λεξικογραφικής ανάλυσης στην προεπεξεργασία του κειμένου, ανέβασε τα ποσοστά της υποστήριξης των κανόνων συσχέτισης. Μέσα από αρκετά πειράματα αποδόθηκαν σε κάθε κείμενο οι 5 ισχυρότεροι κανόνες και η τεχνική προεπεξεργασίας που έφερνε μεγαλύτερη απόδοση στα αποτελέσματα.

5.2 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Η έρευνα θα μπορούσε στο μέλλον να επεκταθεί στην υλοποίηση διαφορετικών αλγορίθμων και με διαφορετικά μέτρα υπολογισμού των ισχυρών κανόνων συσχέτισης. Επίσης θα μπορούσε να ληφθούν διαφορετικές πηγές δεδομένων από άλλα μέσα κοινωνικής δικτύωσης. Τέλος να χρησιμοποιηθεί η έρευνα σε προϊόντα ή εταιρίες για να γνωρίσουν την άποψη των πελατών τους και να μπορέσουν να στηρίξουν την διαφημιστική τους καμπάνια στα συναισθήματα αυτά

6. ΠΑΡΑΠΟΜΠΕΣ

1. Pang B., Lee L., 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, the 42nd Annual Meeting on Association for Computational Linguistics.
2. Mishne A., Gilad D., Glance N., 2006. “Predicting Movie Sales from Blogger Sentiment”, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.
3. Denecke K., 2008. "Using SentiWordNet for multilingual sentiment analysis", IEEE 24th International Conference on Data Engineering Workshop, Cancun, PP 507- 512.
4. Abbasi A., Chen H., Salem A., 2008. “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums”, ACM Transactions on Information Systems (TOIS), Vol. 26, PP 1-34.
5. Singh V., Piryani R., Uddin A., Waila P., 2013. "Sentiment analysis of movie reviews: A new featurebased heuristic for aspect-level sentiment classification," International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), PP 712-717.
6. Tripathy A., Agrawal A., Rath S. K., 2016. “Classification of sentiment reviews using n-gram machine learning approach”, Expert Systems with Applications, Vol. 57, PP. 117-126.
7. Wang S., Manning C., 2012. “Baselines and bigrams: simple, good sentiment and topic classification”, the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Vol. 2, PP 90-94.
8. Zhuang L., Jing F., Zhu X-Y., 2006. “Movie Review Mining and Summarization”, the 15th ACM international conference on Information and knowledge management, PP. 43-50.
9. Narayanan V., Arora I., Bhatia A., 2013. “Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model”, Intelligent Data Engineering and Automated Learning – IDEAL, Springer, Vol. 8206.
10. Kennedy A.,Inkpen D., 2006. “Sentiment Classification of Movie Reviews Using Contextual Valence Shifters”, Computational Intelligence, Vol. 22, PP 110-125.

11. Maas A., Daly R., Pham P., Huang D., Ng A., Potts C., 2011. “Learning word vectors for sentiment analysis”, the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, PP 142-150.
12. Benamara F., Cesarano C., Reforgiato D., “Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone”, Proceedings of International Conference on Weblogs and Social Media, pp. 1-7, 2006.
13. Hummel R.A, Zucker S.W., “On the Foundations of Relaxation Labeling Processes”, Proceedings of International Conference on Computer Vision: Issues, Problems, Principles, and Paradigms, pp. 585-605, 1987.
14. Moghaddam S and Ester M., 2012. “ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews”, Proceedings of 34th International ACM Conference on Research and Development in Information Retrieval, pp. 665-674, 2011.
15. Albornoz J., Plaza L., Gervas P., Diaz,A., 2011. “A Joint Model of Feature Mining and Sentiment Analysis for Product Review Rating”, Proceedings of International Conference on Advances in Information Retrieval, pp. 55-66
16. Ruhi R., Umar A., 2014. "Social Media Analytics as a business intelligence practice: current landscape & future prospects." Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects", Journal of Internet Social Networking & Virtual Communities.
17. Zhunchen L., Osborne M., Wang T., 2012. "Opinion retrieval in twitter." Sixth International AAAI Conference on Weblogs and Social Media.
18. Jilin C., et al., 2013. "CrowdE: Filtering tweets for direct customer engagements." Seventh International AAAI Conference on Weblogs and Social Media.
19. HarishRao M. et al. 2017. “Automatic Product Review Sentiment Analysis Using Vader and Feature Visulaization.”.
20. Alexander P., Paroubek P., 2010. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10.
21. Asur M., Sitaram A., Bernardo A., Huberman S., 2010. "Predicting the future with social media." Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society

22. Akshat, et al., 2012. "Mining sentiments from tweets." Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis.
23. Solomon N., Gray P., 2008. "Business intelligence." Handbook on decision support systems 2. Springer, Berlin, Heidelberg. 175-193.
24. Amine H., Abdelmalek A., Hamou R., Simonet M., 2014. "Detecting opinions in tweets." arXiv preprint arXiv:1402.5123.
25. Bizhanova A., Aizhan C., Uchida O., 2014. "Product reputation trend extraction from twitter." Social Networking 3.04: 196.
26. Das A., Kanti, D. P. Acharjya O., Patra M., 2014. "Opinion mining about a product by analyzing public tweets in Twitter." 2014 International Conference on Computer Communication and Informatics. IEEE.
27. Shruti, et al., 2012. "Text mining for sentiment analysis of Twitter data." Proceedings of the International Conference on Information and Knowledge Engineering (IKE).
28. Kishori K., Pukhraj P., Shrishrimal R., Deshmukh A. 2015. "Twitter sentiment analysis: A review." International Journal of Scientific & Engineering Research 6.4: 957-964.
29. Swapna R. Kharche K., Bijole L., 2015. “Review on Sentiment Analysis of Twitter Data”. International Journal Of Computer Science And Applications Vol. 8, No.2
30. Duy O., et al. 2017. "Next generation business intelligence and analytics: a survey." arXiv preprint arXiv:1704.03402.
31. Kharde K., Vishal A., Sonawane W., 2016. "Sentiment analysis of twitter data: a survey of techniques." arXiv preprint arXiv:1601.06971.
32. Anusha, K. S., Radhika D., 2017. "A Survey on Analysis of Twitter Opinion Mining Using Sentiment Analysis.".
33. Hu M., Minqing H., Liu B., 2004. "Mining opinion features in customer reviews." AAAI. Vol. 4. No. 4.
34. Somprasertsri X., Gamgarn F., Lalitrojwong P., 2010. "Extracting product features and opinions from product reviews using dependency analysis." Seventh International Conference on Fuzzy Systems and Knowledge Discovery. Vol. 5. IEEE.

35. Ayesha, et al., 2013. "Feature level opinion mining of educational student feedback data using sequential pattern mining and association rule mining." International Journal of Computer Applications 81.10.
36. Anwar O., Akib S., Akib E., Ekram T., 2015. "Localized twitter opinion mining using sentiment analysis." Decision Analytics.
37. Bhanap N., Smita M., Kawthekar S., 2017. "Twitter Sentiment Polarity Classification& Feature Extraction."
38. Othman B., Rania, Belkaroui, R., Faiz R., 2017. "Extracting product features for opinion mining using public conversations in Twitter." Procedia computer science 112: 927-935.
39. Richa S., Nigam S., Jain R., 2014. "Mining of product reviews at aspect level." arXiv preprint arXiv:1406.3714.
40. Htay S., Su S., Lynn K., 2013. "Extracting product features and opinion words using pattern knowledge in customer reviews." The Scientific World Journal.
41. Yuanbin W., et al., 2009. "Phrase dependency parsing for opinion mining." Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3-volume 3. Association for Computational Linguistics.
42. Amani K., and Li Y., 2016. "Aspect-based opinion mining using dependency relations." Int J Comput Sci Trends Technol (IJCST): 113-123.
43. Liu Y., Navathe S., Pivoshenko A., Dasigi A., Dingledine R., Ciliax B., 2006. “Text analysis of Medline for discovering functional relationships among genes: evaluation of keyword extraction weighting schemes,” Int. J. Data Mining and Bioinformatics, Vol. 1, No 1.
44. Chen X.,Wu Y., “Personalized knowledge discovery: mining novel association rules from text” Available: www.siam.org/meetings/sdm06/proceedings/067chenx.pdf
45. Ahonen H., Heinonen O., Klemettinen M., Verkamo A., 1997.“Mining in the phrasal frontier,” in Proc. PKDD’97.1st European Symposium on Principle of data Mining and Knowledge Discovery, Norway, June, Trondheim.
46. Mannila H., Toivonen H., Verkamo A., 2012. “Discovery of frequent episodes in event sequences,” Data Mining and Knowledge Discovery, 1(3), 1997b, pp. 259-289.
47. Feldman R.,Dagan L.,1995. “Knowledge discovery in textual databases (KDT)”, in Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining.

48. Norvag K, Eriksen T., Skgstad K., 2006. “Mining association rules in temporal document collections,” Available: http://www.idi.ntnu.no/~noervaag/papers/ISMIS2006.pdf
49. Rajman M., Besancon R., 1997. “Text mining: natural language techniques and text mining applications”, in Proc. 7th working conf. on database semantics (DS-7), Chapan &Hall IFIP Proc. Series. Leysin, Switzerland Oct. 1997, 7-10.
50. Chen C., An-Pin C., 2006. Using data mining technology to provide a recommendation service in the digital library. The Electronic Library. 25(6): 711-734.
51. Combas C., Meskens N., Vandamme J. P., 2007. Using a KDD process to forecast the duration of surgery. International Journal of Production Economics. 112: 279-293.
52. Karen A., Cochran K., 1999. Using data mining and datawarehousing techniques. Industrial Management & Data Systems. 99(5), 189-196.
53. Gargano N., Michael L., Raggad J., Bel G., 1999. Data mining – a powerful information creating tool. OCLC Systems & Services. 15(2), 81-90.
54. Imberman N, Susan P., 2001. Effective Use of the KDD Process and Data Mining for Computer Performance Professionals. in International Computer Measurement Group Conference. Anaheim: USA, 611-620.
55. Ma C., Catherine N., Chou B., David C., Yen V, David C., 2000. Data warehousing, technology assessment and management. Industrial Management & Data Systems. 100(3), 125-135.
56. Min K., Min H., Emam A., 2002. A data mining approach to developing the profiles of hotel customers. International Journal of Contemporary Hospitality Management. 14(6): 274-285.
57. Palace B., 1996. Data Mining: What is Data Mining?
58. Lavrač N., Zupan B., 2005. Data Mining in Medicine. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_52
59. Guang C., WANG k., 2009. 1School of Computer Science and Engineering, Key Laboratory of Computer N

60. Sehral, A. K. (2008). Text Mining: The search for novelty in text. A report submitted in partial fulfillment of the requirements of the Ph.D Comprehensive Examination in the Department of Computer Science.

61. Βαζιργιάννης Μ., Χαλκίδη Μ., (2003).Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, Εκδ. Gutenberg.

62. Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν., Κόκκορας, Φ., 2006. Τεχνητή Νοημοσύνη Γ΄Έκδοση 2006, εκδ. Μ.Γκιούρδας.

63. Kotsiantis S., Kanellopoulos D., 2006. Association Rules Mining: A Recent Overview.

64. Guido R.., 1993. "An Introduction to Python for UNIX/C Programmers". Proceedings of the NLUUG Najaarsconferentie (Dutch UNIX Users Group).

65. Debill E., 2019. "Module Counts". ModuleCounts.

66. Norvig C., 2012. "Quotes about Python". Python Software Foundation.

67. Christopher D. Manning B., Raghavan P., Schütze H., 2008. Introduction to Information Retrieval, Cambridge University Press.

68. Jaaben P., 2018. “Stemming and Lemmatization in Python”.

69. Meer V., 2013. “How we use Python at Spotify”.

70. Recordon H., 2009. “Tornado: Facebook's Real-Time Web Framework for Python”. Facebook for Developers.