Πτυχιακή Εργασία

«Ανάλυση Συναισθήματος με χρήση τεχνικών Μηχανικής Μάθησης σε δεδομένα από το Twitter»

Μπακλέσης Δημήτριος

Επιβλέπων Καθηγητής

Ασημακόπουλος Γεώργιος

Αντίρριο, Απρίλιος 2019

ΠΕΡΙΛΗΨΗ

Η ραγδαία εξέλιξη του διαδικτύου οδήγησε στην ανάγκη για διαχείριση του μεγάλου όγκου πληροφοριών και δεδομένων σε μορφή κειμένου και έτσι δημιουργήθηκαν νέα εργαλεία επικοινωνίας και ανταλλαγής απόψεων. Αυτοματοποιημένες λοιπόν τεχνικές βοήθησαν στην Εξόρυξη Γνώσης από Κείμενο (TextMining) και στην Ανάλυση Συναισθήματος (SentimentAnalysis). Σκοπός της είναι η ανίχνευση της πολικότητας ενός κειμένου, έτσι ώστε να ανακαλυφθεί η υποκειμενική άποψη του συγγραφέα σχετικά με το θέμα του κειμένου, πληροφορία χρήσιμη σε ερευνητικό επίπεδο για κοινωνικά φαινόμενα αλλά και για την ανάπτυξη επιχειρήσεων. Στην παρούσα εργασία μελετώνται μέθοδοι Μηχανικής Μάθησης και τεχνικές Ανάλυσης Συναισθήματος και πραγματοποιείται μία συγκριτική μελέτη μοντέλων κατηγοριοποίησης συναισθήματος, στο προγραμματιστικό περιβάλλον Weka, δεδομένων που προέρχονται από το Twitter. Χρησιμοποιούνται οι αλγόριθμοι *Multinomial Naive Bayes* και *Support Vector Machines*.

ABSTRACT

The rapid development of the internet has led to the need to manage the large amount of information and data in text form, thus creating new tools for communication and exchange of views. Automated techniques therefore helped in Text Mining and Sentiment Analysis. It's purpose is to detect the polarity of a text so as to discover the subjective view of the author on the subject matter, information useful at the research level for social phenomena but also for business development. In this paper we study methods of Machine Learning and Sentiment Analysis techniques and we perform a comparative study of sentiment categorization models for data that are coming from Twitter, in the Weka programming suite. The algorithms that have been used are the *Multinomial Naive Bayes* and *Support Vector Machines*.

ΠΕΡΙΕΧΟΜΕΝΑ

1.ΕΙΣΑΓΩΓΗ

1.1ΓενικήΕισαγωγή

Στην σύγχρονη κοινωνία, η γνώση είναι το επίκεντρο του ενδιαφέροντος και στόχος όλων είναι η απόκτησή της. Γνώση είναι ένα σύνολο επεξεργασμένων δεδομένων και η σωστή κατάρτισή της επιτυγχάνεται μόνο με την ανάλυση αυτών, καθώς και της ερμηνείας τους. Με την ραγδαία εξέλιξη της τεχνολογίας, ήρθαμε αντιμέτωποι με έναν τεράστιο όγκο πληροφοριών που κληθήκαμε να τον διαχειριστούμε.Επιτακτική λοιπόν, ήταν η ανάγκη εξόρυξης γνώσης από κείμενο (textmining) και η ανάλυση συναισθήματος (sentimentalanalysis). Το πεδίο αυτό ερευνά την πολικότητα ενός κειμένου προκειμένου να διερευνηθεί η άποψη του συγγραφέα, πληροφορία χρήσιμη για μεγάλους οργανισμούς και εταιρείες. Αυτό γιατί, δείχνει τι σκέφτονται ή αισθάνονται οι χρήστες και τότε η πρόβλεψη των κινήσεων τους είναι πολύ πιοεύκολη.

1.2 Εισαγωγή στην ανάλυση δεδομένων

Η ανάλυση δεδομένων (dataanalysis) είναι μια διαδικασία συλλογής, καθαρισμού, μετατροπής και μοντελοποίησης δεδομένων για εύρεση χρήσιμων πληροφοριών και συμπερασμάτων.
Η εξόρυξη γνώσης είναι μια τεχνική ανάλυσης δεδομένων με σκοπό την ανακάλυψη γνώσης και της μοντελοποίησής της (εύρεση προτύπων) για πρόβλεψη και όχι τόσο για περιγραφή καταστάσεων.
Η ανάλυση για πρόγνωση (predictiveanalytics) εφαρμόζει στατιστικά μοντέλαπρόβλεψης ή κατηγοριοποίησης και η ανάλυση κειμένου (textanalytics) με στατιστικές και γλωσσολογικές τεχνικές επιτυγχάνει την εξόρυξη και κατηγοριοποίηση πληροφορίας από πηγές αδόμητων δεδομένων.[1]

1.3 Εισαγωγή στον όρο “BigData”

Η ανάλυση τεράστιου όγκου πολύπλοκων δεδομένων έχει οδηγήσει την αφορά στον ορισμό μιας νέας έννοιας για την διαχείριση αυτών, γνωστή πλέον ως «BigData».
Η διαφορά του όρου ανάλυσης δεδομένων με αυτού των bigdata φαίνεται εύκολα από το παρακάτω παράδειγμα.Έστω ότι υπάρχει μια διαρροή σε σωλήνα κήπου ενός κτηρίου και χρησιμοποιούμε ένα κουβά και ένα μονωτικό υλικό για να επιδιορθώσουμε τη ζημιά. Όμως σε λίγο το πρόβλημα επιδεινώνεται και συνειδητοποιούμε ότι χρειαζόμαστε έναν ειδικό, ο οποίος θα φέρει και ειδικά εργαλεία. Συνεχίζουμε βέβαια να χρησιμοποιούμε τον κουβά ως

λύση έως ότου δούμε πως έχει ανοίξει υπερβολικά η τρύπα και είμαστε αντιμέτωποι με εκατομμύρια λίτρα νερού που δεν σώζονται ούτε με περισσότερους κουβάδες. Το πρόβλημα χειροτερεύει με ανεξέλεγκτο νερό να βγαίνει από παντού και μια νέα προσέγγιση της λύσης να είναι πλέον αναγκαία για να προλάβουμε την μεγάλη ταχύτητα και τον όγκο του νερού. Η εικόνα αυτή έρχεται όσο πιο πολύ κοντά γίνεται στον κόσμο των bigdata. Στον κόσμο της αγοράς τώρα, έστω πώς θέλουμε να προβλέψουμε την τιμή της μετοχής μιας εταιρείας μέσω των socialmediaκαι πιο συγκεκριμένα μέσω του Twitterκαι των tweets(δεδομένα από το Twitter). Τα tweetsδείχνουν πόσες φορές έχουν αναφερθεί οι χρήστες στην εταιρεία ή σε κάποιο προϊόν της, τα συναισθήματα των εργαζομένων της και τη γνώμηπιθανών επενδυτών.Δεδομένου όμως πως υπάρχουν προβλήματα, η προσέγγιση της τιμής πρέπει να γίνει όσο πιο ακριβής γίνεται και σε αυτό βοηθούν κατάλληλα μοντέλα για τη σωστή επιλογή δεδομένων. Ορισμένα από τα προβλήματα είναι τα εξής:

- Υπάρχουν περισσότερα από 500 εκατομμύρια tweetsανά δευτερόλεπτο (μεγάλος όγκος και ταχύτητα).
- Πρέπει να γίνεται σωστή κατανόηση του κάθε tweet, έλεγχος αξιοπιστίας και κατηγοριοποίησης του (μεγάλη ποικιλία).
- Πρέπει να αναλύεται το συναίσθημα για παράδειγμα για ένα συγκεκριμένο προϊόν (θετικό ή αρνητικό) (μεγάλη πολυπλοκότητα).
- Η τελευταία διαδικασία πρέπει να γίνεται σε πραγματικό χρόνο (αφού η πιθανότητα να μεταβληθεί το συναίσθημα είναι μεγάλη).

[17]

1.4 Εισαγωγή στην Εξόρυξη Γνώσης Από Κείμενο

Με γνώμονα λοιπόν το πρόβλημα της υπερπληροφόρησης και η ανάγκη να μπορούν πλέον οι χρήστες να αξιοποιούν εύκολα και γρήγορα τις πληροφορίες, κάνει την εμφάνισή της η Εξόρυξη Γνώσης από Δεδομένα (DataMining). Τα δεδομένα αυτά είναι μη δομημένα (όπως είναι τακείμενα, οι εικόνες, τα έγγραφα, οι ιστοσελίδες)και έτσι ειδικότερα η ανάγκη αξιοποίησης δεδομένων σε μορφή κειμένου οδήγησε στην ανάπτυξη τεχνικών και εργαλείων της Εξόρυξη Γνώσης από Κείμενο (TextMining), που θα αναφερθούν στη συνέχεια της εργασίας.

2 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΚΕΙΜΕΝΟ

2.1 Εισαγωγή

Η ανάγκη λοπόν αξιοποίησης δεδομένων σε μορφή κειμένου, οδήγησε στην ανάπτυξη τεχνικών Εξόρυξης Γνώσης από Κείμενο (TextMining).Στόχος της εξόρυξης γνώσηςαπό κείμενο είναι η εύρεση πληροφοριών οι οποίες είναι κρυμμένες σε βάσεις δεδομένων[2].

2.2 Διαδικασία Ανακάλυψης Γνώσης

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων είναι «μία επαναληπτική διαδικασία, πολλών βημάτων, στην οποία απαιτείται η παρέμβαση του χρήστη για τη λήψη κρίσιμων αποφάσεων» (Fayyad, Piatesky-ShapiroandSmyth, 1996). [3]
Ανακάλυψη Γνώσης από Βάσεις Δεδομένων ή Εξόρυξη Γνώσης από Δεδομένα (DataMining) ή Εξαγωγή Γνώσης (KnowledgeExtraction) αποτελεί μια ολοκληρωμένη διαδικασία που περιλαμβάνει την επεξεργασία των δεδομένων, την εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και τέλος την αξιολόγηση και την ερμηνεία των αποτελεσμάτων. Στόχος της Ανακάλυψης Γνώσης είναι η ανάλυση μεγάλου όγκου πρωτογενών δεδομένων, για την ανάδειξη συγκεκριμένων δομών και σχέσεων ανάμεσά τους, έτσι ώστε τα πρότυπα, οι κανόνες ή/και οι περιορισμοί που θα εξαχθούν από τα δεδομένα, να υποστηρίξουν τον άνθρωπο στη διαδικασία λήψης αποφάσεων.
Τα στάδια που την αποτελούν είναι τα εξής:

Εικόνα 1: Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

2.2.1 Επιλογή

Σε αυτό το πρώτο στάδιο γίνεται η συλλογή των δεδομένων από όπου θα εξάγουμε την γνώση. Στο σύνολο αυτό που θα δημιουργηθεί θα γίνει η αναζήτηση των προτύπων και επηρεάζει πολύ όλη την διαδικασία, καθώς εδώ καθορίζεται η ποιότητα των δεδομένων και κατ' επέκταση η αξιολόγηση των αποτελεσμάτων.

2.2.2 Προεπεξεργασία

Το στάδιο αυτό ασχολείται με την αφαίρεση (καθαρισμό) και την επεξεργασία δεδομένων τα οποία ή δεν θα έπρεπε να υπάρχουν ή δεν είναι αρκετά πλήρη (datacleaning). Δεδομένου ότι είναι η πιο χρονοβόρα διαδικασία για την ανακάλυψη γνώσης (απαιτεί περίπου το 50% του συνολικού χρόνου), σκοπός είναι να μειώσουμε αυτό το χρόνο ώστε να δίνεται μεγαλύτερη εμφαση στην εξόρυξη δεδομένων και στην ερμηνεία των αποτελεσμάτων.

2.2.3 Μετασχηματισμός

Κατά τον μετασχηματισμό των δεδομένων, γίνεται μια προσπάθεια μετατροπής τους ώστε να ανήκουν σε ένα κοινό σύνολο, αφού μπορεί να είναι διαφορετικής προέλευσης. Μια τέτοια μετατροπή μπορεί να είναι η μείωση των χαρακτηριστικών των δεδομένων (dimensionalityreduction), επιλέγοντας κάποια από αυτά (featureselection, attributeselection) ή τη μετατροπή συνεχόμενων αριθμητικών τιμών σε διακριτές τιμές (διακριτοποίηση).

2.2.4 Εξόρυξη γνώσης από δεδομένα (DataMining)

Αποτελεί ίσως το βασικότερο στάδο, κατά το οποίο επιλέγεται η κατάλληλη μέθοδος εξόρυξης δεδομένων (summarization, classification, regression, clustering) και αλγόριθμοι που εφαρμόζονται στα μετασχηματισμένα πλέον δεδομένα για επιθυμητά αποτελέσματα (πρότυπα).

2.2.5 Ερμηνεία

Τελικό στάδιο στην διαδικασία ανακάλυψης γνώσης είναι η ερμηνεία και η αξιολόγηση του μοντέλου με την προϋπόθεση τα πρότυπα που έχουν προκύψει να είναι κατανοητά. Για τη σωστή ερμηνεία σημαντικό ρόλο παίζει η παρουσίαση των αποτελεσμάτων με οπτικοποίηση προτύπων και δεδομένων (pattern/visualization) αλλά και η διόρθωση διαφωνιών προηγούμενης πιστευτής γνώσης.[3]

2.3 Εξόρυξη Γνώσης από Κείμενο

Μετά λοιπόν από την Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (KDD) και τηνΕξόρυξη Γνώσηςαπό Δεδομένα (DataMining), ήρθε η ανάγκη για αξιοποίηση του τεράστιου όγκου δεδομένων σε μορφή κειμένου με αυτόματο τρόπο. Για αυτό και περνάμε στην επόμενη εξέλιξη που είναι η Εξόρυξη Γνώσης από Κείμενο. Κατά τη διαδικασία αυτή στόχος είναι η εξόρυξη προτύπων σε μη δομημένα κείμενα και συνδυάζονται τεχνικές από την Εξόρυξη Γνώσης από Δεδομένα, Μηχανική Μάθηση, τη Στατιστική, την Επεξεργασία Φυσικής Γλώσσας (NLP), την Ανάκτηση Πληροφορίας, την Εξαγωγή Πληροφορίας και τη Διαχείριση Γνώσης.[4][5]

2.4 Αναπαράσταση Κειμένου

Η αναπαράσταση ενός κειμένου κατά τη διαδικασία της εξόρυξης εμφανίζει δυσκολίες που αφορούν στην απουσία δομής. Για αυτό το λόγο συμπεριφερόμαστε στο κείμενο σαν μια «σακούλα λέξεων» (bagofwords), η οποία περιέχει όλες τις λέξεις που υπάρχουν στο κείμενο. Ένας τρόπος αναπαράστασης, ο οποίος είναι και ο πιο γνωστός, είναι η διανυσματική αναπαράσταση (vectorrepresentation). Σε αυτήν, το κείμενο αντιστοιχεί σε ένα διάνυσμα όρων (termvector) και κάθε όρος σε ένα μοναδικό χαρακτηριστικό (feature). Κάθε στοιχείο λαμβάνει μια τιμή, η οποία εκφράζει την παρουσίατου όρου στο κείμενο.
Το **Διανυσματικό Μοντέλο Αναπαράστασης Κειμένου** (VectorSpaceModel – VSM) χρησιμοποιείται συχνά στα συστήματα ανάκτησης πληροφορίας και στηρίζεται στην αναπαράσταση κειμένων ως διανύσματα σε έναν πολυδιάστατο Ευκλείδειο χώρο. Αποτελεί γενίκευση του Λογικού Μοντέλου (BooleanModel)*, είναι πιο απλό στη χρήση και πιο αποτελεσματικό.Κάθε όρος αποτελεί ένα χαρακτηριστικό του κειμένου.Κατά την αναπαράσταση ενός κειμένου, κάθε άξονας στο χώρο αντιστοιχεί σε ένα χαρακτηριστικό του κειμένου. Η συντεταγμένη κάθε διανύσματος περιγράφει την εμφάνιση του συγκεκριμένου χαρακτηριστικού στο κείμενο, δηλαδή εκφράζει το βάρος του όρου στο κείμενο, δηλώνοντας το πόσο σημαντικός θεωρείται ο όρος στο συγκεκριμένο κείμενο. Τα βάρη που είναι πραγματικές τιμές και μπορεί να είναι είτε απλά η συχνότητα εμφάνισης της λέξης (που εμφανίζει προβλήματα όπως στην περίπτωση των stopwords που δεν θεωρούνται σημαντικοί όροι), είτε άλλες τιμές. [39]

* Στο Λογικό Μοντέλο (BooleanModel) κάθε κείμενο αναπαριστάται από από ένα σύνολο λογικών τιμών (1 που δηλώνει την εμφάνιση του όρου στο κείμενο και 0 την απουσία του). Παρ'όλη την ευκολία που παρουσιάζει στην κατανόηση του και στον ικανοποιητικό χρόνο αναζήτηση, αδυνατεί να δηλώσει την σημασία της παρουσίας ενός όρου στο κείμενο.

2.5Προσεγγίσεις στην Εξόρυξη Γνώσης από Κείμενο

Η Εξόρυξη Γνώσης από Κείμενο προσεγγίζεται με μεθόδους της Μηχανικής Μάθησης. Στη συνέχεια αναφέρονται κάποιες τεχνικές Εξόρυξης Γνώσης από Κείμενο και μέθοδοι Μηχανικής Μάθησης που χρησιμοποιούνται στην εξόρυξη κειμένου.

- **Κατηγοριοποίηση κειμένου (textclassification)**

Πρόκειται για τη διαδικασία όπου αναθέτει στο κείμενο προκαθορισμένες κατηγορίες/κλάσεις. Μια τέτοια περίπτωση εφαρμογής της κατηγοριοποίησης κειμένου ειναι το «spam*filtering», όπου τα e-mails διαχωρίζονται στις κατηγορίες spam και non-spam (δυαδικό/binary πρόβλημα κατηγοριοποίησης).
Ανήκει στις επιβλεπόμενες μεθόδους μηχανικής μάθησης, γιατί οι κατηγορίες έχουν καθοριστεί πριν την εξέταση των δεδομένων. Δεδομένου τώρα ενός συνόλου εκπαίδευσης (trainingset) εκπαιδεύεται το μοντέλο κατηγοριοποίησης, μέσω στατιστικής ανάλυσης λέξεων και εφαρμόζεται έπειτα στην ταξινόμηση του testset (σύνολο δεδομένου ελέγχου, διαφορετικού του συνόλου εκπαίδευσης)και αξιολογούνται τα αποτελέσματα.

- **Συσταδοποίησηκειμένου (clustering)**

Και εδώ τα δεδομένα χωρίζονται σε ομάδες (συστάδες), όμως δεν είναι εκ των προτέρων καθορισμένες, όπως στην κατηγοριοποίηση. Ανήκει λοιπόν στις μη επιβλεπόμενες μεθόδους μηχανικής μάθησης. Η οργάνωση αυτή γίνεται με βάση τις ομοιότητες των χαρακτηριστικών των δεδομένων(π.χ ομαδοποίηση πελατών βάσει αγοραστικής τους συμπεριφοράς).

- **Εξαγωγή κανόνων συσχέτισης (associationrules)**

Οι κανόνες συσχέτισης ανήκουν στην μη επιβλεπόμενη μάθηση και είναι υπεύθυνοι για την εύρεση συσχετίσεων-κανόνωνπου περιγράφουν μεγάλα τμήματα των δεδομένων μας (π.χ. αυτοί που αγοράζουν το Χ προϊόν, τείνουν να αγοράζουν και το Υ).Έστω τα αντικείμενα Α και Β, τότε ένας κανόνας συσχέτισης εκφράζει τη συνεπαγωγή της εμφάνισης του Α στην εμφάνιση του Β στο ίδιο στιμιότυπο του προβλήματος (Α→Β). Η αξιολόγηση των κανόνων συσχετίσεων γίνεται με τον συντελεστή υποστήριξης s(ποσοστό των συναλλαγών στη βάση δεδομένων που περιέχουν το AUB) και τον συντελεστή εμπιστοσύνης α (κλάσμα των συναλλαγών που περιέχουν το AUB προς τον αριθμό των συναλλαγών που περιέχουν το Α). [2]

*spam ονομάζεται η μαζική αποστολή ηλεκτρονικών μηνυμάτων ή άλλων, σε μια προσπάθεια προώθησης προϊόντων ήιδεών.

- **Περίληψηκειμένου (summarization)**

Η διαδικασία της περίληψης μειώνει το μέγεθος το κειμένου χωρίς να χάνεται το νόημά του. Το πλήθος των λέξεων που θα εξαχθούν για την περίληψη καθορίζεται από τον χρήστη, επομένως και το μέγεθος της περίληψης εξαρτάται από αυτόν.

- **Γλωσσικός προσδιορισμός (languageidentification)**

Με την τεχνική του γλωσσικού προσδιορισμού γίνεται αντιληπτή η γλώσσα στην οποία είναι γραμμένο ένα κείμενο, καθώς και το ποσοστό του κειμένου που είναι γραμμένο σε κάθε γλώσσα στην περίπτωση που το κείμενο έχει γραφτεί σε περισσότερες από μία γλώσσες.

- **Απόδοση κειμένου σε συγγραφέα**

Η τεχνική της απόδοσης κειμένου σε συγγραφέα έχει σκοπό τον προσδιορισμό του συγγραφέα ενός κειμένου.

- **Οπτικοποίηση κειμένου (visualization)**

Με την οπτικοποίηση δίνεται η δυνατότητα της γραφικής απεικόνισης ενός συνόλου κειμένων με στόχο την κατανόηση του θέματος και των βασικών εννοιών από το χρήστη,αφού για παράδειγμα η σημασία του κειμένου αναγνωρίζεται από το μέγεθος στη γραφική απεικόνιση. Αυτό επιτυγχάνεται με την εξαγωγή χαρακτηριστικών γνωρισμάτων και κεντρικών όρων, και έτσι δημιουργείται η γραφική αναπαράσταση των κειμένων.[2]

2.6 Εισαγωγή στην Ανάλυση Συναισθήματος

Όλες αυτοί οι μέθοδοι μηχανικής μάθησης είναι μεγάλης σημασίας και γι’αυτό εφαρμόζονται για την επίτευξη ενός σημαντικού στόχου. Αυτός, που είναι και το θέμα της εργασίας αυτής, αφορά στην Ανάλυση του Συναισθήματος του συγγραφέα του εκάστοτε κειμένου, δηλαδή στη συλλογή χρήσιμων «υποκειμενικών» πληροφοριών για την εξαγωγή σωστών και χρήσιμων συμπερασμάτων που εκμεταλλεύονται διάφοροι επιστημονικοί κλάδοι, όπως θα δούμε παρακάτω.

3 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

3.1 Εισαγωγή

Η Ανάλυση Συναισθημάτων (sentimentanalysis) ή Εξόρυξη Γνώμης (opinionmining) αναφέρεται στην Επεξεργασία της Φυσικής Γλώσσας (NLP), χρησιμοποιώντας μεθόδους κατηγοριοποίησης κειμένου και υπολογιστικής γλωσσολογίας για την εξαγωγή υποκειμενικής πληροφορίας από διάφορες πηγές. [6]
Εφαρμόζεται σε κριτικές ταινιών και στα μέσα μαζικής δικτύωσης με σκοπό τη διαφήμιση και γενικά την οικονομική βελτίωση οργανισμών και εταιρειών.
Γενικότερα, στόχος είναι η ανάλυση του ύφους του συγγραφέα για ένα θέμα. Το ύφος μπορεί να αφορά στην κρίση/γνώμη του, στη συναισθηματική του κατάσταση (δηλαδή το πως νιώθει όταν γράφει) ή το επιτηδευμένο επικοινωνιακό συναίσθημα (δηλαδή εκείνο το συναίσθημα που θέλει ο συγγραφέας να περάσει στον αναγνώστη).
Το ενδιαφέρον για τον τομέα της ανάλυσης συναισθημάτων έγινε μεγαλύτερο με την έντονη χρήση των κοινωνικών μέσων (socialmedia), όπως είναι τα blogs.Μεγάλη συμβολή υπήρξε η εξέλιξη του διαδικτύου, με τη μετάβαση στη νέα γενιά του παγκόσμιου ιστού, Web 2.0. Με την εμφάνιση του Web 2.0 μπορούν πλέον οι χρήστες του διαδικτύου να μοιράζονται πληροφορίες και να συνεργάζονται διαδραστικά. Η διαδικτυακή αυτή πλατφόρμα δίνει τη δυνατότητα στους χρήστες να αλληλεπιδρούν χωρίς ειδικές γνώσεις γύρω από τους υπολογιστές και τα δίκτυα.Η μεγάλη αυτή λοιπόν ευχέρεια στις κριτικές, αξιολογήσεις και γενικά στην online έκφραση οδήγησε τις επιχειρήσεις στο να κινούνται με γνώμονα αυτές τις πληροφορίες, για να προωθήσουν προιόντα και να διαχειριστούν όσο καλύτερα γίνεται τη φήμη τους και επομένως να στραφούν στον τομέα της ανάλυσης συναισθημάτων.

3.2 Εφαρμογές της Ανάλυσης Συναισθήματος

Η Ανάλυση Συναισθήματος και Κειμένων στο διαδίκτυο έχει παίξει πολύ σημαντικό ρόλο και αποτελεί πλέον ένας από τους πιο χρήσιμους ερευνητικούς τομείς λόγω των εφαρμογών της. Συμβάλλει στην ανάλυση των κοινωνικών φαινομένων και τάσεων και συνεπώς αποτελεί αντικείμενο μελέτης για πολλά επιστημονικά πεδία, όπως η οικονομία, η κοινωνιολογία, η πολιτική και η ψυχολογία.
Η Ανάλυση Συναισθήματος έχει την δυνατότητα να συλλάβει την αντίληψη των χρηστών σχετικά με ένα θέμα, να παρακολουθήσει την πορεία τους μέσα στο χρόνο και τότε να συστήσει προιόντα και υπηρεσίες σε άτομα. Πηγή πληροφοριών για όλα αυτά και αιτία πραγματοποίησης τους αποτελούν τόσο τα κοινωνικά δίκτυα και τα προσωπικά ιστολόγια (blogs), όσο και οι ομάδες συζητήσεων (discussionforums). Επίσης, σημαντικό εργαλείο είναι και οι κριτικές/αξιολογήσεις προϊόντων και υπηρεσιών των χρηστών, των οποίων τα δεδομένα αποσαφηνίζουν τις απόψεις τους.Η Εξόρυξη Γνώμης δηλαδή των καταναλωτών, βοηθάει όχι μόνο τις επιχειρήσεις στο σωστό μάρκετινκγ και την προώθηση νέων προϊόντων/υπηρεσιών, στη βελτίωση της ποιότητας των ήδη υπάρχοντων και στην ενημέρωσή τους ως προς την απόδοση των ανταγωνιστών τους,αλλά και τους ίδιους τους καταναλωτές για την λήψη αποφάσεων, όσον αφορά τις αγορές τους.

3.2.1 Κοινωνικά δίκτυα

Σήμερα, τα μέσα κοινωνικής δικτύωσης αποτελούν ένα δημοφιλές επικοινωνιακό εργαλείο μεταξύ των χρηστών του διαδικτύου. Ένα κοινωνικό δίκτυο είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Σήμερα ο όρος αυτός χρησιμοποιείται για να περιγράψει ιστοσελίδες οι οποίες επιτρέπουν την διεπαφή ανάμεσα στους χρήστες, με τις δημοφιλέστερες να είναι το Facebook,το Twitter,το Instagram και το Linkedin (ιστοχώρος επαγγελματικής κοινωνικής δικτύωσης). Οι ιστότοποι κοινωνικής δικτύωσης είναι οργανωμένες ιστοσελίδες στο διαδίκτυο με περισσότερο ομαδοκεντρικό χαρακτήρα που παρέχουν κάποιες βασικές και δωρεάν υπηρεσίες όπως τη δημιουργία προφίλ, το ανέβασμα εικόνων και βίντεο, τον σχολιασμό σε ενέργειες που γίνονται από άλλα μέλη του δικτύου ή μίας ομάδας, την άμεση ανταλλαγή μηνυμάτων και πολλά άλλα.[11]

3.2.2 Αξιοποίηση της Ανάλυσης Συναισθήματος από επιχειρήσεις

Τα κοινωνικά δίκτυα προσφέρουν «χώρους» ανταλλαγής ιδεών και απόψεων για τους χρήστες, παρέχοντας μία σημαντική πηγή δεδομένων για Ανάλυση Συναισθήματος και Εξόρυξη Γνώμης. Η Εξόρυξη Γνώμης, μπορεί να αναδείξει τη συνολική άποψη των χρηστών για ένα θέμα και να συστήσει προϊόντα ή δραστηριότητες στους χρήστες, είτε βάσει των προτιμήσεών τους, είτε με κριτήριο προηγούμενες επιλογές τους.
Τα τελευταία χρόνια η Ανάλυση Συναισθήματος προσελκύει όλο και περισσότερο το ενδιαφέρον της ακαδημαϊκής κοινότητας, αλλά και των επιχειρήσεων χάρη στις εφαρμογές της, κυρίως στον τομέα της Επιχειρηματικής Ευφυίας (BusinessIntelligence).Η ανάλυση συναισθήματος μπορεί να βοηθήσει σε πολλούς τομείς τις επιχειρήσεις. Πιο συγκεκριμένα μπορεί να βελτιώσει την εξυπηρέτηση πελατών. Στην περίπτωση αυτή η ανάλυση συναισθήματος δίνει καλές πληροφορίες για τις παρούσες και για τις μελλοντικές προτιμήσεις των πελατών, τα θέματα ενδιαφέροντος, τις απόψεις για προϊόντα. Αυτό παράλληλα βοηθάει την επιχείρηση να ενημερωθεί κατάλληλα και να ωφεληθεί τόσο από τα θετικά συναισθήματα όσο και από τα αρνητικά για το προϊόν ή την υπηρεσία.
Επιπλέον, οργανισμοί μπορούν να χρησιμοποιήσουν αυτή την πληροφορία από την ανάλυση συναισθήματος για καλύτερο σχεδιασμό μάρκετινγκ ώστε να βελτιωθεί η φήμη της επιχείρησης, γνωρίζοντας τα συναισθήματα και τις προτιμήσεις των ανταγωνιστών.
Επίσης, επιτρέπει στην επιχείρηση να συγκρίνει τις επιδόσεις της με αυτές των ανταγωνιστών. Μια επιχείρηση τότε, θα μπορεί να προβλέπει την μόδα και να σχεδιάσει κατάλληλα προϊόντα, με στόχο την κορυφή του ανταγωνισμού.
Τέλος, η ανάλυση συναισθήματος παρέχει στις επιχειρήσεις αναλυτικές και διορατικές πληροφορίες για τις προτιμήσεις του κοινού. Έτσι, με σωστή χρήση αυτών, δίνεται η δυνατότητα για νέες επιχειρηματικές κινήσεις. Παρέχει λοιπόν την επιχειρηματική ευφυΐα με την οποία θα παρθούν εύστοχες αποφάσεις για την ανάπτυξη της επιχείρησης. [17]

3.3Κατηγοριοποίηση Συναισθήματος

Στην εργασία αυτή θα χρησιμοποιήσουμε μεθόδους επιβλεπόμενης Μηχανικής Μάθησης για την Ανάλυση Συναισθήματος και θα προσεγγίσουμε το πρόβλημα της Ανάλυσης Συναισθήματος, ως πρόβλημα κατηγοριοποίησης/ταξινόμησης.
Κάθε κατηγορία (κλάση) αντιστοιχεί σε ένα συναίσθημα (πολικότητα).
Η Ανάλυση Συναισθήματος μέσω της κατηγοριοποίησης συναισθήματος, διαφέρει από την κατηγοριοποίηση κειμένου. Η κατηγοριοποίηση/ταξινόμηση κειμένου (ανίχνευση θέματος) αναφέρεται στην αντιστοίχιση κειμένου φυσικής γλώσσας σε θεματικές κατηγορίες, οι οποίες ανήκουν σε ένα προκαθορισμένο σύνολο. Το πλήθος και το είδος των κατηγοριών διαφέρει αναλόγως των στόχων του προβλήματος και πολλές φορές μπορεί ένα κείμενο να αντιστοιχηθεί με μία ή περισσότερες επικαλυπτόμενες κλάσεις.
Αυτό δεν συμβαίνει κατά την κατηγοριοποίηση συναισθήματος.
Η κατηγοριοποίηση συναισθήματος, αναφέρεται σε ένα μικρό σύνολο κατηγοριών:

- θετική πολικότητα – αρνητική πολικότητα
- θετική πολικότητα – αρνητική πολικότητα- χωρίς πολικότητα (ουδέτερη)
- 1 αστέρι - 2 αστέρια - 3 αστέρια - 4 αστέρια -5 αστέρια.

Τα tweets, για παράδειγμα ταξινομούνται ακολούθως σε μια από αυτές τις τρεις κλάσεις, ή όπως συνηθίζεται τις περισσότερες φορές μόνο σε positive/negative [17] μετατρέποντας το πρόβλημα της κατηγοριοποίησης σε δυαδικό. Επίσης, επειδή κατά την κατηγοριοποίηση συναισθήματος επιδιώκεται η ανάλυση της πολικότητας ενός κειμένου, οι κατηγορίες είναι ανεξάρτητες μεταξύ τους και αμοιβαία αποκλειόμενες.

3.4 Δυσκολίες και Προκλήσεις

Η πολικότητα ενός κειμένου εξαρτάται από την πολικότητα των επιμέρους λέξεων που περιέχονται σε αυτό. Αναγνωρίζοντας λοιπόν ένα συγκεκριμένο σύνολο λέξεων κλειδιών (keywords), μπορούμε να ανακαλύψουμε την πολικότητα της άποψης που εκφέρεται μέσα στο κείμενο. Αυτή η διαδικασία ενώ είναι εξαιρετικά αποτελεσματική για τη την σωστή ανίχνευση θέματος, δεν αποτελεί πάντα ακριβής τεχνική για την ανάλυση συναισθήματος. Κάποια από τα συνηθισμένα προβλήματα που συναντάμε κατά την ανάλυση συναισθήματος, μέσω ανίχνευσης πολικότητας μεμονωμένων λέξεων είναι τα εξής:

- Το συναίσθημα/άποψη μπορεί να εκφραστεί έμμεσα. Η απουσία λοιπόν συναισθηματικά φορτισμένων λέξεων καθιστά δύσκολη την αναγνώριση του συναισθήματος από τους μεμονωμένους όρους όταν αυτοί ελέγχονται ξεχωριστά για την εύρεση της πολικότητάς τους. Μέρος της δυσκολίας αυτής είναι και ο διαχωρισμός της αντικειμενικής και της υποκειμενικής γνώμης. [6]

- Ιδιαίτερη δυσκολία παρουσιάζει και ο προσδιορισμός του κατόχου του εκφραστή της άποψης (opinionholder), δηλαδή αν η γνώμη ανήκει στο δημιουργό ή στον σχολιαστή π.χ. στην ανάλυση πολιτικών debates.[6]

- Τέλος η πολικότητα του κειμένου μπορεί να επηρεαστεί και από τη σειρά που είναι τοποθετημένες οι λέξεις ή οι φράσεις, την άρνηση (π.χ. η λέξη «καλός» είναι θετικότερη από το αντώνυμό της, «όχι κακός»), την ειρωνεία, τον σαρκασμό, τις γλωσσολογικές ιδιαιτερότητες.
  Όπως αναφέρεται άλλωστε από τους Kim και Hovy, πολλές φορές ακόμη και άνθρωποι διαφωνούν για το συναίσθημα των κειμένων και για το αν μια δήλωση αποτελεί άποψη ή όχι. Έτσι αποδεικνύεται πόσο δύσκολο είναι το έργο ενός υπολογιστή. [7]

Γενικότερα, όσο πιο μικρό είναι το κείμενο, τόσο πιο δύσκολη γίνεται όλη η διαδικασία. Παρ΄όλα αυτά, η ανάλυση συναισθημάτων, όσον αφορά το microblogging (facebook,twitter), παρουσιάζει το Twitter ως έναν αξιόπιστο online δείκτη του πολιτικού κυρίως συναισθήματος με μελέτες να δείχνουν πως το πολιτικό συναίσθημα των tweets αντικατοπτρίζει πολύ καλά τις πολιτικές θέσεις των πολιτικών και γενικότερα του πολιτικό σκηνικό.
Όμως και το Twitter συναντά δυσκολίες στην ανίχνευση πολικότητας ενόςκειμένου.

3.5 ToTwitter

Το twitter λοιπόν, που ανήκει στην κατηγορία των μικρο-ιστολογίων (microblogs), είναι μια πλατφόρμα κοινωνικής δικτύωσης, όπου οι χρήστεςεπικοινωνούν μεταξύ τους και εκφράζουν απόψεις με τη συγγραφή κειμένων.
Δημιουργήθηκε το 2006 και σήμερα έχει πάνω απο 500 εκατομμύρια χρήστες, εκ των οποίων περίπου 320 εκατομμύρια είναι ενεργοί μηνιαίως και οι οποίοι παράγουν 340 εκατομμύρια μηνύματα την ημέρα, συνολικά. [9]
Τα μηνύματα που επιτρέπεται να δημοσιεύονται, τα λεγόμενα «tweets», είναι μέχρι 280 χαρακτήρων. Οι 280 αυτοί χαρακτήρες εκτός από το κυρίως κείμενο (text) και τους συνδέσμους(URLs), μπορεί να περιέχουν και ειδικά σύμβολα, όπως είναι το «@»,το «#» και το «RT». Το πρώτο, συνδυασμένο με ένα όνομα χρήστη (username) χρησιμοποιείται για ειδική αναφορά σε αυτόν τον χρήστη, ενώ το δεύτερο, «hashtag», δηλώνει σε ποιο θέμα αναφέρεται το συγκεκριμένο tweet και στην ουσία ομαδοποιεί tweets που αναφέρονται στο ίδιο θέμα. Τέλος το «ReTweet» δίνει την ευκαιρία στον χρήστη να αναμεταδώσει ένα μήνυμα ενός άλλου χρήστη και έτσι μπροστά από το μήνυμα μπαίνει το σύμβολο αυτό.
Φαίνεται λοιπόν πως το μέσο αυτό, όπως και γενικότερα τα κοινωνικά δίκτυα, αποτελούν πηγές συναισθήματος και τα δεδομένα που εξάγονται από αυτά είναι μέγιστης σημασίας. Συγκεκριμένα το twitter υπερτερεί σε σχέση με άλλα κοινωνικά δίκτυα, λόγω των tweets, με στόχο τον εντοπισμό και την αποσαφήνιση συναισθήματος για συγκεκριμένο θέμα. Αυτό συμβαίνει γιατί:

- Όλο το συναίσθημα περιέχεται σε κείμενο 280 χαρακτήρων, οπότε είναι εύκολο να ανιχνευθεί σε σχέση για παράδειγμα με το Facebook όπου το συναίσθημα υπάρχει και στο «like» και στο «share».

- Υπάρχει κοινωνικό γράφημα (socialgraph)*με συγκεκριμένη δομή, δηλαδή όταν ένας χρήστης «ακολουθεί» έναν άλλον σημαίνει πως «ακολουθεί» τα tweetstου και αυτοί που παρακολουθούνται από άλλους έχουν τους λεγόμενους «followers».

- Τα περισισσότερα δεδομένα του Twitter είναι ελεύθερα στο κοινό (μέσω του StreamingAPI) και επομένως γίνεται εύκολη η συλλογή τους.

- Τα tweets δείχνουν με χρονολογική σειρά την εξέλιξη των γεγονότων, δηλαδή περιέχουν χρονοσφραγίδα (αλληλουχία καταγεγραμμένων γεγονότων=timestamp).

*το κοινωνικό γράφημα πρόκειται για μια αναπαράσταση του κοινωνικού δικτύου, δηλαδή έναν σφαιρικό χάρτη που δείχνει τον τρόπο σύνδεσης των χρηστών στα κοινωνικά δίκτυα και κάθε χρήστης συμβολίζεται με μια τελεία (κόμβος) και η σχέση του με κάποιον άλλον με μια γραμμή (ακμή).

*πρόκειται για μια Διεπαφή Προγραμματισμού Εφαρμογών (ApplicationProgrammableInterface) από την οποία μπορούμε, ως μέλη του Twitter, να έχουμε εύκολα πρόσβαση στα δεδομένα του.

Όμως η Ανάλυση Συναισθήματος στα κοινωνικά δίκτυα γενικότερα, περιλαμβάνουν και κάποιες δυσκολίες οι οποίες αναλύονται παρακάτω:

- Μήκος Κειμένου: Το μήκος του κειμένου στα socialmedia προκαλεί περιορισμούς στην Ανάλυση Συναισθήματος, εξαιτίας της μικρής έκτασης (π.χ. το μέγιστο όριο χαρακτήρων στο twitter που φτάνει τους 280) η οποία μπορεί να αφήσει ανακρίβειες και να μην γίνεται σαφής η πολικότητα του κειμένου. [10]

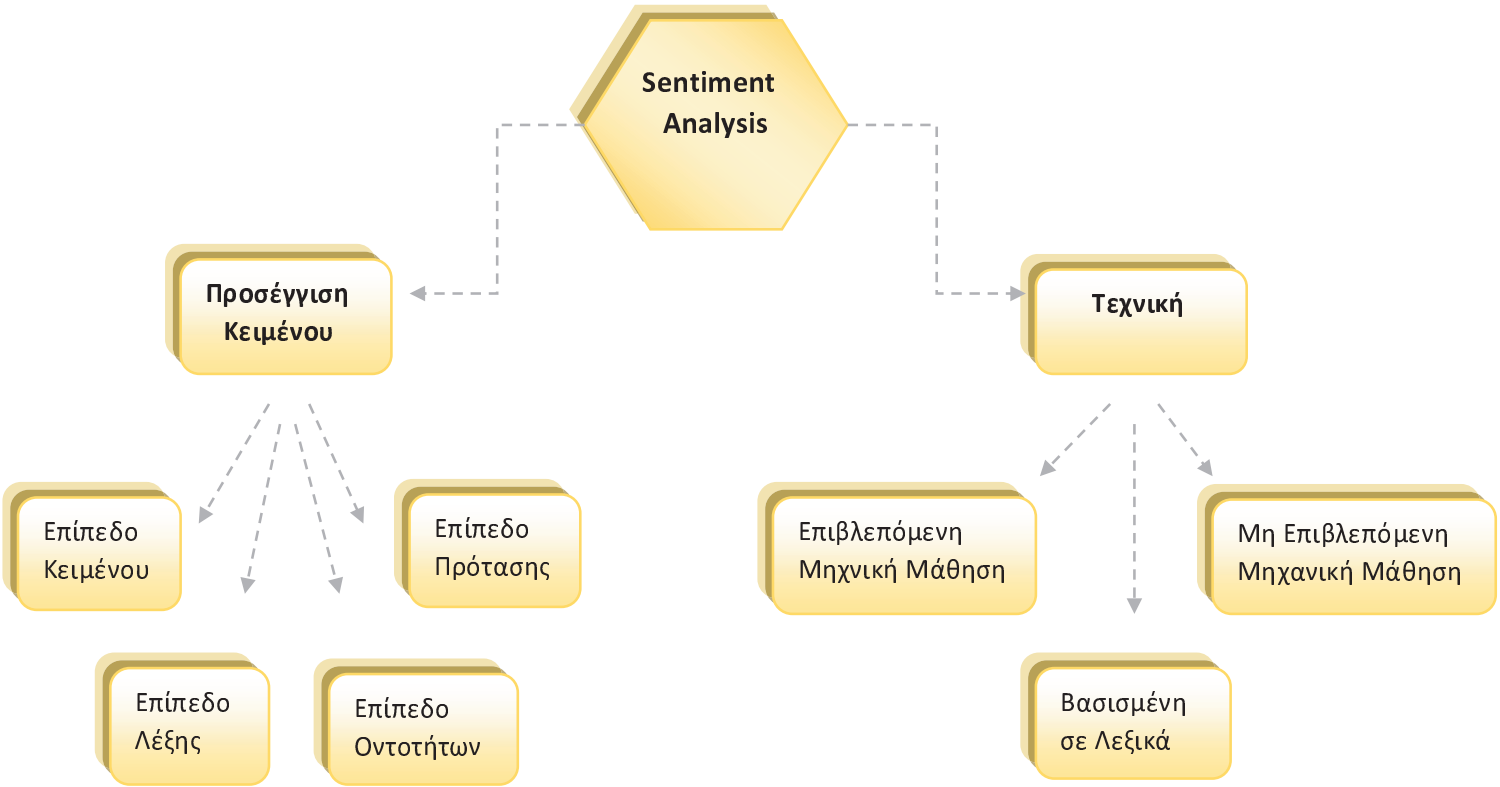
- Λεξιλόγιο: Η έντονη χρήση της γλώσσας αργκό, των νεολογισμών και συντομογραφιών που πλέον χρησιμοποιούνται όλο και περισσότερο (π.χ. thx= thanks, diy=doityourself) μειώνει τις πιθανότητες για ορθή εξαγωγή συμπερασμάτων για τα συναισθήματα του χρήστη μέσω της Εξόρυξης Γνώμης.

- Θόρυβος: Τα συντακτικά, γραμματικά και ορθογραφικά λάθη που κάνουν οι χρήστες καθιστούν δύσκολη την αποσαφήνιση των συναισθημάτων, κάθε κειμένου.

- Χρήση διαφορετικών γλωσσών στο ίδιο κείμενο: Τα κοινωνικά δίκτυα ειναι διεθνή, με κυρίαρχη γλώσσα να επικρατεί να είναι η αγγλική. Αυτό σημαίνει πως οι χρήστες την χρησιμοποιούν όλο και περισσότερο σε συνδυασμό με άλλες γλώσσες, συνήθως της μητρικής τους, στο ίδιο κείμενο. Αυτή η μίξη έχει ως επίπτωση τη δημιουργία επιπλέον εμποδίων στα συστήματα ανίχνευσης συναισθήματος.

3.6 Κατηγοριοποίηση Προσέγγισης Κειμένου

Οι προσεγγίσεις του προβλήματος της Ανάλυσης Συναισθήματος διαφοροποιούνται ως προς το επίπεδο ανάλυσης.Για να πραγματοποιηθεί η εξαγωγή συναισθήματος, υπάρχουν κάποιες τεχνικέςεπεξεργασίας φυσικής γλώσσας και κάποιοι αλγόριθμοι που μπορούν νακατηγοριοποιηθούν με βάση τον τρόπο που προσεγγίζουμε το κείμενο, αλλά και τηνταξινόμηση συναισθήματος που θέλουμε να κάνουμε. Ορίζονται λοιπόν κάποιοι τρόποικατηγοριοποίησης της ανάλυσης συναισθήματος. Επίσης, αναλόγως με τη τεχνική και τοβαθμό που παρεμβαίνει ο άνθρωπος στη διαδικασία, καθορίζονται επιπλέον κάποιεςκατηγορίες.
Τα επίπεδα προσέγγισης κειμένου και οι κατηγορίες των τεχνικών παρουσιάζονται εικονικά στο παρακάτω σχήμα και θα αναλυθούν στη συνέχεια.



Εικόνα 2: Κατηγοριοποίηση Ανάλυσης Συναισθήματος

3.6.1 Ταξινόμηση σε επίπεδο εγγράφου/κειμένου

Αυτή η ταξινόμηση επικεντρώνεται στον προσδορισμό της υποκειμενικής θέσης ενός και μόνο ατόμου, δηλαδή θετική ή αρνητική, ως προς το θέμα που αναλύεται στο κείμενο. Για την διαδικασία αυτή επιλέγονται κείμενα με απόψεις και κρίσεις, εμφανίζοντας πρακτικές δυσκολίες στις περιπτώσεις όπου το αντκείμενο σχολιασμού είναι παραπάνω από ένα (π.χ. σύγκριση δύο προϊόντων). Μέρος της ταξινόμησης σε επίπεδου κειμένου είναιη γραμματική και συντακτική ανάλυση του κειμένου (PartofSpeechtagging–POS*) και σημαντικό ρόλο παίζουν οι συντακτικές σχέσεις και το φαινόμενο της άρνησης (π.χ. «όχι καλός»), το οποίο πολλές φορές δίνει λανθασμένα αποτελέσματα, όπως στην περίπτωση αναπαράστασης του κειμένου ως σύνολο λέξεων, όπως θα αναφέρουμε λεπτομερειακά στη συνέχεια. [12]

3.6.2 Ταξινόμηση σε επίπεδο πρότασης

Η ανάλυση σε αυτό το επίπεδο ασχολείται με την πρόταση και τον προσδιορισμό θετικής, αρνητικής ή ουδέτερης στάσης που εκφράζει. Έχοντας ως δεδομένο ότι υπάρχει μόνο μια άποψη σε κάθε πρόταση, οι προτάσεις μπορούν να χαρακτηριστούν απευθείας ως θετικές ή αρνητικές. Στόχος αυτής της προσέγγισης είναι ο διαχωρισμός των «αντικειμενικών» προτάσεων από εκείνων με «υποκειμενικό» χαρακτήρα και στη συνέχεια ταξινομεί τις δεύτερες ως θετικές ή αρνητικές. Και εδώ λαμβάνονται υπόψην η γραμματική και συντακτική ανάλυση των λέξεων της πρότασης, το φαινόμενο της άρνησης και η σημασιολογία-αμφισημία των λέξεων (π.χ η λέξη «python» μπορεί να αναφέρεται ή στο γνωστό φίδι ή στη γλώσσα προγραμματισμου). [15] [16]

3.6.3 Ταξινόμηση σε επίπεδο λέξης

Το επίπεδο αυτό χρησιμοποιείται για την ταξινόμηση επιπέδου πρότασης ή κειμένου με γνώμονα τις λέξεις γνώμης (opinionwords), θεωρώντας τες ως τους πιο σημαντικούς δείκτες συναισθημάτων [14]. Μια λίστα από τέτοιες λέξεις ονομάζεται «λεξικό συναισθημάτων». Τέτοια λεξικά προκύπτουν από την επεξεργασία είτε μεγάλων σωμάτων ηλεκτρονικών κειμένων (textcorpora)*, είτε γλωσσολογικών πόρων (π.χ. λεξικά), προκειμένου να επεκταθεί μια αρχική λίστα με λέξεις γνώμης (seedwords).Στην περίπτωση που τα λεξικά δημιουργούνται από σώματα κειμένου (textcorpora), η επέκταση της λίστας επιτυγχάνεται με χρήση συντακτικών μοτίβων τα οποία ικανοποιούνται μέσα σε αυτά τα κείμενα ή και με χρήση πληροφοριών που προκύπτουν από τη συχνότητα δάφορων λεξικών μοτίβων. [16] Στην περίπτωση που τα λεξικά βασίζονται σε γλωσσολογικούς πόρους, η επέκταση της λίστας επιτυγχάνεται με συνώνυμα, αντώνυμα και την ιεραρχία των λέξεων μέσα σε γλωσσολογικούς θησαυρούς, όπως είναι το WordNet.

3.6.4 Ταξινόμηση σε επίπεδο οντότητας και χαρακτηριστικών

Η ταξινόμηση αυτού του επιπέδου αφορά την ίδια την άποψη και όχι την ανάλυση δομικών στοιχείων της γλώσσας (κείμενο, πρόταση, φράση).
Πολλές φορές χρειαζόμαστε μια λεπτομερή ανάλυση, ώστε να διαχωρίζονται όλα τα χαρακτηριστικά στα οποία εναφέρεται μια άποψη (θετική ή αρνητική), όπως είναι η εύρεση του συναισθηματικού τους φορτίου. Αυτό πολλές φορές δεν είναι δυνατόν να επιτευχθεί με την ταξινόμηση σε επίπεδο κειμένου ή πρότασης, καθώς δεν γίνεται ούτε να εντοπιστούν οι μεταβλητές (opiniontargets) στις οποίες αναφέρεται μια γνώμη, ούτε και συνεπώς να προσδιοριστεί σε κάθε μια από αυτές ένα ξεχωριστό συναίσθημα. Μια άλλη δυσκολία σε αυτά τα επίπεδα είναι πως στην περίπτωση που ένα υποκείμενο έχει μια άποψη (θετική ή αρνητική) για μια οντότητα, δεν μπορούμε να γνωρίζουμε αν θα έχει την ίδια άποψη για κάθε μεμονωμένο χαρακτηριστικό της.[20]
Η συγκεκριμένη λοιπόν ταξινόμηση στηρίζεται στο γεγονός ότι μια άποψη αποτελείται από ένα συναίσθημα (sentiment) και έναν στόχο (target) στον οποίο απευθύνεται και συνήθως αναπαριστάταιμέσω οντοτήτων (entities). Σκοπός της ανάλυσης αυτής, λοιπόν, είναι να αναζητά τις απόψεις προς τα αντικείμενα-στόχους, καθώς και τα επιμέρους χαρακτηριστικά τους (aspects). Επίσης, ρόλο σε αυτήν την ανάλυση παίζουν και άλλα στοιχεία, όπως το πρόσωπο που εκφράζει την άποψη (opinionholder) και ο χρόνος έκφρασης (time). Για την ανάλυση του επιπέδου οντότητας και χαρακτηριστικών χρειάζεται η κατηγοριοποίηση των χαρακτηριστικών (featuresentimentclassification) σε θετικό-αρνητικό-ουδέτερο, που επιτυγχάνεται κυρίως με τεχνικές επιβλεπόμενης μηχανικής μάθησης και τη δημιουργία λεξικών πόρων.[13]

3.7 Εισαγωγή ΣτιςΤεχνικές Ανάλυσης Συναισθήματος

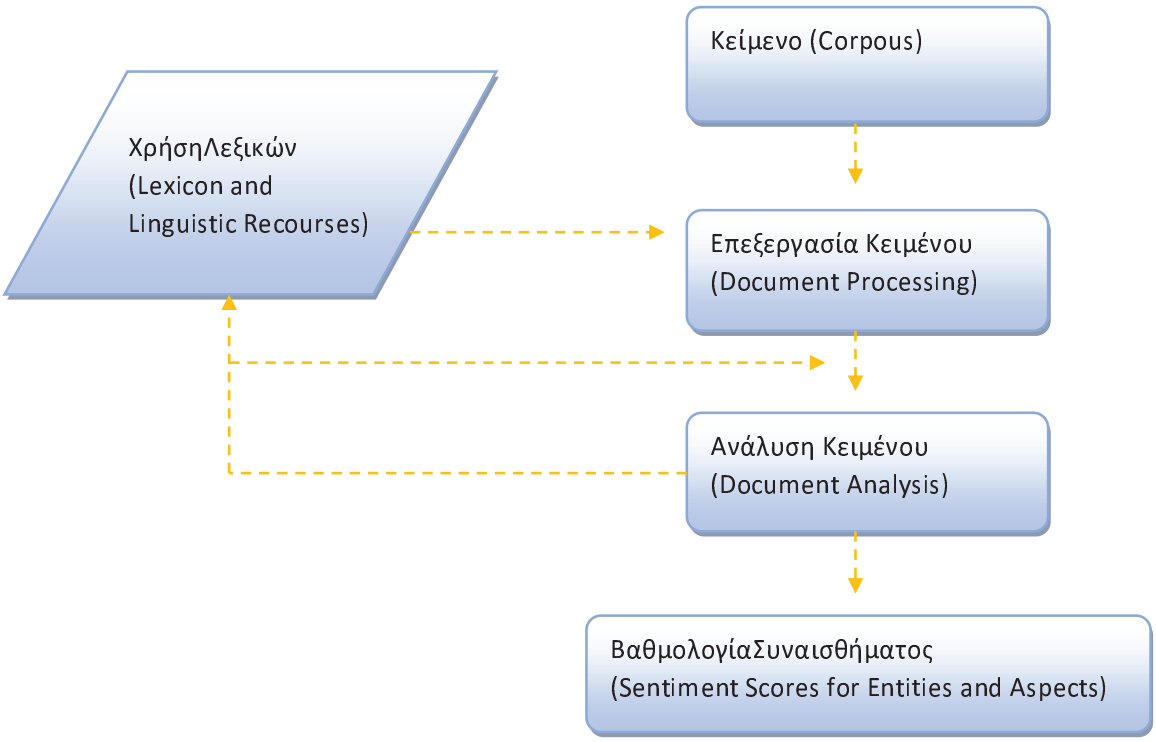
Με βάση την τεχνική που χρησιμοποιούμε, προσδιορίζονται κάποιες κατηγορίες για τηνανάλυση συναισθήματος. Οι κυριότερες από αυτές, όπως φαίνεται και στο παραπάνω σχήμα, είναιοι τεχνικές με Επιβλεπόμενη Μηχανική Μάθηση, Μη Επιβλεπόμενη Μηχανική Μάθηση, καιοι τεχνικές Βασισμένες σε Λεξικά, καθώς και ο συνδυασμός αυτών. Εμείς θα ασχοληθούμε με τις τεχνικές με Λεξικά και με Επιβλεπόμενη Μηχανική Μάθησηκαι θα αναλυθούνσταεπόμενακεφάλαια.

*Η PartofSpeechTagging είναι μια διαδικασία μετατροπής μιας πρότασης από μια λίστα λέξεων σε λίστα πλειάδων. Κάθε πλειάδα είναι της μορφής (λέξη, ετικέτα) και κάθε λέξη του κειμένου σημειώνεται ως προς τοτι μέρος του λόγου είναι (ουσιαστικό, ρήμα, υποκείμενο κ.τ.λ).

*Στη γλωσσολογίατο corpustext είναι ένα μεγάλο και δομημένο σύνολο κειμένων (σήμερα συνήθως αποθηκεύεται και επεξεργάζεται ηλεκτρονικά). Χρησιμοποιούνται για στατιστικές αναλύσεις και δοκιμές υποθέσεων, ελέγχοντας τα περιστατικά ή γλωσσικούς κανόνες σε μια συγκεκριμένη γλωσσική επικράτεια.

4 ΤΕΧΝΙΚΕΣ ΜΕ ΛΕΞΙΚΑ

Οι τεχνικές βασισμένες σε λεξικά, χρησιμοποιούν έτοιμα λεξικά συναισθήματος τα οποία περιέχουν διάφορους όρους του κειμένου, οι οποίοι χαρακτηρίζονται και έτσι προκύπτει η συνολική πολικότητα.Οι τεχνικές αυτές παρουσιάζουν αρκετά πλεονεκτήματα στη χρήση τους, αφού παρουσιάζουν αποτελέσματα υψηλής ακρίβειας (στις περιπτώσεις που εφαρμόζονται σε θέματα όπου τα λεξικά περιέχουν το λεξιλόγιο που χρησιμοποείται από το εκάστοτε κείμενο) και καλύπτουν μεγάλη ποικιλία θεμάτων, καθώς δεν χρειάζονται τα trainingsets. Με τις τεχνικές αυτές, όπου γίνεται χρήση λεξικών, διαχειριζόμαστε το κείμενο ως ένα σύνολο από ανεξάρτητες μεταξύ τους λέξεις χωρίς να μας ενδιαφέρει η σύνταξη, η γραμματική ή σειρά τους (bagofwords).
Για την ανάλυση συναισθήματος των λέξεων που περιέχουν συναίσθημα (sentimentwords) με λεξικά, απαραίτητο βήμα είναι και η απόδοση βαθμολογίας-ετικέτας σε κάθε λέξη, που δείχνουν κατά πόσο το νόημα της λέξης ταιριάζει σε συγκεκριμένες κατηγορίες με συνηθέστερες το θετικό, αρνητικό και ουδέτερο συναίσθημα. Βέβαια υπάρχουν και λεξικά με περισσότερες κατηγορίες (χαρά, ενθουσιασμός, λύπη κ.τ.λ.) και διαβαθμίσεις (πολύ θετικό, πολύ αρνητικό), με αρκετές περιπτώσεις να είναι αυτές που λαμβάνουν υπ'όψην τους τα βαθμωτά επίθετα (συγκριτικού-υπερθετικού βαθμού).[21]Κάθε λέξη από το κείμενο,αναζητείται και αντιστοιχείται με εκείνη του λεξικού και κρατείται ηβαθμολογία της. Τελικά, το συνολικό συναίσθημα του κειμένου προσδιορίζεται από τοάθροισμα των βαθμολογιών των επιμέρους λέξεων.
Εδώ παρουσιάζεται σχηματικά η διαδικασία ανάλυσης συναισθήματος με χρήσηλεξικών:



Εικόνα 3: Σύστημα Ανάλυσης Συναισθήματος με χρήση λεξικού

4.1 Δημιουργία Λεξικών

Οι μέθοδοι που είναι βασισμένες σε λεξικά αποτελούν τον καλύτερο τρόπο για την εκτέλεση της μη επιβλεπόμενης μηχανικής μάθησης. Εδώ, όπως προαναφέραμε, δημιουργούνται λεξικά που περιέχουν λέξεις (προκαθορισμένης σημασιολογίας), οι οποίες αντιπροσωπεύουν θετική ή αρνητική άποψη (opinionwords).
Ορίζονται τρεις βασικές μέθοδοι για τη δημιουργία τέτοιων λεξικών:

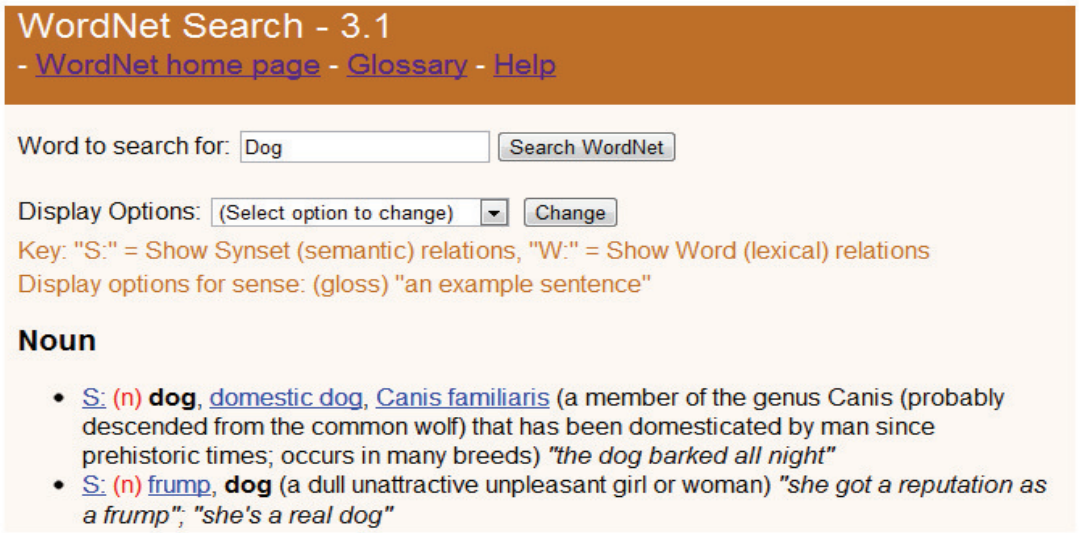
- Οι μέθοδοι βασισμένες σε λεξικά (disctionary-basedmeyhods).Αυτές χρησιμοποιούν λεξικά (π.χ. WordNet) για να καθορίσουν την πολικότητα της λέξης από άλλες παρόμοιες, σε σημασιολογικό/γλωσσικό επίπεδο, λέξεις. Επομένως, ο συναισθηματικός προσανατολισμός των λέξεων αυτών θα είναι γνωστός και η διαδικασία επιλογής τέτοιων λέξεων ονομάζεται bootstrapping, με μοναδικό μειονέκτημα αυτό των σημασιολογικών εννοιών μιας λέξης που δυσκολεύουν τον εντοπισμό τους σε κάθε χρήση της.
- Οι μέθοδοι των σωμάτων κειμένων (corpus-basedmethods).Αυτές βρίσκουν την πολικότητα της λέξης από ένα συγκεκριμένο ηλεκτρονικό σώμα (corpus), μελετώντας τη σχέση ανάμεσα στις λέξεις και κάποια λέξη συναισθήματος (seed). Έτσι, με βάση το συντακτικό και τα συμφραζόμενα δημιουργείται το λεξικό γύρω από αυτή τη λέξη, το οποίο εμπλουτίζεται μετά με σχετικές λέξεις όσον αφορά τη σημασιολογία και τη συναισθηματική πολικότητα.
- Η μη αυτοποιημένη αναζήτηση και συλλογή όρων.Είναι χρονοβόρα, αλλά συνδυάζεται με τις παραπάνω και διορθώνει τυχόν λάθηαυτών. Το λεξικό συναισθήματος εδώ,δημιουργείται με τη χειροκίνητη εισαγωγή λέξεων.

4.2 Λεξικά

4.2.1 WordNet

Το WordNet είναι μια λεξικολογική βάση δεδομένων για την αγγλική γλώσσα, το οποίο δημιουργήθηκε το 1986. Κατηγοριοποιεί τα ουσιαστικά, ταρήματα, τα επίθετα και τα επιρρήματα σε ομάδες συνωνύμων (synsets) ή ιεραρχίας. Περιέχει ορισμούς, παραδείγματα χρήσης, έναν αριθμό από σχέσεις μεταξύ συνωνύμων (π.χ. hypernyms) και έναν αριθμό από έννοιες. [22]
Η έννοια μιας λέξης του WordNet αποτελείται από:

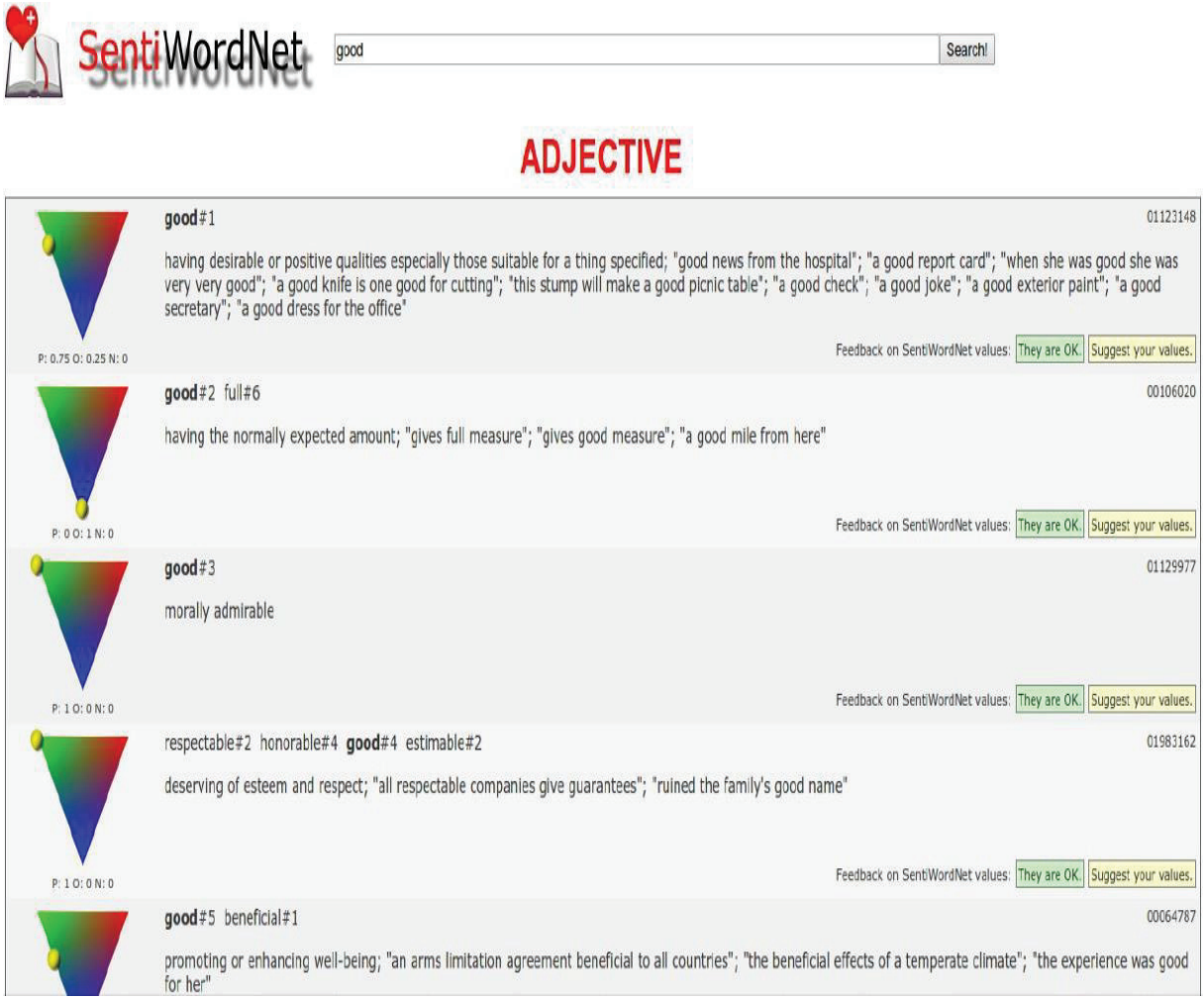
- έναν αριθμό πουσυμβολίζει τη συχνότητα εμφάνισης του όρου με τη συγκεκριμένη έννοια και έτσι μπορούμε να ανακαλύψουμε την πιο δημοφιλή έννοια για κάθε λέξη (MostFrequentSense ή FirstSense).Πολλές φορές μια τέτοια πληροφορία για μια λέξη βοηθάει στηναποσαφήνισή της και αφού μελετηθεί και ο αριθμός εμφάνισής της στο κείμενο, διευκρινίζεται και η πολικότητα του κειμένου (π.χ. με τη λέξη «bad»).
- ένα σύνολο συνωνύμων (synsets) της συγκεκριμένης ερμηνείας της λέξης.
- ένα σύνολο από φράσεις της καθομιλουμένης που περιέχουν τη λέξη με τη συγκεκριμένηέννοια.



Εικόνα 4: Παράδειγμα λειτουργίας WordNet

4.2.2 SentiWordNet

Το SentiWordNet που αναπτύχθηκε το 2010 χρησιμοποιεί κυρίως τα σύνολα συνωνύμων (synsets) του WordNet. Είναι αποτέλεσμα της αυτόματης επισήμανσης των synsets του WordNet σύμφωνα με τις βαθμολογίες: θετικό, αρνητικό, ουδέτερο που περιγράφουν το συναίσθημα τωνόρων που το αποτελούν. Τασύνολα συνωνύμων (synsets) συμβάλλουν, ως τα βασικά στοιχεία του λεξικού, στο να ανακαλυφθούν όλες οι διαφορετικές έννοιες του ίδιου όρου, δυνατότητα που δεν θα μας έδινε η χρήση των ίδων των όρων. Οι τιμές των βαθμολογιών για κάθε συνόλο συνωνύμων ανήκουν στο διάστημα [0.0, 1.0] και τοάθροισμα τους πρέπει να είναι πάντα ίσο με 1. [23]



Εικόνα 4: ΠαράδειγμαλειτουργίαςSentiWordNet

4.2.3 LinguisticInquiryandWordCount

Το LIWC λεξικό δημιουργήθηκε το 2007 και αποτελεί τη βάση της διαδικασίας της ανάλυσης κειμένου, υποστηρίζοντας 82 γλώσσες. Το συνθέτουν 4.500 λέξεις και ρίζες αυτών και κάθε μία ορίζει κατηγορίες (συναίσθημα, μέρος του λόγου κτλ). Για παράδειγμα, η λέξη «δάκρυ» ανήκει σε 3 λεκτικές κατηγορίες: λύπη, αρνητικό συναίσθημα και ουσιαστικό, οι οποίες κατηγοριοποιούνται ιεραρχικά. Κάθε λέξη εξετάζεται μεμονωμένα και όταν βρεθεί εκείνη που αντιστοιχεί σε κάποια κατηγορία του λεξικού, τότε αυξάνεται κατά 1 ο μετρητής της κατηγορίας αυτής.Έπειτα, το λεξικό δημιουργεί και προσθέτει λήμματα (WordCollection). Στη συνέχεια, γίνεται βαθμονόμηση των αξιολογητών, δηλαδή η βαθμολόγηση και αξιολόγηση κάθε κατηγορίας από τρεις κριτές,οι οποίοι κρίνουν αν μια λέξη θα αφαιρεθεί ή θα προστεθεί σε μια κατηγορία.Ακολουθεί η ψυχομετρική αξιολόγηση, όπου διαγράφονται ή αντικαθίστανται οι λιγότερο συχνές κατηγορίες με νέες, ανάλογα με την ανάλυση των κειμένων.
Τέλος, γίνονται προσθήκες, από το γραπτό ή προφορικό λόγο, με αποτέλεσμα να αλλάζει η δομή του λεξικού.[24]

LIWC

TRADITIONAL LIWC DIMENSION	YOUR DATA
I-WORDS (I, ME, MY)	0.6
SOCIAL WORDS	11.8
POSITIVE EMOTIONS	4.2
NEGATIVE EMOTIONS	1.6
COGNITIVE PROCESSES	9.8
SUMMARY VARIABLES	
ANALYTIC	77.1
CLOUT	89.8
AUTHENTICITY	23.1
EMOTIONAL TONE	73.9

Εικόνα 5: Παράδειγμα λειτουργίας LinguisticInquiryandWordCount

4.3 Προβλήματα των μεθόδων με λεξικά και ανάγκη για άλλες τεχνικές

Οι μέθοδοι όμως αυτοί, παρ'όλα τα πλεονεκτήματα που εμφανίζουν, πολλές φορές μπορεί να μην μας δώσουν τα επιθυμητά αποτελέσματα.

- Αυτό αρχικά, μπορεί να συμβεί αν δεν λάβουμε υπ'όψην μας το φαινόμενο της άρνησης, την ειρωνεία, την ένταση των λέξεων, το θέμα, τη σειρά των λέξεων και τους ιδιωματισμούς της γλώσσας. Για παράδειγμα:
  1) στην άρνηση άλλο αποτέλεσμα δίνει η λέξη «good» και άλλο η φράση «notgood».
  2) όσον αφορά τη σειρά λέξεων οι φράσεις «Youhavetakentherightdecision, heisnotagoodperson» και « Youhaven'ttakentherightdecision, heisagoodperson» δίνουν διαφορετικές έννοιες και το bagofwords εδώ θα δώσει λανθασμένα αποτελέσματα.
- Επίσης επειδή το πλήθος των λέξεων στα λεξικά είναι περιορισμένο, οι νεολογισμοί και οι συντομογραφίες που εμφανίζονται, όπως στο Twitter, δεν αποδίδονται εύκολα.
- Τέλος, το συναίσθημα που έχει προσδιοριστεί στις λέξεις από τα λεξικά είναι συγκεκριμένο και σταθερό, χωρίς να λαμβάνεται υπ'όψην το γενικότερο πλαίσιο μέσα στο οποίο χρησιμοποιούνται και έτσι οδηγούμαστε σε λανθασμένα αποτελέσματα. Δηλαδή, στη λέξη «great» πρέπει να προσδίδεται θετική ετικέτα όταν αναφέρεται στη λέξη «party», ενώ αρνητική όταν αναφέρεται στη λέξη «problem».

4.4 Εισαγωγή στη Μηχανική Μάθηση

Συμπεραίνουμε από τα παραπάνω πως είναι δύσκολο να δημιουργηθεί ένα ολοκληρωμένο λεξικό γνώμης και συνεπώς, σε πολλές περιπτώσεις με ελλιπή λεξικά, η ανάλυση συναισθήματος δεν δίνει ορθά αποτελέσματα. Δημιουργήθηκε λοιπόν η ανάγκη για χρήση και άλλων μεθόδων, όπως είναι η επιβλεπόμενη και η μη επιβλεπόμενη μηχανική μάθηση και αλγορίθμων που βοηθάνε στην εκπαίδευση δεδομένων και στην ανάλυση πολικότητας.

5 ΤΕΧΝΙΚΕΣΕΠΙΒΛΕΠΟΜΕΝΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

5.1 Εισαγωγή

Για τη μελέτη του προβλήματος της κατηγοριοποίησης συναισθήματος, θα χησιμοποιηθούν στατιστικές μέθοδοι που περιλαμβάνουν μοντέλα Bayes και Μηχανές Διανυσμάτων Υποστήριξης. Η προσέγγιση του προβλήματος με στατιστικές μεθόδους είναι η πιο γνωστή στην Ανάλυση Συναισθήματος και χρησιμοποιεί αλγορίθμους Μηχανικής Μάθησης (MachineLearning), σε συνδυασμό με κείμενα, που έχουν κατηγοριοποιηθεί χειροκίνητα, ώστε να εκπαιδεύσει μία μηχανή, για να μπορεί να ανακαλύψει την πολικότητα νέων κειμένων.

5.2 Ταξινομητές-Μοντέλα Επιβλεπόμενης Μάθησης(NB,SVM,ME)

Ένας ταξινομητής (classifier) είναι ένα μαθηματικό εργαλείο το οποίο, με τη βοήθεια των χαρακτηριστικών, ταξινομεί μια είσοδο, δηλαδή της αναθέτει μία ετικέτα/label. Παρ'όλο που υπάρχουν στη διαθεσή μας πολλοί αποτελεσματικοί αλγόριθμοι επιβλεπόμενης μάθησης, κανένας από αυτούς δεν είναι το ίδιο αποδοτικός σε όλα τα προβλήματα επιβλεπόμενης μάθησης.

- ***NaiveBayes***

Ο αλγόριθμος*NaiveBayes*, όπως και όλοι οι naïveBayesclassifiers, βασίζεται στην εφαρμοφή του θεωρήματος Bayes και αποτελεί την πιοαπλή (λόγω μικρού trainingset) τεχνική για την κατασκευή ταξινομητών, δηλαδή μοντέλων πουαναθέτουν ετικέτες κλάσης σε οντότητες προβλημάτων που αναπαρίστανται από διανύσματα με τιμές χαρακτηριστικών (τα χαρακτηριστικά αφορούν κυρίως τις συχνότητες εμφάνισης λέξεων), όπου οι ετικέτες αυτές καθορίζονται από ένα πεπερασμένο σύνολο.Δεν είναι ένας μεμονωμένος αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μιαοικογένεια αλγορίθμων που βασίζεται σε ένα κοινό πρότυπο, με βασική υπόθεση αυτή της ανεξαρτησίας των μεταβλητών των χαρακτηριστικών μιας κλάσης.
Στα μοντέλα Bayes, μετά την εκπαίδευσή τους για να μπορέσει να γίνει η αρχικοποίηση των παραμέτρων τους, υπολογίζεται η πιθανότητα κάθε κατηγορίας, με βάση τα δεδομένα και με χρήση του κανόνα Bayes, στην οποία θα ανήκει ένα νέο άγνωστο στιγμιότυπο του προβλήματος.

Γενικά, ο*NaïveBayes* είναι ένα μοντέλο δεσμευμένης πιθανότητας που ορίζεται ως εξής:
Έστω ότι έχουμε δύο τυχαίες μεταβλητές A,B, με τον κανόνα του Bayes ορίζεται η πιθανότητα να συμβεί το γεγονός A δεδομένου ότι έχει συμβεί το B (εκ των υστέρων πιθανότητα του A), δηλαδή η δεσμευμένη πιθανότητα του A δοθέντος του B, P(A\B) που υπολογίζεται ως:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$[2]

- P(B\A) η πιθανότητα του B δεδομένου του A να είναι αληθής.
- P(A) και P(B) οι (εκ των προτέρων) πιθανότητες του A και του B αντίστοιχα, που είναι ανεξάρτητες μεταξύ τους.

Όσον αφορά το πρόβλημα της κατηγοριοποίησης συναισθημάτων, η εκ των υστέρων αυτή πιθανότητα ανάθεσης μιας κατηγορίας σε ένα στιγμιότυπο, με γνωστές τις τιμές των featurestου, υπολογίζεται σύμφωνα με τον κανόνα του Bayesαπό τον τύπο

$$P(c|x_1,x_2,\ldots,x_n)= \frac{P(x1,x2,\ldots,xn|c)\,P(c)}{P(x1,x2,\ldots,xn)}$$[2]

Όπου:

- c είναι η κατηγορία και παίρνει μία διακριτή τιμή από το πεπερασμένο σύνολο C των δυνατών κατηγοριών που υπάρχουν.
- $\{x_1,x_2,\ldots,x_n\}$ είναι το διάνυσμα των χαρακτηριστικών x.
- P(c) η εκ των προτέρων πιθανότητα κάθε κατηγορίας.
- $P(x_1,x_2,\ldots,x_n)$ η πιθανότητα εμφάνισης των δεδομένων.
- Η πιθανότητα $P(x_1,x_2,\ldots,x_n|c)$ υπολογίζεται εύκολα, αφού κάθε χαρακτηριστικό (τυχαία ματαβλητή) θεωρείται ανεξάρτητο οποιουδήποτε άλλου στην ίδια κατηγορία, όπως δηλώνει άλλωστε και η έννοια της «αφελούς» υπόθεσης («naïve»).
  Άρα:
  $P(x_1,x_2,\ldots,x_n|c)= P(x_1|c)\,P(x_2|c,x_1)\,\ldots\,P(x_n|c,x_1,x_2,\ldots,x_{n-1})$

Επομένως απλοποιείται και η εκ των υστέρων πιθανότητα κάθε χαρακτηριστικούx$_i$και άρα:

$$P(c|x_1,x_2,\ldots,x_n) = \frac{P(X_1,X_2,\ldots,X_n|c)\,P(c)}{P(X_1,X_2,\ldots,X_n)} = \frac{P(c)\,P(x_1|c)\,P(x_2|c,x1)\,\ldots\,P(x_n|c,x_1,x_2,\ldots,x_{n-1})}{P(X_1,X_2,\ldots,X_n)} =$$

$$= \frac{P(c)\,P(x_1|c)\,P(x_2|c)\,\ldots\,P(x_n|c)}{P(X_1,X_2,\ldots,X_n)} = \frac{P(c)\prod_{i=1}^{n}P(x_i|c)}{P(X_1,X_2,\ldots,X_n)}$$

Με $\prod_{i=1}^{n} P(xi|c)$ η πιθανοφάνεια των δεδομένων (x) της κατηγορίας, όπου $P(x_i|c)$υπολογίζεται μέσω της κατανομής των πιθανοτήτων των x_i.

Αφού υπολογιστούν ξεχωριστά οι πιθανότητες να ανήκει το νέο στιγμιότυπο σε μια γνωστήκατηγορία, ο αλγόριθμος δίνει ως κατηγορία διάταξης, αυτή με τη μεγαλύτερη πιθανότητα.Στόχος είναι να επιλέγεται κάθε φορά η πιο πιθανή ετικέτα για είσοδο, με γνωστή την εκ των προτέρων πιθανότητα κάθε ετικέτας και αναλόγως των χακτηριστικών υπολογίζεται η πιθανότητα να προσδιοριστεί με την συγκεκριμένη ετικέτα ή με κάποια διαφορετική.

Ο υπολογισμός των πιθανοτήτων των x_i δεδομένης της κατηγορίας c, $P(x_i|c)$, διακρίνεται σε δύο περιπτώσεις και υπολογίζεται ως εξής:

- Στην περίπτωση που τα χαρακτηριστικά των στιγμιότυπων λαμβάνουν συνεχείς τιμές, υλοποείται το μοντέλο***GaussianNaïveBayes***όπου η πιθανότητα $P(x_i|c)$ υπολογίζεται με την κανονική κατανομή. Στην διάρκεια της εκπαίδευσης αφού επιλέξουμε για την κατηγορία εκείνα τα στιγμιότυπα που έχουν επιλεγεί χειροκίνητα να ανήκουν σε αυτή, υπολογίζουμε για κάθε ένα χαρακτηριστικό μιας κατηγορίας το μέσο όρων των τιμών του (μ_c) και τη διασπορά του (σ_c^2) για να προκύψει η κανονική κατανομή του άρα και η πιθανότητά του μέσω του τύπου:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\exp-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}$$

- Στην περίπτωση που τα χαρακτηριστικά των στιγμιότυπων λαμβάνουν διακριτές τιμές (δηλαδή αναπαριστούν συχνότητες $\{x_1,x_2,\ldots,x_n\}$ γεγονότων {1,2,..,n}), με πιθανότητες να ανήκουν στην κατηγορία c $\{p_{c1},p_{c2},\ldots,p_{cn}\}$, υλοποείται το μοντέλο ***MultinomialNaiveBayes*** όπου η κατανομή είναι πολυωνυμική.Τα $P(x_i|c)$ υπολογίζονται, από το trainingset, ως η συχνότητα εμφάνισης του γεγονότος i στην κατηγορία c ως εξής:

  $$P(x_i|c) = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

  Αν ένας όρος δεν υπάρχει στο trainingsetή σε κάποιο στιγμιότυποτης κατηγορίας, τότε μηδενίζεται η τελική πιθανότητα$P(c|x_1,x_2,\ldots,x_n)$ανάθεσης της κατηγορία αυτής στο στιγμιότυπο, που περιέχει τον όρο, πρόβλημα που αντιμετωπίζεται με την προσθήκη μίας τιμής σε όλες τις πιθανότητες, ώστε να μην πραγματοποιείται αυτός ο μηδενισμός. [36]

- ***SupportVectorMachines***

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVMs) είναι μια μέθοδος κατηγοριοποίησης που ανήκει στις μηχανές εκμάθησης (learningmachines) για επεξεργασία δεδομένων, περισσότερο αποδοτική στην κατηγοριοποίηση εικόνων. Πρόκειται για έναν μη πιθανοτικό (δυαδικό) ταξινομητή, καθώς κατηγοριοποιεί κάθε στοιχείο σε μία από τις δύο κλάσεις. Τα δεδομένα αναπαριστώνται με σημεία στο χώρο και τοποθετούνται έτσι ώστε να υπάρχει όσο το δυνατό μεγαλύτερο διαχωριστικό κενό ανάμεσα στα δεδομένα από διαφορετικές κατηγορίες. Η σωστή επιλογή της κατηγορίας από την αρχή, επιτυγχάνεται με τον εντοπισμό του σημείου στο μέρος που την υποδηλώνει. Σκοπός είναι να κατασκευαστεί ένα υπερεπίπεδο (hyperplane)* που να διαχωρίζει τα δεδομένα και να μεγιστοποιεί κάθε φορά την απόσταση διαχωρισμούμεταξύ των θετικών και αρνητικών στοιχείων, ελαχιστοποιώντας το σφάλμα ταξινόμησης. Το υπερεπίπεδο εκείνο που θα καταφέρει την μέγιστη απόσταση μεταξύ του ίδιου και του κοντινότερου σημείου (δηλαδή πρακτικά των δύο κλάσεων), ονομάζεται υπερεπίπεδο μέγιστου διαχωρισμού (maximummarginhyperplane), συμβολίζεται ως $\rightarrow_w$ και είναι το ζητούμενο. Ένας τέτοιος γραμμικός ταξινομητής ονομάζεται ταξινομητής μέγιστου διαχωρισμού.Τα διανύσματα τωνπιο κοντινών στοιχείων στο υπερεπίπεδο αυτό είναι τα διανύσματα υποστήριξης (supportvectors), δηλαδή εκείνα τα στιγμιότυπα εκπαίδευσης πουεπιλέγονται από κάθε κατηγορία και που ορίζουν το μέγιστο περιθώριο (mergin) των δύο κατηγοριών. Έτσι δημιουργείται μια γραμμική συνάρτηση διάκρισης (discriminantfunction)που θα κάνει τον βέλτιστο διαχωρισμό.

Ο SVM, ως γραμμικός ταξινομητής, έχει την δυνατότητα να μαθαίνει και να προβλέπει ανεξάρτητα των διαστάσεων του χώρου των χαρακτηριστικών και του πλήθους των χαρακτηριστικών, χρησιμοποιώντας τις υποθετικές συναρτήσεις. Επομένως, έχει τη δυνατότητα να εφαρμόζεται σε πολλά διαφορετικά προβλήματα ταξινόμησης και έτσι προτιμάται, λόγω αποτελεσματικότητας, ταχύτητας και ικανότητας να επιτύγχει τόσο γραμμική αλλά και μη γραμμική κατηγοριοποίηση (μετασχηματίζοντας το χώρο των χαρακτηριστικών του προβλήματος, σε έναν χώρο μεγαλύτερης διάστασης).
Συγκεκριμένα έστω ότι $c_j \in \{1,-1\}$, δηλαδήοι κλάσεις που μπορούν να πάρουν θετική ή αρνητική ετικέτα, είναι οι σωστές κλάσεις για τα κείμενα, τότε έχουμε τη λύση από την παρακάτω εξίσωση.

$$\underset{w}{\rightarrow} = \sum_j a_j c_j d_j, a_j \geq 0 \text{[37]}$$

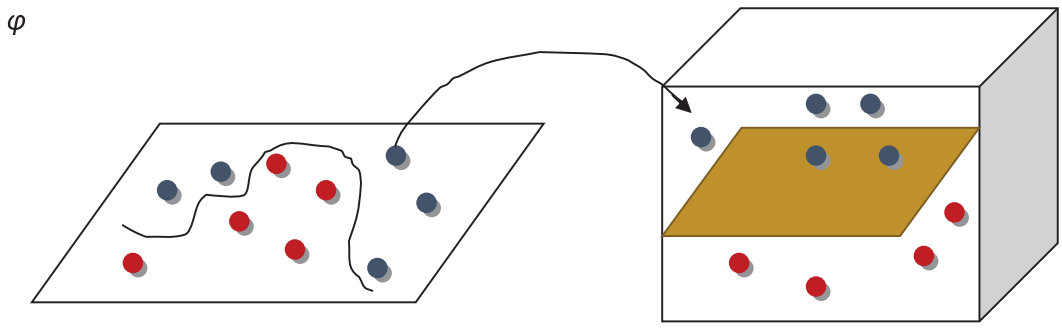
όπου:

- $\underset{w}{\rightarrow}$να είναι το διάνυσμα του υπερεπιπέδου μέγιστου διαχωρισμού.
- c_jείναι μια κλάση, θετική (1) ή αρνητική (-1).
- a_j, μια ποσότητα μεγαλύτερη ή ίση του μηδενός, που υπολογίζεται λύνοντας ένα πρόβλημα βελτιστοποίησης.
- d_jείναι ένα έγγραφο κειμένου. Ταd_jγια τα οποία τα a_jείναι μεγαλύτερα από το μηδέν, είναι τασupportvectors.

*hyperplaneστον n-διάστατο χώρο είναι ένα επίπεδο υποσύνολο n-1 διάστασης που χωρίζει το χώρο σε δύο μέρη.

Αφού εξεταστεί η πλευρά toumaximummarginhyperplaneστην οποία βρίσκονται τα δεδομένα, γίνεται η κατηγοριοποίησή τους.Όπως ήδη αναφέραμε, τονέο προς μελέτη στιγμιότυπο n-δεδομένων (σημεία) αντιμετωπίζεται σαν ένα διάνυσμα n-διαστάσεωνκαιστόχος είναι να χωρίσουμε αυτά τα σημεία με ένα (n–1)-διάστατουπερεπίπεδο (γραμμικός ταξινομητής).

Ένα απλό παράδειγμα αναπαράστασης της λειτουργίας των SVMs είναι το εξής:
Έστω ότι έχουμε 2 σύνολα από μπάλες διαφορετικού χρώματος (μπλεκαι κόκκινες) πάνω στο τραπέζι και θέλουμε να τις χωρίσουμε με βάση το χρώμα. Παίρνουμεμια βέργα και την βάζουμε ανάμεσα τους πάνω στο τραπέζι. Παρ'όλο που οι μπάλες έχουν διαχωριστεί σωστά, βάζουμε επιπλέον μπάλες πάνω στο τραπέζι και 1 μπάλα είναι στην λάθος μεριά του τραπεζιού και συνεπώς υπάρχει μια καλύτερηθέση να τοποθετήσουμε την βέργα διαχωρισμού. Ο SVM προσπαθεί συνέχεια να τοποθετήσειτην βέργα στην καλύτερη δυνατή θέση έτσι ώστετο κενό σε κάθε μια πλευρά της από τα αντίστοιχα σύνολα με μπάλες να είναι όσο το δυνατό μεγαλύτερο. Στο τέλος, η βέργατοποθετείται σε ένα σημείο που διαχωρίζει απόλυτα τα σύνολα αυτά.
Έστω τώρα ότι έρχεται κάποιοςκαι τοποθετεί τις μπάλες σε σημεία που δεν γίνεται να διαχωριστούνε πλήρωςμε γραμμική βέργα πάνω στο επίπεδο. Τότε αναποδογυρίζουμε το τραπέζι και πετάμε τις μπάλες στον αέρα. Βρισκόμενοι σε μια άλλη διάστασηκαι με ιδιαίτερες τεχνικές, παίρνουμε ένα χαρτί και χωρίζουμε τα 2 σύνολα και πάλι πλήρωςμε μια επιφάνεια.Αν βλέπαμε τις μπάλες προηγουμένως ο μόνος τρόπος να διαχωριστούν ήταν με μιακαμπύλη γραμμή, ωστόσο στον αέρα αυτό μπορεί πολύ εύκολα να επιτευχθεί με μια ίσιαεπιφάνεια. Οι μπάλες είναι τα δεδομένα, η ράβδος oclassifier,η μέγιστη δυνατή απόσταση trickoptimization, το αναποδογύρισμα του τραπεζιού kernellingκαι το κομμάτι χαρτί hyperplane.



Εικόνα 6: Πρακτική απεικόνιση του SVM

- **MaximumEntropy**

Ο ταξινομητής μέγιστης εντροπίας έχει κοινά χαρακτηριστικά με τον ταξινομητή NaiveBayes, καθώς ο δεύτερος είναι πιο ειδικός από τον πρώτο, με κάποιες βασικές διαφορές. Αρχικά, στην περίπτωση του NaiveBayes προσδιορίζεται μια ξεχωριστή παράμετρος για κάθε ετικέτα (εκ των προτέρων πιθανότητα) και μια άλλη για κάθε ζεύγος χαρακτηριστικού-ετικέτας (δηλαδή για το πόσο συμβάλλει το κάθε χαρακτηριστικό στην πιθανότητα της ετικέτας). Σε αντίθεση με τον NaiveBayes, ο MaximumEntropy δίνει τη δυνατότητα στο χρήστη να επιλέξει τους συνδυασμούς ετικετών και τα χαρακτηριστικά εκείνα που θα έχουν δικές τους παράμετρους, αφού υπάρχει πιθανότητα συσχέτισης μιας ετικέτας με περισσότερα από ένα χαρακτηριστικά ή και το αντίθετο, χρησιμοποιώντας μια παράμετρο.
ΣτονMaximumEntropy δεν λαμβάνεται ως προϋπόθεση η ανεξαρτησία των χαρακτηριστικών, οπότε αυξάνεται το πλήθος των προβλημάτων που μπορούν να λυθούν. Δεδομένου λοιπόνότι δεν υπολογίζονται οι πιθανότητες για τον καθορισμό των παραμέτρων (αφού η ανεξαρτησία των χαρακτηριστικών δεν αποτελεί υπόθεση), καθιστάται δύσκολη η εύρεση των εξαρτήσεων για όλους τους συνδυασμούς των χαρακτηριστικών και άρα η εύρεση εκείνων των παραμέτρων που θα κάνουν τον ταξινομητή αποτελεσματικό γίνεται με επαναληπτικές μεθόδους βελτιστοποίησης, που είναι πάντα αποτελεσματικές.
Η διαδικασία λοιπόν του αλγορίθμουμε στόχο την μεγιστοποίηση της πιθανότητας του συνόλου εκπαίδευσης, την σωστή δηλαδή ανάθεση ετικετών στα δεδομένα, αναλύεται παρακάτω.
Αφού στην αρχή εισαχθούν στις παραμέτρους τυχαίες τιμές, γίνεται μια συνεχή και χρονοβόρα αλλαγή των τιμών αυτών μέσα στις επαναλήψεις, έτσι ώστε να βελτιωθούν όσο γίνεται.
Κάθε συνδυασμός από ετικέτες και χαρακτηριστικά που λαμβάνει δική του παράμετρο ονομάζεται jointfeature και είναι ιδιότητα των τιμών με ετικέτα, ενώτα απλά χαρακτηριστικά είναι ιδιότητα των τιμών χωρίς ετικέτα.Κάθε ετικέτα λαμβάνει μια βαθμολογία για μια συγκεκριμένη είσοδο, που εξαρτάται από τα jointfeatures, και είναι το γινόμενο των παραμέτρων που συσχετίζονται με τα jointfeatures και εφαρμόζονται στην είσοδο και την ετικέτα.Η πιθανότητα της κλάσηςc, δοθέντος του κειμένουdκαι των βαρών λ υπολογίζεται από τον τύπο:

$$P(c|d,\lambda)=\frac{\exp\sum_i \lambda_i f_i\,(c,d)}{\sum_{c'\in C}\exp\sum_i \lambda_i f_i(c',d)}\quad[37]$$

Όπου:

- $f_i(c,d)$ είναι ένα κοινό χαρακτηριστικό που ορίζουμε για την κάθε λέξη μιας κλάσης και παίρνει την τιμή 1 αν το πλήθος των λέξεων του κειμένου είναι θετικό και 0 σε άλλη περίπτωση.
- μέσω επαναληπτικής βελτιστοποίησης, προσδιορίζεται ένα χαρακτηριστικό-βάρος λ για κάθε κοινό χαρακτηριστικό και παίρνει μεγάλη τιμή για να μεγιστοποιηθεί η πιθανότητα των δεδομένων εκπαίδευσης.

5.3 Πρόβλεψη

Μετά το πέρας της διαδικασίας της εκπαίδευσης του ταξινομητή, ξεκινά η διαδικασία της ταξινόμησης της άγνωστης εισόδου (ένα κείμενο ελέγχου-testset) στις συγκεκριμένες, ήδη καθορισμένες, κατηγορίες.Ο ταξινομητής υπολογίζειτην πιθανότητα να προσδιοριστεί το κείμενο με μια ετικέτα και η μεγαλύτερη πιθανότητα θα είναιτο τελικό αποτέλεσμα του ταξινομητή, που θα κρατήσουμε. Για την αξιολόγηση τηςεπίδοσης του ταξινομητή, χρησιμοποιούμε διάφορες μετρικές που σχετίζονται με την ορθότητα (accuracy), την ακρίβεια (precision), την ανάκληση (recall), την εξειδίκευση (specificity) και το F-Measure και αξιολογούν τους αλγόριθμους μηχανικής μάθησης.

- Η συνολική***Ορθότητα*** πρόβλεψης (***Accuracy***), ηπιο συχνή και απλή μετρική, υπολογίζει το ποσοστό των στιγμιότυπων του συνόλου ελέγχου που ταξινομήθηκαν στην σωστή κατηγορία. Αφορά το πλήθος των σωστών ταξινομημένων προτύπων προς το συνολικό πλήθος.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Αν μας ενδιαφέρει το πόσο καλά μαθαίνει το σύστημά μας την κάθε κατηγορία θα πρέπει να χρησιμοποιήσουμε διαφορετικές μετρικές απόδοσης:

- Η ***Ακρίβεια*** (***Precision***)για την θετική π.χ. κατηγορία είναι ο αριθμός των σχετικών εγγράφων (εκείνων που ανήκουν στην ίδια κλάση και συγκεκριμένα εκείνων που ανήκουν στη θετική κλάση) που ανακτώνται από μια αναζήτηση προς το συνολικό αριθμό των εγγράφων(TP και FP,δηλαδή εκείνων που ταξινομήθηκαν στη θετική κλάση) που ανακτώνται από την εν λόγω αναζήτηση, δηλαδή το ποσοστό των σωστών ταξινομήσεων στην θετική κατηγορία,και εκφράζει την πιθανότητα ότι ένα (τυχαία επιλεγμένο) ανακτώμενο έγγραφο είναι σχετικό (μέση πιθανότητα σχετικής ανάκτησης).

$$\text{precision} = \frac{TP}{TP+FP}$$

- Η ***Ανάκληση*** (***Recall***) (αντιστόφως ανάλογη της ακρίβειας)για την θετική κατηγορία είναι ο αριθμός των σχετικών εγγράφων που ανακτώνται από μια αναζήτηση διαιρούμενο με το συνολικό αριθμό των υφιστάμενων σχετικών εγγράφων (TP και FN, δηλαδή όλων εκείνων που υπάρχουν και όντως ανήκουν στην θετική κλάση), δηλαδή το ποσοστό των σωστά ταξινομημένων δειγμάτωνμεταξύ όλων των δειγμάτων που ανήκουν στην κατηγορία αυτή, και εκφράζει την πιθανότητα να ανακτηθεί ένα σχετικό έγγραφο (τυχαία επιλεγμένο) σε μια αναζήτηση (μέση πιθανότητα πλήρους ανάκτησης).

$$\text{recall} = \frac{TP}{TP + FN}$$

- Η ***Εξειδίκευση*** (***Specificity***)το αντίστοιχο της ανάκλησης για την αρνητική κατηγορία, ως:

$$\text{specificity}=\frac{\text{TN}}{\text{TN}+\text{FP}}$$

- Το ***F-Measure***, συνδυάζοντας τις μετρικές της ακρίβειας και της ανάκλησης, προκύπτει ο αρμονικός μέσος των δύο και παρουσιάζει μια ολική εκτίμηση των μοντέλων.

$$\text{F-Measure}=\frac{2\times\text{recall}\times\text{precision}}{\text{recall}+\text{precision}}$$

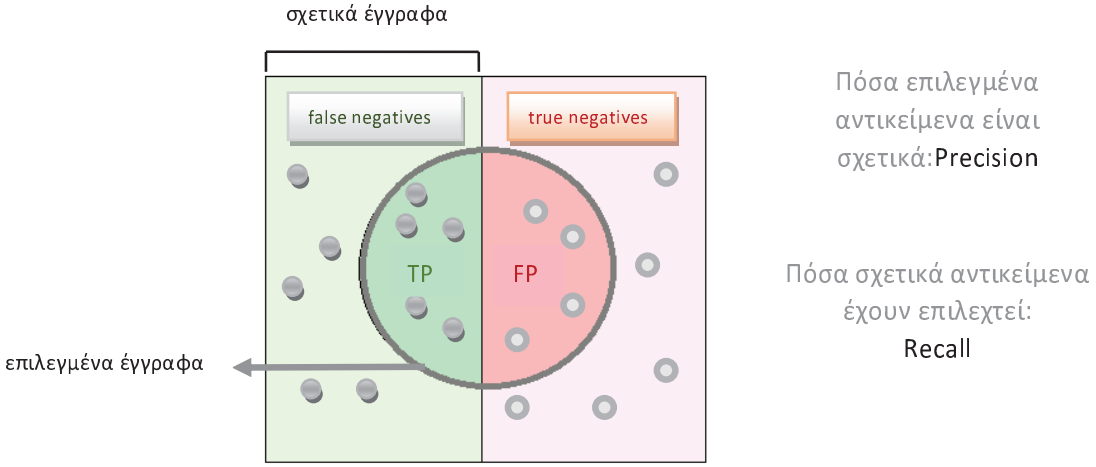
Όπου τα μεγέθη TP, TN, FP, FN αφορούν τις σωστές ή λανθασμένες ταξινομήσεις. Δηλαδή:

Το *TP*(TruePositives) αφορά τη σωστή ταξινόμηση, δηλαδή το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία «θετικό» και ταξινομήθηκανστην κατηγορία «θετικό».

Το *TN*(TrueNegatives) αφορά τη σωστή ταξινόμηση με το πλήθος των στιγμιοτύπων που ανήκουν στην κατηγορία «αρνητικό» να ταξινομούνται στην κατηγορία «αρνητικό».

Το *FP*(FalsePositives), δηλαδή ηλανθασμένη ταξινόμηση όπου το πλήθος των στιγμιοτύπων που ανήκουνστην κατηγορία «αρνητικό», ταξινομήθηκαν στην κατηγορία «θετικό».

Το*FN*(FalseNegatives) είναι η λανθασμένη ταξινόμηση, όπου το πλήθος των στιγμιοτύπων που ανήκουνστην κατηγορία «θετικό», ταξινομήθηκαν στην κατηγορία «αρνητικό».



Εικόνα 7: Precision και recall

5.3 Εφαρμογή

Με βάση τις παραπάνω μεθόδους της Επιβλεπόμενης Μηχανικής Μάθησης, θα εφαρμόσουμε δύο αλγόριθμους (*MultinomialNaïveBayes*, *SupportVectorMachine*) στο προγραμματιστικό περιβάλλον Weka και χρησιμοποιώντας arff αρχεία από το Twitter θα μελετήσουμε τα αποτελέσματά τους και τα ποσοστά ακρίβειάς τους σε διαφορετικές συνθήκες κάθε φορά.

6 ΜΟΝΤΕΛΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

6.1 Εισαγωγή στη Μηχανική Μάθηση

Η μηχανική μάθηση στηρίζεται στην στατιστική, καθώς και τα δύο πεδία χρησιμοποιούν ανάλυση δεδομένων και πολλές φορές ταυτίζεται και με τη εξόρυξη δεδομένων (datamining), η οποία βέβαια επικεντρώνεται περισσότερο στη διερευνητική ανάλυση δεδομένων. Η ανάπτυξή της, μέσω της δημιουργίας και εξέλιξης μεθόδων που έχουμε ήδη αναφέρει, αποτελεί έναν από τους βασικότερους λόγους ανάπτυξης του τομέα της ανάλυσης συναισθημάτων.
Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται ανάλογα με το επιθυμητό αποτέλεσμα του αλγορίθμου. Οι κατηγορίες αυτές είναι:

- Επιβλεπόμενη μάθηση (supervisedlearning)και η πιο δημοφιλής, όπου ο αλγόριθμος κατασκευάζει μια συνάρτηση με σκοπό να απεικονίσει δεδομένες εισόδους σε γνωστές εξόδους (σύνολο εκπαίδευσης), έτσι ώστε να επιτευχθεί η γενίκευση της συνάρτησης αυτής για την πρόβλεψη άγνωστων εξόδων (σύνολο ελέγχου) των δεδομένων που εισέρχονται.
- Μη επιβλεπόμενη μάθηση(unsupervisedlearning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει τις εξόδους για το σύνολο εκπαίδευσης, με σκοπό να μάθει περισσότερα για τις εισόδους αυτέςκαι να μπορέσει να βγάλει αποτελέσματα[25]. Είναι μια πιο δύσκολη διαδικασία, κατά την οποία μέσω κάποιων αλγορίθμων εξάγεται μια συνάρτηση με σκοπό να ανακαλύψει τα κρυμμένα χαρακτηριστικά και τη δομή των δεδομένων, που δεν είναι προσδιορισμένα με μία ετικέτα-κλάση. Δεν πραγματοποιείται κάποια αξιολόγηση ως προς την ακρίβεια του αποτελέσματος, αφού τα χαρκτηριστικά δεν ανήκουν σε κατηγορίες, όμως έχει τη δυνατότητα να ερμηνεύει και να βρίσκει λύσεις σε μια απεριόριστη ποσότητα δεδομένων, γεγονός που την διαφοροποιεί από τις επιβλεπόμενες τεχνικές ανάλυσης [38]. Χρησιμοποιεί όμως τεχνικές για να μπορέσει να ερμηνεύσει χαρακτηριστικά κλειδιά των δεδομένων με μεθόδους εξόρυξης γνώσης στο στάδιο της προεπεξεργασίας με την πιο σημαντική και ταυτόχρονα δημοφιλής τεχνική μη επιβλεπόμενης μηχανικής μάθησης να είναι εκείνη της συσταδοποίησης (clustering).

6.2 Επιβλεπόμενη Μηχανική Μάθηση

Πρόκειται για τη διαδικασία κατά την οποία παράγεται μια συνάρτηση απόκατηγοριοποιημένα δεδομένα. Η επιβλεπόμενη μηχανική μάθηση (supervisedmachinelearning), η πιο γνωστή τεχνική κατηγοριοποίησης (classification) συναισθήματος, είναι η μάθηση που στηρίζεται στην κατηγοριοποίηση των αντικειμένων εισόδου, δηλαδή έχοντας ένα προκαθορισμένο σύνολο από κλάσεις, τοποθετεί τα αντικείμενα προς εξέταση σε κάθε μια από τις κλάσεις αυτές (π.χ. θετικά, αρνητικά, ουδέτερα). Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο με τα σωστά στιγμιότυπα για τη εκπαίδευση του αλγορίθμου (trainingset). Κάθε στιγμιότυπο είναι ένα ζευγάρι που περιέχει ένα αντικείμενο εισόδου (συνήθως ένα διάνυσμα χαρακτηριστικών, δηλαδή μια πιο απλή και εύχρηστη αναπαράσταση του κειμένου) και μια γνωστή τιμή της εξόδου (σήμα ελέγχου). Ο ταξινομητής (classifier) προσπαθεί να εντοπίσει τις διαφορές ανάμεσα στα διανύσματα αυτά (δηλαδή τα κείμενα) που ανήκουν σε διαφορετικές κατηγορίες και για αυτό πρέπει να χρησιμοποιηθούν σύνολα εκπαίδευσης.
Ένας αλγόριθμος επιβλεπόμενης μάθησης μελετά τα δεδομένα εκπαίδευσης και τις κλάσεις τους, έτσι ώστε να εξάγει μια συνάρτηση που να κατηγοριοποιεί νέα (άγνωστα) δεδομένα, δηλαδή να κάνει σωστή απονομή ετικετών κλάσης σε άγνωστα συμβάντα. Μερικοί από τους δημοφιλέστερους αλγορίθμους επιβλεπόμενης μάθησης είναι οι NaiveBayes, MaximumEntropy και SupportVectorMachines (SVM).
Οι τεχνικές επιβλεπόμενης μηχανικής μάθησης παρ'όλο που παρουσιάζουν υψηλή ακρίβεια σε σχέση με τις τεχνικές μη-επιβλεπόμενης μάθησης, εμφανίζουν κάποιες δυσκολίες που αφορούν τόσο το μεγάλο χρονικό διάστημα που χρειάζονται για να να δημιουργήσουν το σύνολο εκπαίδευσης του ταξινομητή (για την εύρεση κατάλληλων τιμών), όσο και το ποσοστό ακρίβειας που είναι εξαρτώμενο από το εκάστοτε σύνολο εκπαίδευσης.

6.2.1 Διαδικασία Επιβλεπόμενης Μηχανικής Μάθησης

Εδώ θα περιγράψουμε τα βήματα επίλυσης προβλήματος στη περίπτωση μιας επιβλεπόμενης μηχανικής μάθησης και θα αναφέρουμε μερικούς από τους πιο συχνούς αλγορίθμους επιβλεπόμενης μηχανικής μάθησης.Σαν πρώτο στάδιο, έχοντας το σύνολο δεδομένων (dataset), δημιουργούμε δύο νέα σύνολα-υποκατηγορίες του αρχικού που αφορούν το σύνολο εκπαίδευσης (trainingset) και το σύνολο ελέγχου (testset). Το trainingseteίναι το σύνολο εκείνο που δίνεται σαν είσοδος στον ταξινομητή, για να το μελετήσει και οι τιμές που έχουν επιλεγεί για να το απαρτίζουν θα πρέπει να είναι όσο πιο αντιπροσωπευτικές του πληθυσμού γίνεται για να φτάσουμε στην όσο το δυνατόν υψηλότερη ακρίβεια. Το testset χρησιμοποιείται για τον έλεγχο μετά την εκπαίδευση του ταξινομητή και πριν την πρόβλεψη των αποτελεσμάτων. Υπολογίζουμε δηλαδή την απόδοση της συνάρτησης κατηγοριοποίησης, εφαρμόζοντας την στο δείγμα δεδομένων ελέγχου (testset), το οποίο θα πρέπει να είναι διαφορετικό από το trainingset.

6.2.1.1 Δημιουργία dataset από Twitter με το Orange

Το Orange είναι ένα ανοιχτού κώδικα και δωρεάν λογισμικό, γραμμένο σε Python. Χρησιμοποιείται για μηχανική μάθηση και εξόρυξη δεδομένων. Διαθέτει ένα γραφικό περιβάλλον προγραμματισμού με σκοπό την ανάλυση και την απεικόνιση δεδομένων. Μπορεί να χρησιμοποιηθεί και ως μία βιβλιοθήκη της Python. Το πρόγραμμα διατηρείται και αναπτύσσεται από το Εργαστήριο Βιοπληροφορικής του τμήματος Υπολογιστών και Πληροφορικής στο Πανεπιστήμιο της Λιουμπλιάνα.

Τα εργαλεία του λογισμικού ονομάζονται widgets και χρησιμοποιούνται είτε για απεικόνιση στοιχείων, επιλογή υποσυνόλων, προεπεξεργασία ή για την εμπειρική αξιολόγηση αλγορίθμων μάθησης και την προγνωστική μοντελοποίηση.

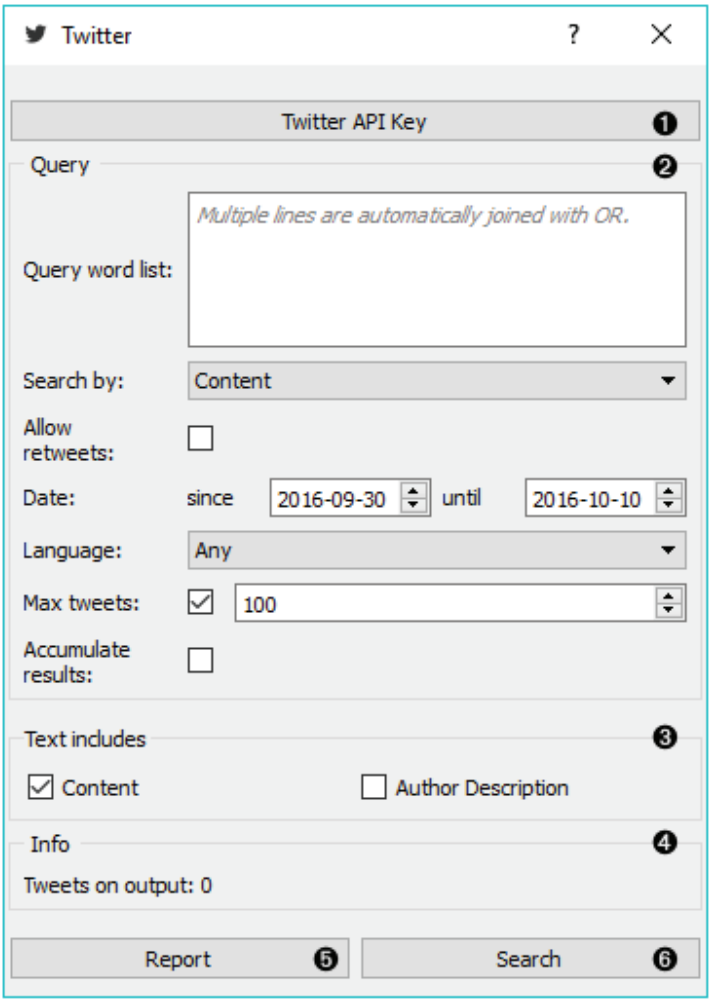
Ο οπτικός προγραμματισμός υλοποιείται μέσα από ένα γραφικό περιβάλλον στον οποίο οι ροές εργασίας δημιουργούνται από τη σύνδεση προκαθορισμένων ή σχεδιασμένων από το χρήστη widgets, ενώ οι προχωρημένοι χρήστες μπορούν να χρησιμοποιούν το Orange, ως μία βιβλιοθήκη της Python για το χειρισμό των δεδομένων, και τη τροποποίηση των widget.

Το widget Twitter επιτρέπει την αναζήτηση μηνυμάτων tweets μέσω του Twitter API. Η αναζήτηση μπορεί να γίνει ανά περιεχόμενο, συγγραφέα ή και τα δύο και να συσσωρευτούν τα αποτελέσματα αν θέλουμε να δημιουργήσουμε ένα ευρύτερο σύνολο δεδομένων.

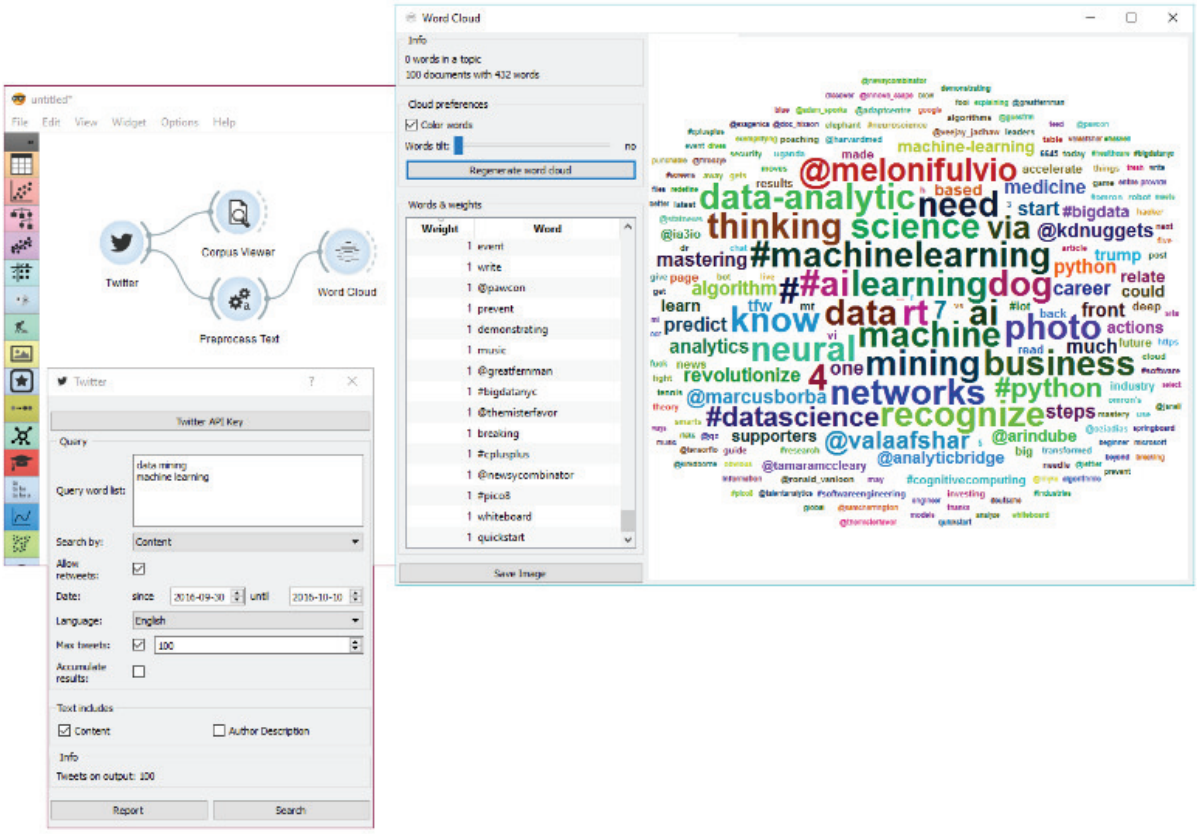
[40]

Οι παράμετροι που πρέπει να προσδιοριστούν για την αναζήτηση είναι οι εξής:

- Querywordlist, όπουοιαναζητήσεις συνδέονται αυτόματα με OR.
- Searchby, όπου καθορίζει κανείς αν θέλει να ψάξει ανά περιεχόμενο, συγγραφέα (το twitterusername χωρίς @) ή και τα δύο.
- Allow retweets
- Date, όπου ορίζεται το χρονικό διάστημα από το οποίο θέλει κανείς τα tweets.
- Language, δηλαδή η γλώσσαστην οποία ανακτώνται τα retweets.
- Maxtweets, δηλαδή το μέγιστο όριο των ανακτώμενων tweets.
- Accumulateresults, τοοποίοαν ενεργοποιηθεί θα έχουμε το περιθώριο για νέες αναζητήσεις μαζί με τις προηγούμενες, και θα προστεθούν τα αποτελέσματα.



Για παράδειγμα, θα ψάξουμε τώρα για tweets που περιέχουν σαν περιεχόμενο τις λέξεις ‘datamining’ ή ‘machinelearning’. Θα επιτρέπονται τα retweets και το όριο των tweets προς αναζήτηση, στην αγγλική γλώσσα, θα ανέρχεται στα 100.



Αρχικά βλέπουμε το αποτέλεσμα (τα κείμενα και σχετικές πληροφορίες) στο CorpusViewer. Κατόπιν, προεπεξεργαζόμαστε τα tweets με μετατροπή τους σε πεζά γράμματα, αφαίρεση url συνδέσμων, tokenizing και αφαίρεση των stopwords και σημείων στίξης.
Τα αποτελέσματα τώρα φαίνονται στο WordCloud, όπου εμφανίζονται οι πιο δημοφιλείς λέξεις στον τομέα του datamining και του machinelearning.

6.2.1.2 Προεπεξεργασία

Στην προεπεξεργασία, τροποποιούμε κατάλληλα το κείμενο (corpus) και τα χαρακτηριστικά του για να γίνει η εκπαίδευση και να το επεξεργαστούμε με τις μεθόδους επιβλεπόμενης μηχανικής μάθησης. Για την σωστή προσαρμογή του εκάστοτε λοιπόν κειμένου λαμβάνονται υπ'όψην ορισμένα χαρακατηριστικά, που μπορεί να υπάρχουν σε περιπτώσεις που το περιβάλλον που το περιέχει είναι το Twitter, και εκτελούνται οι εξής διαδικασίες που αφορούν το φιλτράρισμα λέξεων που δεν συνεισφέρουν στο νόημα της πρότασης:

- Αφαιρούνται [26] ή αντικαθίστανται με λέξεις-κλειδιά οι αναφορές προς άλλους χρήστες (@username) και τα hashtags[27] (ή αφαιρείται μόνο ο χαρακτήρας «#» [28]).
- Αφαιρούνται ή αντικαθίστανται τα emoticons με λέξεις συναισθήματος (μέσω ενός έτοιμου λεξικού emoticons) και οι συντομογραφίες με την πλήρη μορφή τους. [28]

- Αφαιρούνται τα retweets (δηλαδή το «RT»), κοινέςκαι συνηθισμένες λέξεις, επαναλαμβανόμενοι χαρακτήρες (π.χ. «looooooove») και άρθρα (a,an,the). [26][27] [28]
- Συνενώνονται οι λέξεις άρνησης (no,not) με την προηγούμενη ή επόμενη λέξη. [26]
- Πραγματοποιείται tokenization. Πρόκειται για μια απαραίτητη διαδικασία που αφορά την τμηματοποίηση του κειμένου σε παραγράφους, προτάσεις, λέξεις, δηλαδή στα λεγόμενα tokens για να μπορέσουν αυτά να μελετηθούν ξεχωριστά.
- Εφαρμόζεται lemmatization, όπου επιστρέφεται το θέμα της λέξης (η λέξη από την οποία προέρχεται και δεν είναι πάντα η ρίζα) [28] [29] και έτσι αποθηκεύονται λέξεις με το ίδιο νόημα. Δηλαδή η λέξη «saw», που σημαίνει «πριόνι» ή «είδα» (αόριστος του ρήματος «see»), θα μας δώσει την ίδια τη λέξη (στην περίπτωση που σημαίνει «πριόνι») ή το ρήμα «see» και με βάση το lemma θα μπει σε λίστα με τα συνώνυμά της.

6.2.1.3 Χαρακτηριστικά-Features

Τα χαρακτηριστικά είναι ιδιότητες που βοηθούν στην αποσαφήνιση μιας οντότητας και συνεπώς στη σωστή ταξινόμηση. Για παράδειγμα, για την οντότητα «υπολογιστής», τα χαρακτηριστικά είναι «οθόνη», «πληκτρολόγιο», «ποντίκι» κτλ. Πρέπει και αυτά βέβαια να είναι στην κατάλληλη μορφή (διανύσματα ή δυαδικά μεγέθη) για να μπορούν να χρησιμοποιηθούν ως είσοδος στον ταξινομητή. Τέτοια χαρακτηριστικά είναι:

- Τα tokens, δηλαδή οι λέξεις που προέρχονται από το tokenization στη διαδικασία της προεπεξεργασίας του κειμένου.
- N-grams*, δηλαδή μονογράμματα, διγράμματα και ο συνδυασμός αυτών. Ο ταξινομητής SupportiveVectorMachineέχει αποδειχθεί ο πιοαποτελεσματικός για τη χρήση των μονογραμμάτων.[34]
- Οι Part-Of-Speech taggers. Το POStagging είναι όπως προαναφέραμε το πρόβλημα της γραμματικής επισημείωσης. Πολλές φορές όμως, μια λέξη μπορεί να έχειπαραπάνω από μία ερμηνείες (π.χ. η λέξη «προβλέψεις» που μπορεί να εμφανιστείείτε σαν ρήμα είτε σαν ουσιαστικό). Έτσι, οι partofspeechtaggers, μελετώντας συνολικά το κείμενο και τη γλώσσα στην οποία είναι γραμμένο, μπορούν να αναγνωρίσουν περισσότερες ιδιότητες της λέξης, όπως πτώση, γένος, αριθμός κλπ.
- Τα σημεία στίξης.
- Το πλήθος των θαυμαστικών. [30]
- Τα κεφαλαία γράμματα. [30][31]
- Το πλήθος των επαναλαμβανόμενων γραμμάτων.[32]
- Η άρνηση που μπορεί να αλλάξει την πολικότητα του κειμένου και όταν μελετάται μεμονωμένα, ως χαρακτηριστικό, δεν φέρει ορθά αποτελέσματα.[34]

6.3 Προγραμματιστικό Περιβάλλον – Weka

Στην συγκεκριμένη εργασία, το προγραμματιστικό περιβάλλον το οποίο στηρίχτηκε ήταν το λογισμικό Weka (WaikatoEnvironmentforKnowledgeAnalysis). Πρόκειται μια δημοφιλή πλατφόρμα, ελεύθερου λογισμικού, μηχανικής μάθησης γραμμένο στην γλώσσα «Java», το οποίο αναπτύχθηκε στο Πανεπιστήμιο «Waikato» της Νέας Ζηλανδίας. Περιέχει εργαλεία οπτικοποίησης, γραφικά στοιχεία και αλγορίθμους για την ανάλυση δεδομένων και την πρόβλεψη μοντέλων. Οι βασικές λειτουργίες του είναι η εξόρυξη δεδομένων (δηλαδή η προεπεξεργασία τους με χρήση συγκεκριμένων φίλτρων), η ομαδοποίηση, η ταξινόμηση και η απεικόνιση των αποτελεσμάτων. Οι τεχνικές στο Wekaβασίζονται στο γεγονός ότι τα δεδομένα είναι διαθέσιμα ως ένα απλό αρχείο, όπου κάθε σημείο δεδομένων αποτελείται από έναν σταθερό αριθμό των χαρακτηριστικών (αριθμητικά, ονομαστικά κ.τ.λ.), έχοντας πρόσβαση σε SQLβάσεις δεδομένων.

.

6.3.2 Αρχεία στο Weka

Τα αρχεία που εισάγουμε στο Wekaπρέπει να είναι σε μορφήARFF (Attribute - RelationFileFormat). Πρόκειται γιααρχεία κειμένου χαρακτήρων, το οποίο περιλαμβάνει μία σειρά από
instances που περιγράφονται από χαρακτηριστικά(attributes)και προτιμώνται λόγω εξοικονόμησης μνήμης, ταχύτητας και αποτελεσματικότητας ως προς την ανάλυση, καθώς περιέχουν μεταδεδομένα (metadata) για τα columnheaders. [38]
Παράδειγμα τέτοιου αρχείου είναι το εξής:

@RELATIONtrenornoah

@ATTRIBUTEtweetidNUMERIC
@ATTRIBUTEtimestampDATE "yyyy-MM-ddHH:mm"
@ATTRIBUTEtextSTRING
@ATTRIBUTEclass {positive,negative}

@DATA
986213433996726273,"2018-04-17 17:03", 'Congrats @Trevornoah on the launch of your foundation.I can't wait what you accomplish',positive
991929735918833665,"2018-05-02 11:37",'This week's most influential person. I recognize the purest form of African talent #trevornoah',positive
990633886077079554,"2018-04-29 09:48",'It all went downhill. That's the last time I've watched #Trevornoah's degrading show',negative

Γενικότερα:

- Οι γραμμές που ξεκινάνε με % είναι σχόλια τα οποία δεν υπολογίζονται κατά τη διαδικασία φόρτωσης του αρχείου, έτσι ώστε το νόημα του κειμένου να είναι πιο κατανοητό.

- Οι γραμμές που ξεκινάνε με *@relation*, είναιυποχρεωτικές και περιγράφουν το αρχείο.

- Η δήλωση των attributes γίνεται σύμφωνα ως εξής:
@attribute<attribute_name><datatype>.
Το όρισμα *<attribute_name>* είναι το όνομα του χαρακτηριστικού, το οποίο πρέπει να ξεκινάει με γράμμα και το όρισμα *<datatype>* καθορίζει τον τύπο του χαρακτηριστικού,δηλαδή να είναι αριθμητικό (numeric) με πραγματικές ή ακέραιες τιμές, ονομαστικό (*<nominal-specification>*), αλφαριθμητικό (string) ή ημερομηνία (date [*<date-format>*]).
[38]

6.4 Συλλογή Δεδομένων

Η λειτουργία και η απόδοση του εκάστοτε μοντέλου κατηγοριοποίησης επηρέαζεται άμεσα και διαφέρει αναλόγως του συνόλου των δεδομένων που μελετάται. Εδώ μελετώνται δεδομένα από μηνύματα του Twitter, με το διαχρονικό μειονέκτημα του μικρού μεγέθους τους και της χρήσης νεολογισμών, συντομογραφιών και emoticons.

6.4.1 Δεδομένα από το Twitter

Το σύνολο δεδομένων που χρησιμοποιήθηκε περιέχει 2000 tweets, από το Twitter(http://twitter.com/). Τα δεδομένα που ανακτήθηκαν έχουν χωριστεί χειροκίνητα σε δύο κατηγορίες, θετικά και αρνητικά (1000 θετικά και1000 αρνητικά). Κάθε tweet αποτελεί και ξεχωριστό αρχείο κειμένου. Κάθε στιγμιότυπο του συνόλου δεδομένων περιλαμβάνει:

- το θέμα του tweet (#topic)
- το συναίσθημα του tweet
- το tweetid
- την ημερομηνία του δημοσιεύτηκε το tweet
- το κείμενο του tweet

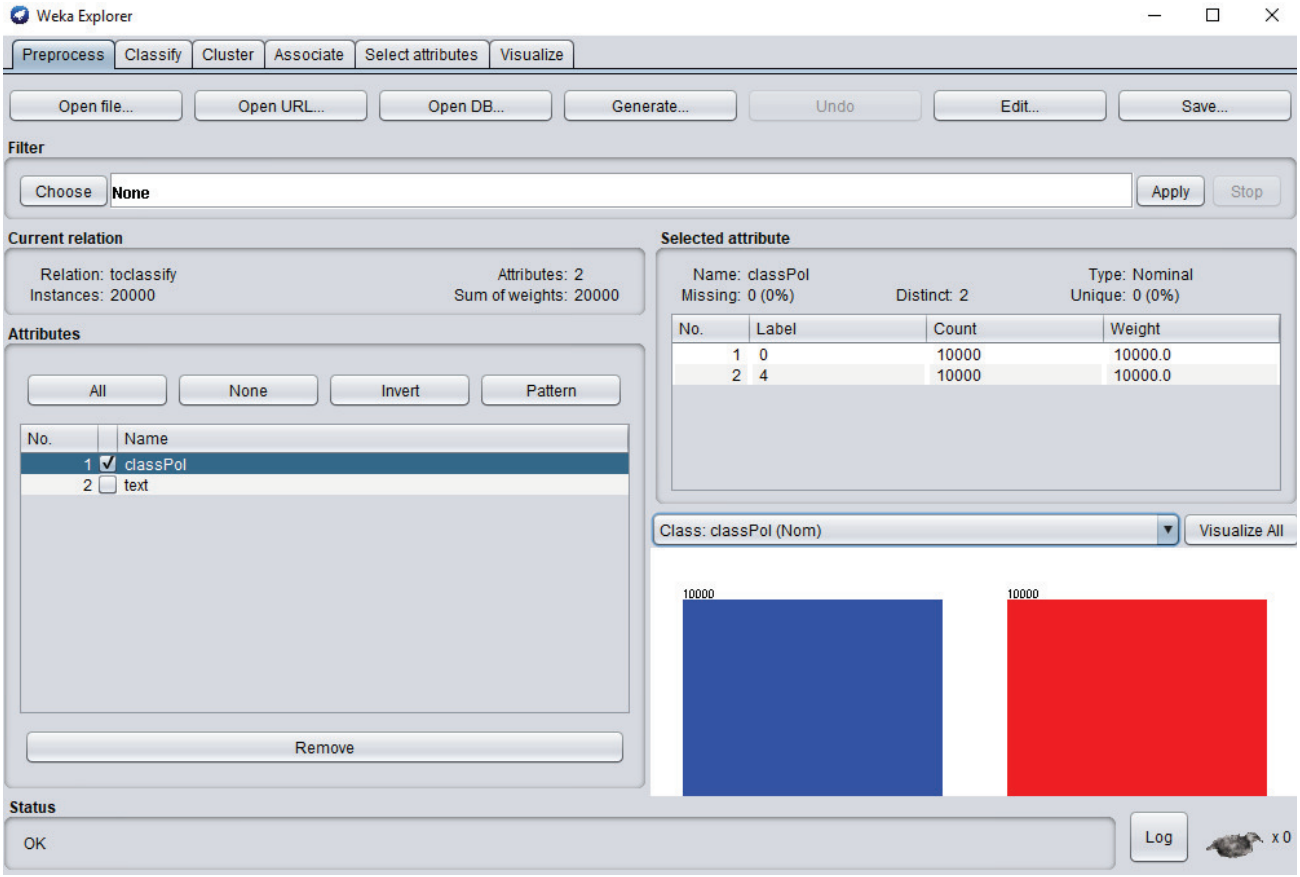
Εικόνα 8: Στιγμιότυπο από tweet του BillGates

6.5 Περιγραφή Μοντέλου Κατηγοριοποίησης

Τα μοντέλα κατηγοριοποίησης συναισθήματος πραγματοποιήθηκαν στο προγραμματιστικό περιβάλλον Weka. Ακολουθούν παρακάτω τα βήματα που υλοποιήθηκαν.

6.5.1 Εισαγωγή των δεδομένων

Αρχικά, εισάγουμε τα δεδομένα στο Weka. Πρόκειται για ένα αρχείο αποτελούμενο με 2000 φακέλους κειμένου και χωρισμένο στις δύο υποκατηγορίες pos και neg (οι τιμές της κλάσης). Από τη στιγμή που γίνεται η φόρτωση των δεδομένων, το Weka αυτόματαδημιουργεί μια συσχέτιση με δύο χαρακτηριστικάμε το πρώτο να περιέχει τα δεδομένα κειμένου και το δεύτερο να είναι η κλάση του εγγράφουπου έχει οριστεί από την υποκατηγορία του φακέλου (ροσή neg).Τοιστόγραμμα δείχνει την διανομή των κλάσεων (μπλε = neg, κόκκινο = θετικά).



Εικόνα 9: Εισαγωγή δεδομένων και διανομή κλάσης

6.5.2 Προεπεξεργασία Κειμένου στο Weka

Πριν την ταξινόμηση, χρειάζεται να γίνουν ορισμένες διεργασίες που αφορούν κυρίως τα γλωσσικά περιεχόμενα του κειμένου. Σκοπός είναι η μετατροπή κάθε κείμενο σε διάνυσμα, στο οποίο κάθε έγγραφο αντιπροσωπεύεται από την συχνότητα εμφάνισης κάποιων σημαντικών όρων. Πραγματοποιείται, λοιπόν:

- Word parsing και tokenization

Εδώ κάθε έγγραφο αναλύεται με σκοπό την εξαγωγή των όρων. Είναι αναγκαίος ο καθορισμός των χαρακτήρωνκαι η διαδικασία «tokenization» σε περιπτώσεις τονισμένων λέξεων, συζευγμένων και ακρωνυμίων.

- Αφαίρεση των stop-words

Σε αυτή τη φάση, αφαιρούνται οι πιο συχνά εμφανιζόμενες λέξεις (π.χ. άρθρα), καθώς δεν συνεισφέρουν στην ταξινόμηση του κειμένου λόγω απουσία υποκειμενικής ισχύος. Βέβαια

υπάρχει ο κίνδυνος να παραμείνουν σπάνιες λέξεις οι οποίες όμως δεν θεωρούνται αντιπροσωπευτικές.

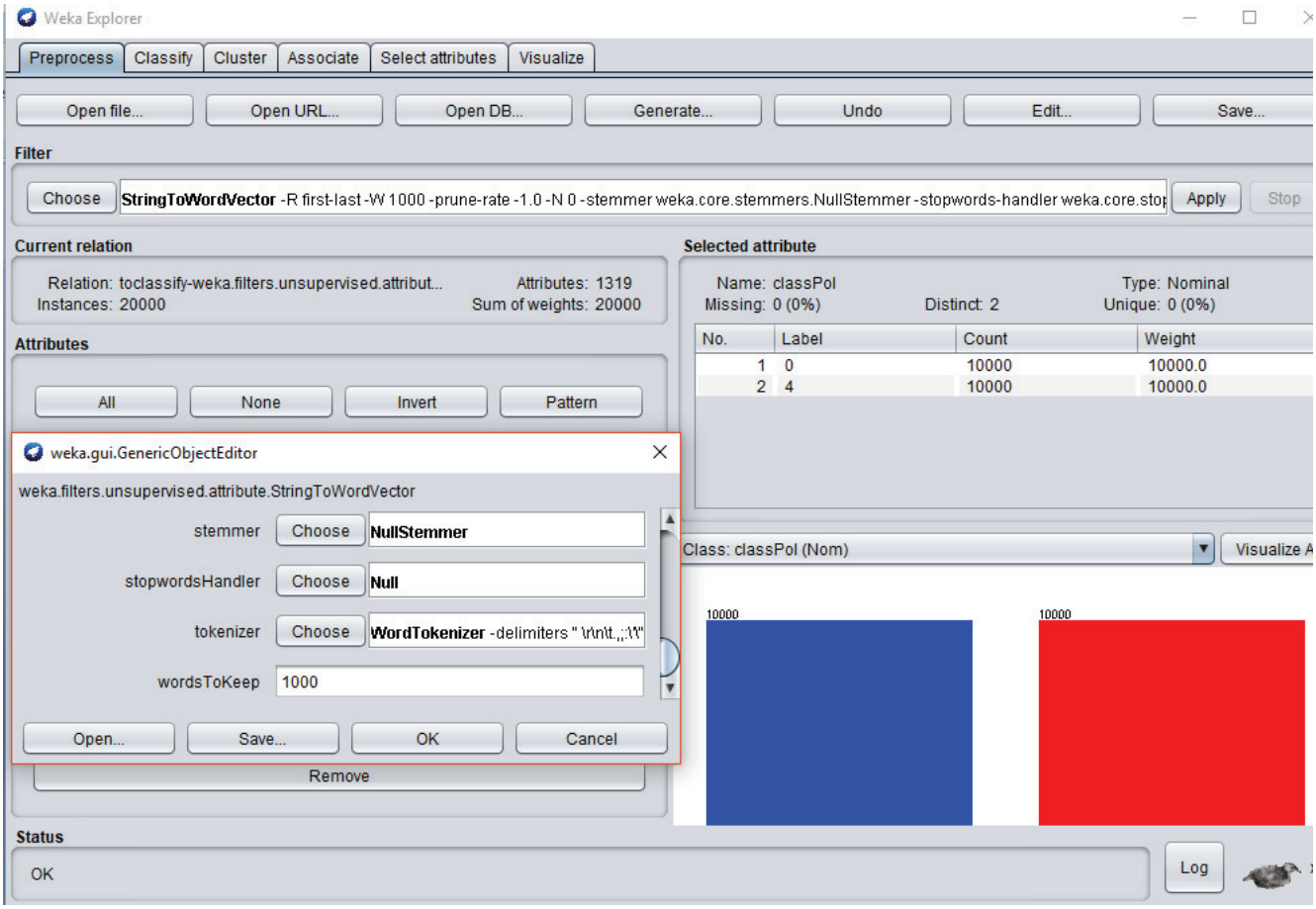
- Lemmatization και stemming

Όπως έχουμε ήδη αναφέρει, κατά την τεχνική του lemmatization, καθορίζεται το λήμμα της λέξης, δηλαδή το κοινό θέμα (η κοινή ρίζα) που έχει με άλλες παραγώγους της. Οπιο απλός τρόπος για να επιτευχθεί κάτι τέτοιο είναι η αφαίρεση του επιθέματος με stemmingαλγορίθμους.

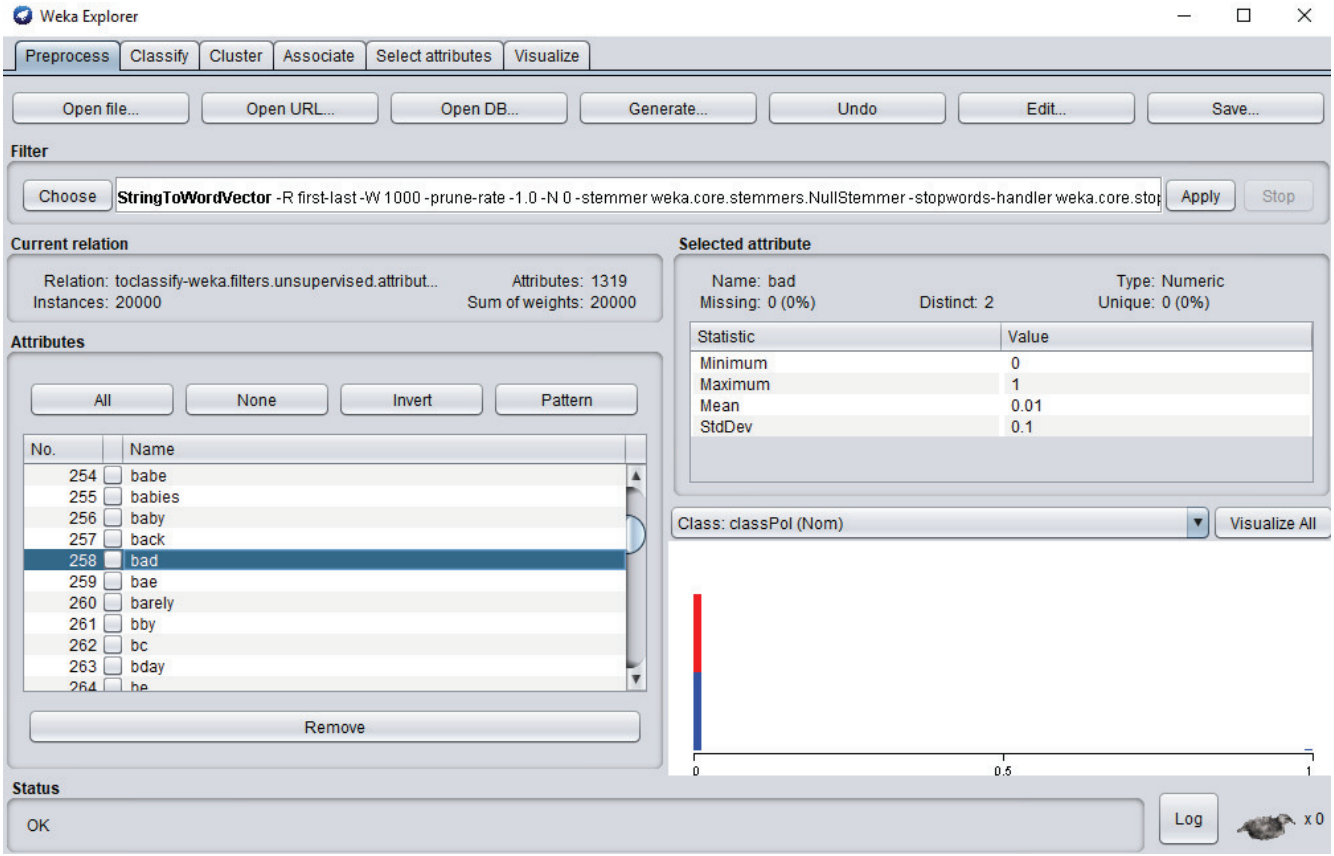
- Term selection/feature extraction

Προκειμένου να γίνει πιο αποτελεσματική η κατηγοριοποίηση του κειμένου, οι όροι που έχουν επιλεγεί από όλες τις προηγούμενες φάσεις πρέπεινα περάσουν από μία επιπλέον επεξεργασία φιλτραρίσματος, για τηναφαίρεση εκείνων που έχουν μειωμένη ικανότητα προβλεψιμότητας ή είναιάμεσα συνδεδεμένοι με άλλους όρους.

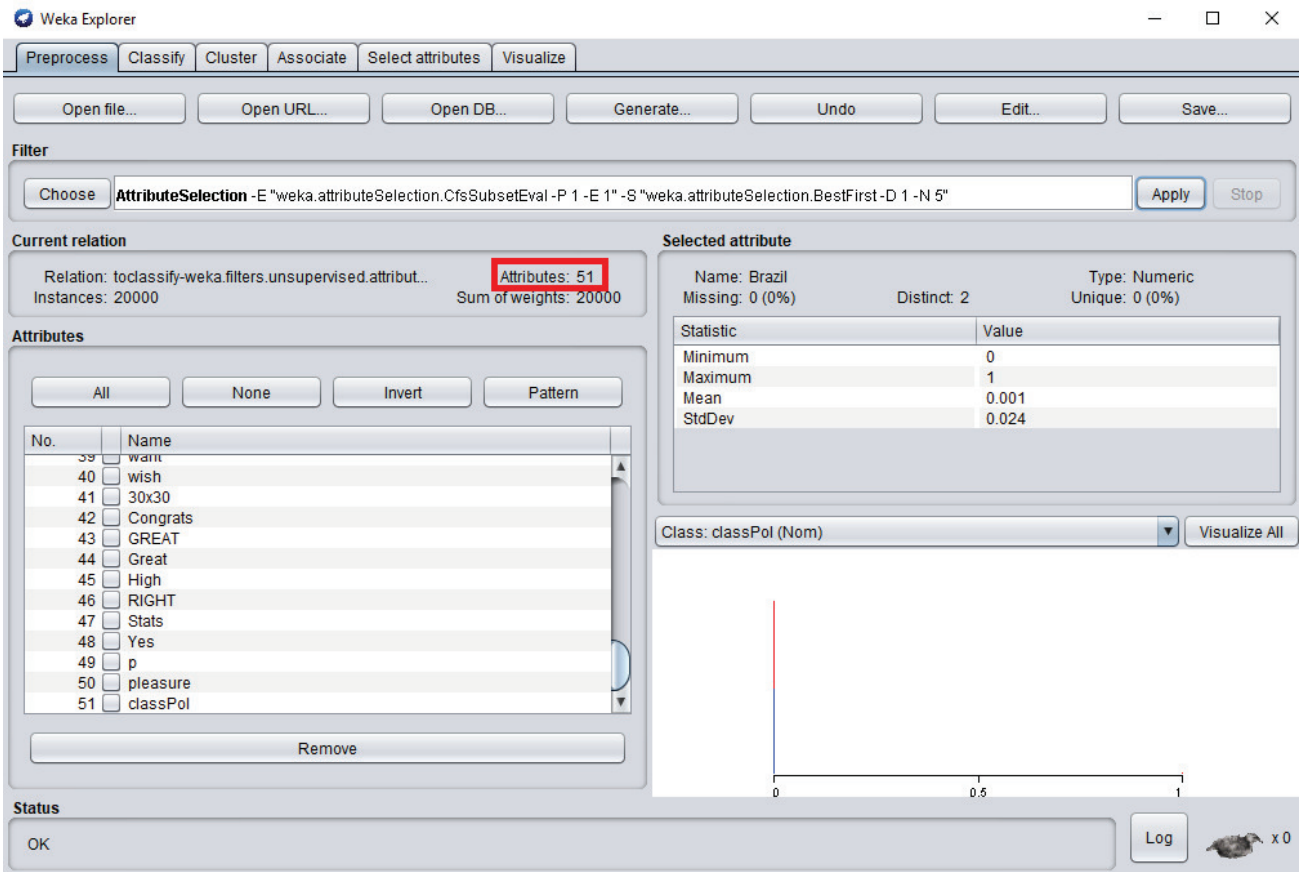
Με το φίλτρο «StringToWordVector», που μετατρέπει το κείμενο σε διάνυσμα, δίνεται η δυνατότηταγιαεφαρμογή του tokenizer, καθορισμό ή μη μιας stop-wordsλίστας και ενός stemmer.Κάθε διαφορετική επιλογή θα έχει και διαφορετικό αποτέλεσμα, οπότε θα επιλέξω όλους τους δυνατούς συνδυασμούςανάμεσα της stop-wordsλίστας του stemmerκαι του attributeselection για να διερευνήσουμε κατά πόσο επηρεάζονται τα αποτελέσματα των μοντέλων.



Εικόνα 10: Επιλογή του φίλτρου «StringToWordVector», χωρίς τη χρήση stemmerκαι stop-words λίστας



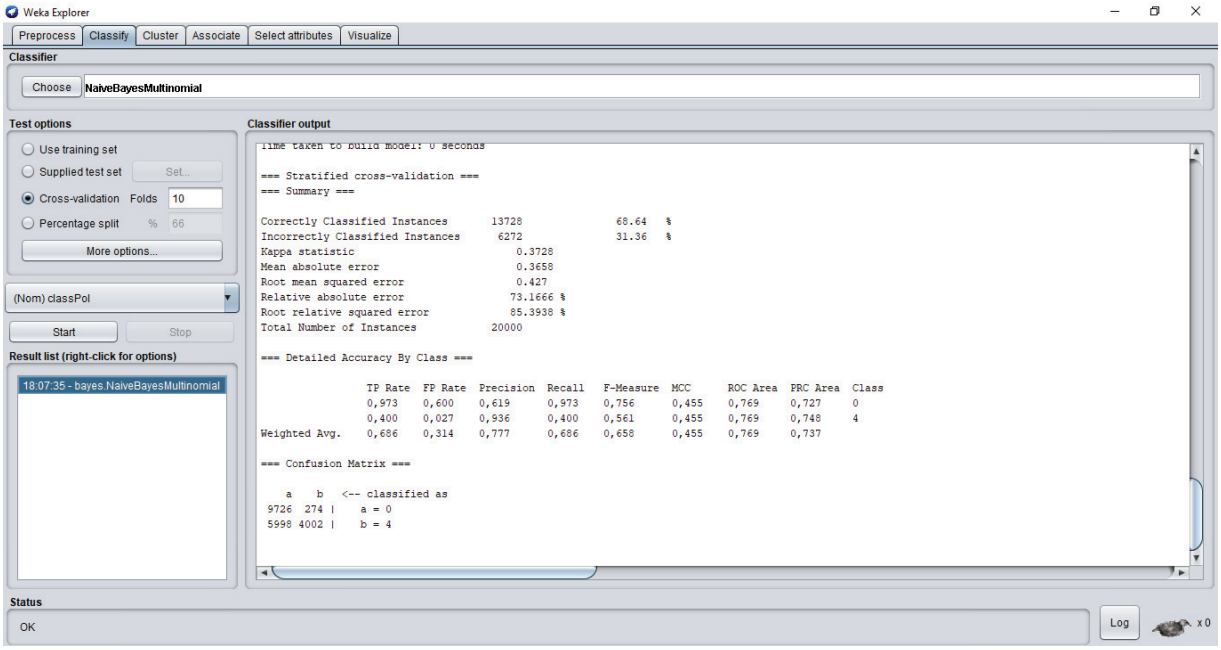
Εικόνα 11: Επιλογή όρου που δείχνει τη διανομή της λέξης «bad» στα έγγραφα με συχνότερη εμφάνιση (τιμή 1) στα αρνητικά σχόλια (μπλε γραμμή)



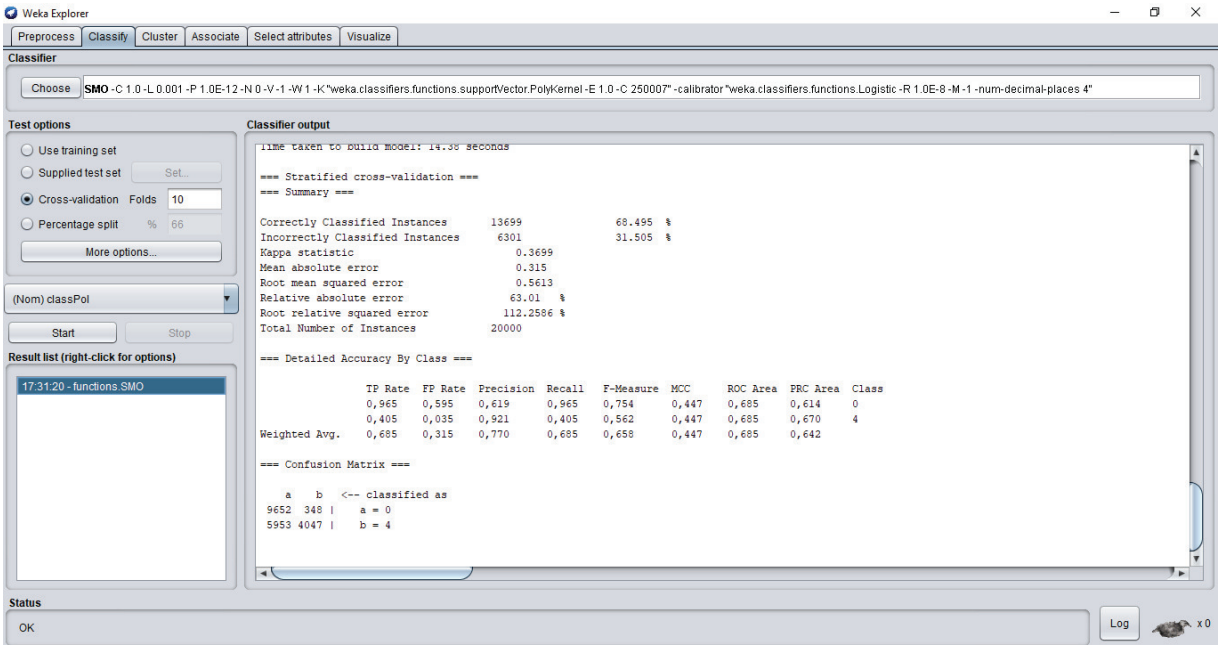
Εικόνα 12:Τα 1319 attributes, μειώθηκαν σε 51,μετάτηντελευταίαφάσητουattributeselectionμετοφίλτρο «AttributeSelection» καιτηνεπιλογήτηςμεθόδου «CfsSubsetEval» (Correlation-basedfeaturesubsetselection) πουεπιλέγειεκείναταattributesπουείναιστενάσυνδεδεμέναμετιςκλάσεις και πολύ λιγότερο με τα υπόλοιπα.

6.5.3 Επιλογή αλγορίθμου

Οι αλγόριθμοι που έχουν επιλεγεί, για την υλοποίηση των μοντέλων, προέρχονται από την επιβλεπόμενη μηχανική μάθηση. Ο ***MultinomialNaiveBayes***και ο αλγόριθμος ***SVM***θα μελετηθούν στη συνέχεια της εργασίας. Μετά την ολοκλήρωση τους, θα συγκριθούν τα αποτελέσματά τους.



Εικόνα 13: Αποτελέσματα του αλγόριθμου ***MultinomialNaiveBayes***, χωρίς stemmer και χωρίς αφαίρεση stop-wordsκαι με attributeselection



Εικόνα 14: Αποτελέσματα του αλγόριθμου ***SupportVectorMachine***, χωρίς stemming και χωρίς αφαίρεσηstop-wordsκαι με attributeselection

Stemmer	Stopwords Removal	Attribute Selection	TP Rate	FP Rate	Precision	Recall	FMeasure	Ποσοστό Σωστής Ταξινόμησης
no	no	yes	0.973	0.600	0.619	0.973	0.756	***58.64%***
no	yes	yes	0.971	0.794	0.550	0.971	0.702	***58.815%***
yes	no	yes	0.973	0.600	0.619	0.973	0.756	***58.64%***
yes	yes	yes	0.971	0.794	0.550	0.971	0.702	***58.815%***
no	no	no	0.839	0.275	0.753	0.839	0.794	***78.205%***
no	yes	no	0.818	0.420	0.661	0.818	0.731	***69.93%***
yes	no	no	0.839	0.275	0.753	0.839	0.794	***78.205%***
yes	yes	no	0.818	0.420	0.661	0.818	0.731	***69.93%***

Εικόνα 15: Αποτελέσματα του αλγόριθμου ***MultinomialNaiveBayes***

Stemmer	Stopwords Removal	Attribute Selection	TP Rate	FP Rate	Precision	Recall	FMeasure	Ποσοστό Σωστής Ταξινόμησης
no	no	yes	0.965	0.595	0.619	0.965	0.754	***68.495%***
no	yes	yes	0.219	0.037	0.856	0.219	0.349	***59.105%***
yes	no	yes	0.965	0.595	0.619	0.965	0.754	***68.495%***
yes	yes	yes	0.219	0.037	0.856	0.219	0.349	***59.105%***
no	no	no	0.840	0.285	0.747	0.840	0.791	***77.775%***
no	yes	no	0.645	0.258	0.714	0.645	0.678	***69.36%***
yes	no	no	0.840	0.285	0.747	0.840	0.791	***77.775%***
yes	yes	no	0.645	0.258	0.714	0.645	0.678	***69.36%***

Εικόνα 16: Αποτελέσματα του αλγόριθμου***SupportVectorMachines***

7 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η Ανάλυση Συναισθήματος αποτελεί σημαντική βοήθεια και απαραίτητο εργαλείο για τους χρήστες για την σωστή διαχείριση και επεξεργασία του τεράστιου όγκου πληροφοριών και δεδομένων που έχουν πρόσβαση, καθώς και για την εξαγωγή γνώσης μέσα από τα δεδομένα σε μορφή κειμένου, διαδικασίες που έγιναν δύσκολο μέχρι τώρα να πραγματοποιηθούν λόγω της δραματικής ανάπτυξης του Διαδικτύου.

Στην διπλωματική αυτή εργασία μελετώνται μοντέλα Κατηγοριοποίησης Συναισθήματος, σε σχόλια χρηστών στο Διαδίκτυο, ειδικότερα σε μηνύματα στο Twitter, με τη συμβολή και χρήση τεχνικών από την επιβλεπόμενη Μηχανική Μάθηση. Οι αλγόριθμοι που αξιοποιήθηκαν και υλοποιήθηκαν στο προγραμματιστικό περιβάλλον Wekaήταν οι *MultinomialNaïveBayes* και *SupportVectorMachines* και πραγματοποιήθηκε σύγκριση των αποτελεσμάτων τους στην Ανάλυση Συναισθήματος. Κατά τη διαδικασία της ταξινόμησης, πειραματιστήκαμε με διάφορες τιμές των παραμέτρων (stopwords, stemmer και attributeselection), για την εύρεση της καλύτερης δυνατής απόδοσης του κάθε μοντέλου κατηγοριοποίησης.

Τα πειραματικά αποτελέσματα έδειξαν πως και οι δύο αλγόριθμοιεπιτυγχάνουν υψηλά ποσοστά κατηγοριοποίησης, με τον ταξινομητή*MultinomialNaïveBayes*να είναι πιο γρήγορος, όσον αφορά τον χρόνο εκπαίδευσης και απάντησηςκαι να πετυχαίνει το καλύτερο ποσοστό, με τις μετρικέςαξιολογήσεις του να μην πέφτουν κάτω από το ***50%***.

Συνοψίζοντας, διαπιστώνουμε πως και οι δύο αλγόριθμοι από την επιβλεπόμενη Μηχανική Μάθηση, *MultinomialNaïveBayes*και *SupportVectorMachines* είναι κατάλληλοι για τη δημιουργία μοντέλων Κατηγοριοποίησης Συναισθήματος με την ιδανική επιλογή του αλγορίθμου να εξαρτάται από τα ποιοτικά και ποσοτικά χαρακτηριστικά του συνόλου δεδομένων, που έχουμε κάθε φορά προς εξέταση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1]*https://en.wikipedia.org/wiki/Data_analysis.*

[2]**Dunham MargaretH.***"Data Mining Introductory and Advanced Topics".* s.l. : Pearson Education Inc, 2002.

[3]**Fayyad Usama, Piatetsky-Shapiro Gregory, and Smyth Padhraic.***"From Data Mining to Knowledge Discovery in Databases".* s.l. : AI Magazine, Volume 17, Number 3, 1996.

[4]**Feldman Ronen and DaganIdo.***"Knowledge Discovery in Textual Databases (KDT)".* s.l. : Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, pp.112-117, 1995.

[5]**Nahm Un Yong, Mooney Raymond J.***"Text Mining with Information Extraction".* s.l. : Proceedings of the 18th National Conference on Artificial Intelligence (AAAI), pp.60-67, 2002.

[6]**Pang Bo and LeeLillian.**"*Opinion Mining and Sentiment Analysis".* s.l. : Foundations and Trends in Information Retrieval, Vol.2, pp.1-135, 2008.

[7] **Soo-Min Kim and EduardHovy.**"*Identifying and analyzing judgment opinions".* s.l. : Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006.

[8] https://en.wikipedia.org/wiki/Social_graph

[9] https://en.wikipedia.org/wiki/Twitter

[10] **Adam Bermingham and Alan F Smeaton.**"*Classifying sentiment in microblogs: is brevity an advantage?*"s.l. : In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM, 2010.

[11] https://el.wikipedia.org/wiki/Κοινωνικό_δίκτυο

[12]**Agarwal Aproorv, Xie Boyi, Vovsha Ilia, Rambow Owen and Passonneau Rebecca. "***Sentiment Analysis of Twitter Data".* s.l. : Proceedings of the Workshop on language in Social Media pp.30-38, 2011.

[13]**Liu, Bing.**"*Sentiment Analysis and Opinion Mining".* s.l. : Morgan & Claypool Publishers, 2012.

[14]**Meena Arun., Prabhakar T.V.**"*Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis".* s.l. : Proceedings of the 29th European Conference on IR Research, pp.573-580, 2007.

[15]**Wilson Theresa, Wiebe Janyce and Hoffmann Paul. "***Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis".* s.l. : In HLT and EMNLP, 2005.

[16]**Turney Peter D.**"*Thumbs Up or Thumbs Down?"*Philadelphia : Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002.

[17]https://www.qubole.com/blog/sentiment-analysis/

[18]https://en.wikipedia.org/wiki/Big_data

[19]**Speriosu, Michael.**"*Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph".* University of Texas at Austin : s.n.

[20]**Hu, X., Tang, L.,Tang, J. and Liu, H.**"*Exploiting social relations for sentiment analysis in microblogging".* 2013.

[21]**Wiebe and Hatzivassiloglou.**"*Effects of Adjective Orientation and Gradability on Sentence Subjectivity".* 2000.

[22]**George Miller and Christiane Fellbaum. "***WordNet. An electronic lexical database".* Cambridge : MA: MIT Press, 1998.

[23]**Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani.**"*SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining".* 2010.

[24]**James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales and Roger J. Booth.**"*The Development and Psychometric Properties of LIWC 2007".* The University of Texas at Austin and The University of Auckland, New Zealand : s.n.

[25]**Brownlee, Jason. "***Supervised and Unsupervised Machine Learning Algorithms".* Machine Learning Algorithms, 2016.

[26]**Alexander Pak and PatrickParoubek.**"*Twitter as a corpus for sentiment analysis and opinion mining".* In LREC, 2010.

[27]**Efthymios Kouloumpis, Theresa Wilson and Johanna Moore.**"*Twitter sentiment analysis: The good the bad and the omg!"*s.l. : ICWSM, 11:538–541, 2011.

[28]**Reynier Ortega, Adrian Fonseca and Andres Montoyo."***Ssa-uo: unsupervised twitter sentiment analysis".* In Second Joint Conference on Lexical and Computational Semantics (* SEM), volume 2, pages 501–507 : s.n., 2013.

[29]**Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu and Tiejun Zhao.**"*Target dependent twitter sentiment classification".* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 151–160. Association for Computational Linguistics, : s.n., 2011.

[30]**Dmitry Davidov, Oren Tsur and Ari Rappoport."***Enhanced sentiment learning using twitter hashtags and smileys".* In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. Association for Computational Linguistics : s.n., 2010.

[31]**Adam Bermingham and Alan FSmeaton.**"*Classifying sentiment in microblogs: is brevity an advantage"*In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836 : s.n., 2010.

[32]**Akshi Kumar and Teeja MarySebastian.**"*Sentiment analysis on twitter".* IJCSI International Journal of Computer Science Issues, 9(3):372–378, 2010.

[33] **Luciano Barbosa and JunlanFeng. "***Robust sentiment detection on twitter from biased and noisy data".* In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44 : Association for Computational Linguistics, 2010.

[34]**Alec Go, Richa Bhayani and Lei Huang.**"*Twitter sentiment classification using distant supervision".* s.l. : CS224N Project Report, Stanford, pages 1–12, 2009.

[35]https://en.wikipedia.org/wiki/Unsupervised_learning

[36]**Ashraf M. Kibrira, Eibe Frank, Bernhard Pfahringer and Gooffrey Holmes. "***Multinomial Naive Bayes for Text Categorization Revisited".*

[37]**Bo Pang and Lillian Lee, Shivakumar Vaithyanathan.**"*Thumbs up? Sentiment Classification using Machine Learning Techniques".*

[38] http://weka.wikispaces.com/ARFF%20%28book%20version%29

[39] http://repfiles.kallipos.gr/html_books/246/Ch4.html

[40] https://en.wikipedia.org/wiki/Orange_(software)