ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναλύοντας τις Πτυχές Ανάλυσης Συναισθήματος

Συγγραφέας:
Αϊβατογλου Γεωργιος

Επιβλέποντες:
Χοτστης Ηλιας
Βαβαλης Εμμανουηλ

Βόλος, Ιανουάριος 2018

To my family and friends

Περίληψη

Η ανάλυση συναισθήματος είναι η διαδικασία κατά την οποία γίνεται αναγνώριση συμπεριφοράς, συναισθήματος και κρίσης. Τα τελευταία χρόνια με την ανοδική πορεία των κοινωνικών δικτύων και γενικότερα του Internet, οι άνθρωποι μοιράζονται διαδικτυακά πολλές πληροφορίες καθιστώντας την ανάλυση αυτών απαραίτητη [3]. Γενικά οι άνθρωποι εμπιστεύονται τις κριτικές προϊόντων και υπηρεσιών στο Internet όσο εμπιστεύονται και την κρίση κάποιου κοντινού τους ανθρώπου [26]. Βέβαια υπάρχουν αρκετά εμπόδια που πρέπει να αντιμετωπιστούν αφού η κατανόηση της ανθρώπινης γλώσσας από τους υπολογιστές δεν είναι και τόσο εύκολη υπόθεση.

Η παρούσα διπλωματική πραγματεύεται το πρόβλημα της ανάλυσης συναισθήματος και συγκεκριμένα της αυτόματης κατηγοριοποίησης κειμένου ως θετικό ή αρνητικό με γνώμονα την άποψη του συγγραφέα πάνω στο θέμα συζήτησης. Πιο συγκεκριμένα γίνεται ανάλυση δεδομένων από το κοινωνικό δίκτυο Twitter και από Rss Feeds διαδικτυακών εφημερίδων με σκοπό την εξαγωγή συμπερασμάτων.

Με βάση τα παραπάνω, έγιναν δύο διαφορετικές υλοποιήσεις. Η πρώτη έχει να κάνει με ανάλυση tweets για εξαγωγή άποψης και κρίσης σχετικά με το MacBook της Apple. Πέρα από την εξαγωγή της συνολικής κοινής γνώμης σχετικά με αυτό το προϊόν, δόθηκε βάση στα tweets με αρνητική διάθεση και βρέθηκαν τα συγκεκριμένα θέματα στα οποία οι χρήστες εκφράζουν δυσαρέσκεια. Αυτό έχεις ως σκοπό να δώσει στις εταιρείες μια προδιάθεση σχετικά με τις αλλαγές που πρέπει να γίνουν σε μελλοντικά προϊόντα. Η δεύτερη υλοποίηση έγινε προκειμένου να γίνει έρευνα για τυχόν συσχέτιση μεταξύ διαφόρων οικονομικών άρθρων και της τιμής μιας μετοχής την ίδια χρονική περίοδο [4] [9]. Πραγματοποιήθηκαν τέσσερις διαφορετικές μέθοδοι. Οι πρώτες τρεις σχετίζονται με αλγορίθμους επιβλεπόμενης μηχανικής μάθησης, ενώ η τέταρτη σε αλγόριθμο βασισμένο σε λεξικό.

Συμπερασματικά φαίνεται ότι στην πρώτη υλοποίηση ένα μοντέλο επιβλεπόμενης μηχανικής μάθησης είναι αρκετά έμπιστο στην κατηγοριοποίηση tweets, με το ποσοστό ακρίβειας να κυμαίνεται στο 75%. Η δεύτερη ανάλυση αντιθέτως αποδεικνύει ότι από την μία ένα μοντέλο μπορεί να δώσει κάποια ικανοποιητικά αποτελέσματα αλλά από την άλλη η τιμή μιας μετοχής μεταβάλλεται από παράγοντες που δεν είναι εύκολα ελέγξιμοι.

Λέξεις Κλειδιά : Μηχανική Μάθηση, Ανάλυση Άποψης, Συναισθηματική Ανάλυση, Επιβλεπόμενη Μηχανική Μάθηση, Ανάλυση Άποψης Βασισμένη σε Λεξικό

Abstract

Sentiment Analysis is the process whereby objective information such as behavior, emotion and judgment are automatically identified. In modern years, with the upward trend of social media and internet, people share a great deal of information online, making the analysis of them vital [3]. In general people trust online reviews as much as they trust the judgment of a friend [26]. Of course there are major hurdles that need to be addressed, as comprehension of human language by computers is not that easy.

This thesis deals with the problem of Sentiment Analysis and in particular with the automatic categorization of text as positive or negative according to the author's point of view. Particularly, we analyze data from Twitter and websites in order to draw conclusions.

Based on the above, two different implementations have been developed on Sentiment Analysis. The first one has to do with tweets and Apple's MacBook. In addition, except only finding the public opinion about this product, we are focusing on the negative tweets and the issues with which users express dissatisfaction. This is intended to give companies a predisposition for changes on upcoming products. Through the second implementation we investigate any correlation between financial articles and the price of a share over the same period [4] [9]. Four different methods have been developed. The first three are supervised machine learning algorithms and the fourth one is sentiment analysis based on lexicon.

In conclusion, supervised machine learning models are enough to reliably categorize tweets, with accuracy percentage ranging to 75%, while the second analysis shows that it could have some satisfactory results but the price of a share depends on many factors that are not easy controllable.

Keywords : Machine Learning, Opinion Analysis, Sentiment Analysis, Naive Bayes, Supervised Machine Learning, Sentiment Analysis Based on Lexicon

Περιεχόμενα

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Η γλώσσα είναι ένας τρόπος επικοινωνίας μεταξύ των ανθρώπων όταν θέλουν να εκφράσουν αυτό που νιώθουν. Ένα μέρος αυτής της επικοινωνίας είναι και το συναίσθημα που κρύβει η κάθε λέξη, το οποίο αποδίδει έμφαση στα λεγόμενα. Κατά την διάρκεια της επικοινωνίας ο ομιλητής δημιουργεί ένα μήνυμα και το αποστέλλει στον ακροατή και αυτός με την σειρά του το αποκωδικοποιεί. Η επικοινωνία αυτή μπορεί να είναι μέσω φωνής ή και κειμένου ανάλογα την περίσταση. Η ψηφιοποίηση των μηνυμάτων υπάρχει εδώ και αρκετά χρόνια, όμως απαιτείται ανθρώπινος παράγοντας ώστε να είναι σε θέση να αποκωδικοποιήσει και τα κατανοήσει αυτά τα μηνύματα. Η αυτοματοποίηση αυτής της διαδικασίας είναι στο επιστημονικό επίκεντρο τα τελευταία χρόνια, με σκοπό να μπορούν οι υπολογιστές να αναλύσουν αυτά τα μηνύματα και να πάρουν ανάλογες αποφάσεις σχετικά με το συναίσθημα που εκφράζουν. Για το λόγο αυτό η ανάλυση συναισθήματος είναι ένας πολύ σημαντικός τομέας στην επιστήμη των υπολογιστών όπου χρησιμοποιείται όλο και περισσότερο για την αντιμετώπιση προκλήσεων που έχουν να κάνουν με την επικοινωνία των ανθρώπων.

Την εποχή που διανύουμε περνάμε όλο και περισσότερο χρόνο αλληλεπιδρώντας με υπολογιστές. Συγκεκριμένα χρησιμοποιώντας το Internet, το οποίο έχει εξελιχθεί σε ταχύτητα, είναι πολύ εύκολο για τον καθένα πια να αναζητήσει πληροφορίες. Όμως, η αναζήτηση πληροφοριών σε αυτό το χαώδες σύνολο δεδομένων δεν είναι εύκολη, ειδικά όταν απαιτούμε αμεσότητα και εγκυρότητα στις πληροφορίες που μας ενδιαφέρουν [12]. Σε αυτό το πρόβλημα γίνεται προσπάθεια επίλυσης μέσω της ανάλυσης συναισθήματος ώστε τα δεδομένα πριν φτάσουν στον τελικό χρήστη να έχουν φιλτραριστεί, παρέχοντας του μόνο αυτά που τον ενδιαφέρουν.

Την τελευταία δεκαετία έχει σημειωθεί ραγδαία αύξηση στον συγκεκριμένο τομέα εξαιτίας δύο παραγόντων: της υπολογιστικής ισχύς που διαθέτουμε και του τεράστιου όγκου δεδομένων που αναρτάται καθημερινά στο Internet. Σε συνδυασμό, ερευνητές

είχαν την δυνατότητα να δημιουργήσουν τεράστια λεξικά αποτελούμενα από χιλιάδες λέξεις το κάθε ένα σε διάφορες γλώσσες, δίνοντάς την δυνατότητα δημιουργίας αλγόριθμων μηχανικής μάθησης για την κατανόηση των γλωσσών και της συναισθηματικής πολικότητας των λέξεων. Φυσικά κάθε γλώσσα έχει την δική της σύνταξη και γραμματική και γι΄ αυτό θα πρέπει να δίνεται ιδιαίτερη προσοχή στους κανόνες της για να έχουμε τα αποτελέσματα που θέλουμε.

Σε αυτή την διπλωματική εργασία γίνεται προσπάθεια επεξεργασίας και ανάλυσης δεδομένων με σκοπό την εκπαίδευση αλγόριθμων ώστε να βρίσκονται σε θέση να αναγνωρίζουν, να κατανοούν και να αντιδρούν σε συναισθηματικά ερεθίσματα που τους παρέχονται από πληθώρα πληροφοριών. Φυσικά τα αποτελέσματα δεν ευνοούν μόνο ανεξάρτητους αλλά και εταιρείες οι οποίες μπορούν να τα εκμεταλλευτούν για να έχουν μια αντικειμενική εικόνα για αυτές αλλά και για ανταγωνιστές [5].

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

2.1 Κοινωνικά δίκτυα

Τα τελευταία χρόνια τα κοινωνικά δίκτυα (Social Networks) έχουν κατακτήσει ένα σημαντικό κομμάτι της χρήσης του internet φέρνοντας μεγάλες αλλαγές στον τρόπο με τον οποίο οι άνθρωποι μοιράζονται αυτό που αισθάνονται. Αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνήσουν και να μοιραστούν την άποψη τους χωρίς περιορισμούς. Κατά την πλειοψηφία παρέχονται δωρεάν στο ευρύτερο κοινό με μοναδική προϋπόθεση ο χρήστης να δημιουργήσει ένα προσωπικό λογαριασμό. Τα πιο γνωστά κοινωνικά δίκτυα είναι το Facebook, το Instragram, το LinkedIn και το Twitter και κάθε ένα ξεχωριστά έχει τον δικό του τρόπο προσέγγισης και χρήσης. Έχουν γίνει απαραίτητο στοιχείο του marketing σε όλο τον κόσμο και χρήζουν ανάλυσης και εκμετάλλευσης από κάθε εταιρεία ανεξαιρέτως [14]. Τα χαρακτηριστικά που οδήγησαν σε αυτήν την αναγκαιότητα είναι ο πολύ μεγάλος όγκος χρηστών που διαθέτουν, η εύκολη και γρήγορη διαφήμιση μέσα από αυτά και ο μεγάλος αριθμός προσωπικών πληροφοριών που μοιράζονται οι χρήστες καθημερινά.

Συγκεκριμένα το Twitter είναι ένας χώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες την δημοσιοποίηση σύντομων μηνυμάτων, έως 280 χαρακτήρες, τα οποία ονομάζονται tweets [17]. Η υπηρεσία ξεκίνησε τον Μάρτιο του 2006 και απαριθμεί περίπου 330 εκατομμύρια συνδρομητές. Επιτρέπει την αλληλεπίδραση με τα δεδομένα του μέσω μιας διεπαφής (Application programming interface, API). Η συγκεκριμένη διεπαφή ενώ παρέχει τις περισσότερες λειτουργίες του Twitter στο κοινό, δεν επιτρέπει την πρόσβαση στον κώδικα που υλοποιεί αυτές τις υπηρεσίες. Για την εκμετάλλευση αυτής της διεπαφής χρειάζεται μια γλώσσα προγραμματισμού όπως php, ruby ή python για την δημιουργία αιτήσεων και τα δεδομένα που θα επιστραφούν θα είναι σε μορφή αρχείου JSON.

2.2 Εξόρυξη δεδομένων

Η επιστήμη που ασχολείται με την εξόρυξη πληροφοριών από τα κοινωνικά δίκτυα και όχι μόνο, ονομάζεται εξόρυξη δεδομένων (Data Mining). Είναι ένας επιστημονικός κλάδος που βασίζεται σε αρχές στατιστικής, τεχνητής νοημοσύνης και μηχανικής μάθησης και στόχος της είναι η συλλογή πληροφοριών σε δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει κατάλληλες αποφάσεις.

Όταν γίνεται αναφορά σε δεδομένα κοινωνικών δικτύων, είναι αντιληπτό ότι ο όγκος των πληροφοριών είναι τεράστιος. Στο μέλλον, δεδομένου ότι η ιδέα του Internet of Things γίνεται εντονότερη, ο κλάδος των Big Data γίνεται όλο και σημαντικότερος. Η ιδέα πίσω από το Internet of Things είναι η σύνδεση και η ανταλλαγή πληροφοριών όλων των ηλεκτρονικών συσκευών μεταξύ τους μέσω του Internet.

Η συλλογή δεδομένων από κοινωνικά δίκτυα, οχήματα, αισθητήρες ή μικροσυσκευές γίνονται ζωτικής σημασίας για κάθε εταιρεία που θέλει να είναι ανταγωνιστική. Πιο συγκεκριμένα οι εταιρείες χρησιμοποιούν δεδομένα από τα κοινωνικά δίκτυα ώστε να είναι σε θέση να κατηγοριοποιήσουν τους πελάτες τους και να καταλάβουν την αγοραστική τους συνήθεια ώστε να τους προωθούν συγκεκριμένα προϊόντα μέσω σχετικών διαφημίσεων, με στόχο την επίτευξη κέρδους. Η ανάγκη όμως για δεδομένα δεν σταματάει στα κοινωνικά δίκτυα. Οι περισσότερες σύγχρονες εταιρείες αντλούν πληροφορίες από το ίδιο τους το ιστορικό, στηρίζοντας το μελλοντικό τους πλάνο εξ' ολοκλήρου σε αυτές. Για αυτό τον λόγο δημιουργήθηκαν και νέοι όροι που συνδυάζουν τον τομέα της ανάλυσης πληροφοριών και των επιχειρήσεων όπως Επιχειρησιακή Νοημοσύνη (Business Intelligence) και Επιχειρησιακές Αναλύσεις (Business Analytics).

Επιχειρησιακές αναλύσεις (ΕΑ) είναι τεχνικές και τεχνολογίες που χρησιμοποιούνται για την εξερεύνηση ιστορικών πληροφοριών μιας εταιρείας με σκοπό την δημιουργία νέων χρήσιμων στοιχείων για την βελτίωση της απόδοσης και του μελλοντικού επαγγελματικού πλάνου. Από την άλλη, ο όρος Επιχειρησιακή νοημοσύνη (ΕΝ), αν και είναι πολύ κοντά στον προηγούμενο, διαφέρει στο ότι επικεντρώνεται περισσότερο στο ερώτημα του πώς θα γίνει βελτίωση. Οι τεχνικές αυτές δίνουν την δυνατότητα σε μια εταιρεία να εφαρμόζει μετρήσεις σε τεράστια σύνολα δεδομένων, να καλύπτει αναζητήσεις σε αυτά, να αντλεί δεδομένα, να παρακολουθεί την επίδοσή της, καθώς και μελλοντικές προβλέψεις και αναλυτικές προδιαγραφές νέων υπηρεσιών. Με άλλα λόγια, τόσο οι ΕΑ όσο και η ΕΝ αντιμετωπίζουν τα ίδια προβλήματα, αλλά αν υπάρχει απόθεμα τεράστιου όγκου δεδομένων για ανάλυση με σκοπό την εξαγωγή συμπερασμάτων, τότε τα εργαλεία και οι τεχνικές που θα χρησιμοποιηθούν εμπίπτουν στην ΕΝ.

2.3 Ανάλυση συναισθήματος

Για να γίνει εφικτή αυτή η εκμετάλλευση των προσωπικών πληροφοριών που μοιράζονται οι άνθρωποι στα κοινωνικά δίκτυα δεν επαρκεί μόνο η συλλογή αυτών των δεδομένων αλλά και η ανάλυση τους. Μία ανάλυση η οποία θα μπορούσε να κατανοήσει την σημασία αυτών των δεδομένων που δημοσιοποιούν οι χρήστες και να παράγει αποτελέσματα, ονομάζεται ανάλυση συναισθήματος (Sentiment Analysis) και είναι ένας τομέας της επιστήμης των υπολογιστών και της γλωσσολογίας. Κάνει χρήση της ανάλυσης φυσικής γλώσσας με σκοπό τον εντοπισμό, την αναγνώριση, την εξαγωγή και την μελέτη σημαντικών πληροφοριών από κείμενο και ήχο. Όπως προδίδει και το όνομα της, δεν χρησιμοποιείται μόνο στον εντοπισμό συναισθήματος, αλλά και στην αναγνώριση αντικειμενικότητας ή υποκειμενικότητας αλλά ακόμα και στον καθορισμό γνώμης για διάφορα χαρακτηριστικά. Αυτά τα χαρακτηριστικά μπορεί να είναι προϊόντα ή υπηρεσίες παρέχοντας έτσι ένα πολύ ισχυρό εργαλείο σε επιχειρήσεις κάθε μορφής. Δίνει την δυνατότητα εκτίμησης ελαττωμάτων σε προϊόντα ώστε να γίνει διόρθωση στη νέα έκδοσή τους. Τέλος, θα μπορούσε να γίνει χρήση ακόμα και για ανάλυση προηγούμενων εγχειρημάτων άλλων εταιρειών ώστε να αποφευχθούν παρόμοια λάθη.

2.3.1 Πολικότητα λέξεων

Η εύρεση της πολικότητας σε μία πρόταση αρκετές φορές μπορεί να μην είναι εύκολη υπόθεση. Οι περισσότερες από αυτές είναι πολύπλοκες συνδυάζοντας περισσότερα του ενός νοήματα. Για παράδειγμα στην πρόταση *η μπαταρία του κινητού Α είναι καλύτερη από την μπαταρία του κινητού Β* γίνεται κατανοητό ότι η μπαταρία του κινητού Β μάλλον δεν είναι και τόσο καλή, τουλάχιστον σε σχέση με του Α. Επίσης πολλές λέξεις έχουν διαφορετική συναισθηματική πολικότητα ανάλογα με το θέμα αναφοράς. Δηλαδή χαρακτηρίζοντας μικρό σε διαστάσεις ένα κινητό τηλέφωνο μάλλον είναι θετικής σημασίας, ενώ αντίθετα χαρακτηρίζοντας μικρό ένα δωμάτιο ξενοδοχείου μάλλον αρνητικής. Ένα μέρος του λόγου που χρήζει ιδιαίτερης αντιμετώπισης είναι η άρνηση. Συνήθως η άρνηση μπορεί να αντιστρέψει ολόκληρο το νόημα μιας πρότασης επηρεάζοντας την πολικότητα. Ομοίως και οι λέξεις που δείχνουν ένταση και συχνότητα.

Ακόμα μία ενδιαφέρουσα πρόκληση είναι η ειρωνεία. Ένας άνθρωπος μπορεί να την κατανοήσει συνδυάζοντας πολλές πληροφορίες όπως γνώση πάνω στο αντικείμενο συζήτησης και αρκετή εμπειρία. Όμως ένας υπολογιστής δεν μπορεί με ευκολία να την διακρίνει. Ένας πιθανός τρόπος αντιμετώπισης που μπορεί να χρησιμοποιηθεί, χωρίς όμως να λειτουργεί πάντα σωστά, είναι αναλύοντας αναφορές για το θέμα συζήτησης όπου αν η πληθώρα από αυτές είναι αρνητικές και μία είναι εξαιρετικά θετική, τότε μάλλον πρόκειται για ειρωνεία, χωρίς όμως αυτό να είναι βέβαιο.

2.4 Ανάλυση κειμένου

Η ανάλυση ενός κειμένου μπορεί να γίνει με τρεις διαφορετικούς τρόπους:

- Επίπεδο κειμένου (Document Level)
- Επίπεδο πρότασης (Sentence Level)
- Επίπεδο χαρακτηριστικών (Feature-Aspect Level)

Η ανάλυση σε επίπεδο κειμένου χρησιμοποιείται για την απόφανση συναισθηματικής πολικότητας ενός ολόκληρου κειμένου. Όπως είναι κατανοητό το κείμενο αντιμετωπίζεται σαν ενιαία οντότητα με συγκεκριμένο θέμα αναφοράς. Η ανάλυση σε επίπεδο πρότασης περιπλέκει λίγο την ανάλυση χωρίζοντας το συνολικό κείμενο σε μικρότερες προτάσεις και γίνεται ανάλυση σε κάθε πρόταση ξεχωριστά. Τέλος, η ανάλυση σε επίπεδο χαρακτηριστικών προσπαθεί να εντοπίσει τα θέματα για τα οποία εκφέρεται άποψη μέσα στο κείμενο [22].

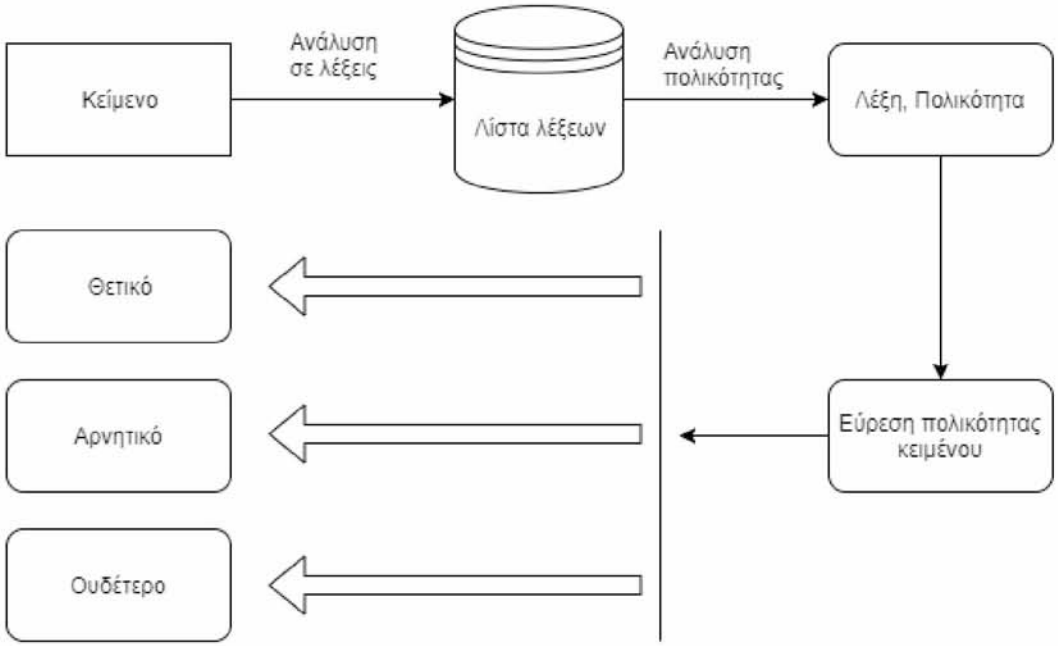
2.4.1 Ανάλυση σε επίπεδο κειμένου

Αυτή είναι η απλούστερη μορφής ανάλυσης. Ολόκληρο το κείμενο θεωρείται ως μία μονάδα πληροφορίας που έχει μια συναισθηματική πολικότητα ως προς ένα θέμα, όπως ένα βιβλίο ή μία ταινία. Η ανάλυση αυτής της μορφής δεν προτείνεται όταν στο κείμενο υπάρχουν αναφορές σε παραπάνω του ενός θέματα και φυσικά αποφεύγεται σε αναλύσεις κοινωνικών δικτύων. Έτσι η κατηγοριοποίηση του κειμένου ως θετικό ή αρνητικό γίνεται αναλύοντας ολόκληρο το κείμενο θεωρώντας το αντικείμενο συζήτησης μοναδικό. Αυτού του είδους η προσέγγιση βοηθά στη λήψη αποφάσεων παρέχοντας την συνολική πολικότητα του κειμένου μετρώντας το σύνολο των θετικών και αρνητικών απόψεων μέσα σε αυτό.

2.4.2 Ανάλυση σε επίπεδο πρότασης

Η ανάλυση επιπέδου πρότασης είναι η πιο συνήθης και έχει πολύ ικανοποιητικά αποτελέσματα. Πραγματοποιείται διασπώντας το κείμενο σε προτάσεις θεωρώντας την κάθε μία ως ξεχωριστή οντότητα με το δικό της αντικείμενο συζήτησης και την δικής της πολικότητα. Επίσης μία πρόταση μπορεί να χαρακτηρισθεί ως υποκειμενική ή αντικειμενική. Ο πρώτος χαρακτηρισμός αποδίδεται σε προτάσεις που περιέχουν πολικότητα και πρέπει να λαμβάνονται υπόψιν στην ανάλυση, ενώ ο δεύτερος περιέχει πληροφορία που δεν χρησιμοποιείται αφού στηρίζεται σε γεγονότα χωρίς να αποδίδεται γνώμη.

Σχήμα 2.1: Ανάλυση σε επίπεδο κειμένου



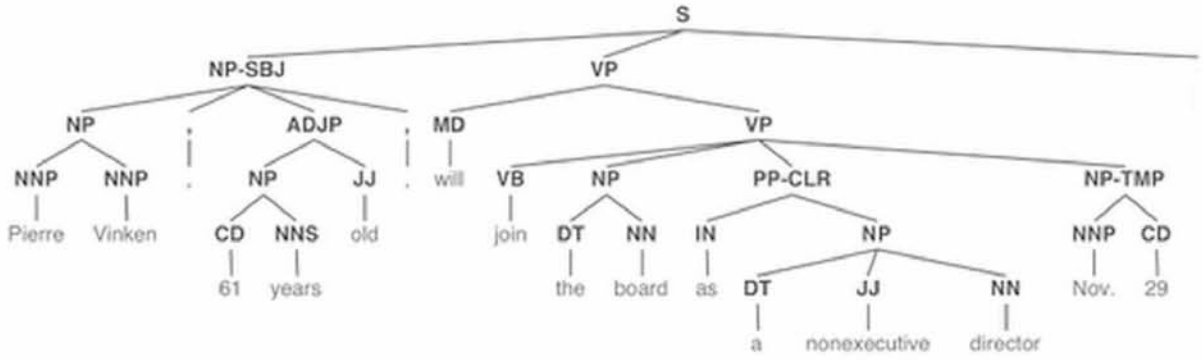
2.4.3 Ανάλυση σε επίπεδο χαρακτηριστικών

Σε επίπεδο κειμένου και σε επίπεδο πρότασης η ανάλυση δεν ανακαλύπτει τι ακριβώς αρέσει και τι όχι από το αντικείμενο συζήτησης. Γι΄ αυτό τον λόγο η ανάλυση σε επίπεδο χαρακτηριστικών δεν αναζητά δομές λόγου όπως κείμενα και προτάσεις, αλλά επικεντρώνεται στην αναζήτηση της άποψης [16]. Βασίζεται στην ιδέα ότι μία άποψη αποτελείται από συναισθηματική πολικότητα, αρνητική ή θετική, και έναν στόχο. Στην ανάλυση αυτής της μορφής γίνεται προσπάθεια εντοπισμού των χαρακτηριστικών ενός προϊόντος από ένα κομμάτι λόγου. Για παράδειγμα στην πρόταση *Η μπαταρία του κινητού Α διαρκεί πάρα πολύ*, το κινητό Α είναι το θέμα συζήτησης, η μπαταρία είναι το χαρακτηριστικό για το οποίο γίνεται λόγος και το 'διαρκεί πάρα πολύ' είναι η άποψη.

2.5 Ανάλυση σε μέρη του λόγου

Η ανάλυση σε μέρη του λόγου (Part of Speech Tagging) θεωρείται βασικό κομμάτι στην επεξεργασία φυσικής γλώσσας. Σκοπό έχει να αποφανθεί για κάθε λέξη μέσα στο κείμενο σε ποιο μέρος του λόγου ανήκει. Μιλώντας όμως για κείμενα κοινωνικών δικτύων, η δυσκολία πολλαπλασιάζεται αφού η γλώσσα που χρησιμοποιείται εκεί είναι η καθομιλουμένη. Λόγω της μεγάλης σε έκταση χρήσης της συναισθηματικής ανάλυσης, υπάρχουν πολλά εργαλεία και βιβλιοθήκες που βοηθούν και επιταχύνουν αυτή την διαδικασία. Ένα από αυτά είναι και το Natural Language Toolkit, NLTK. Πρόκειται για μία πλατφόρμα που βοηθά στην υλοποίηση προγραμμάτων σχετικά με την επεξεργασία φυσικής γλώσσας σε περιβάλλον της python.

Σχήμα 2.2: Διάσπαση και ανάλυση κειμένου σε μέρη του λόγου



2.6 Ανάλυση σε λεκτικές μονάδες

Η ανάλυση σε λεκτικές μονάδες (Tokens) είναι η διαδικασία μετατροπής μίας πρότασης ή ενός κειμένου σε μία σειρά μεμονωμένων λέξεων. Είναι πολύ σημαντική στην ανάλυση συναισθήματος από κείμενο από τη στιγμή που το συναίσθημα προσδιορίζεται από συγκεκριμένες λέξεις και δεν είναι εύκολα αντιληπτό από έναν υπολογιστή. Η επιλογή οριοθέτησης στις περισσότερες περιπτώσεις είναι το κενό ανάμεσα από τις λέξεις. Ακόμα πρέπει να λαμβάνεται υπόψιν το θαυμαστικό ως έμφαση της λέξης που το εμπεριέχει και το ερωτηματικό ως αβεβαιότητα. Το κόμμα, η άνω τελεία και οι παρενθέσεις είναι επίσης σημάδια που δείχνουν ότι οι λέξεις έχουν μια συνοχή μεταξύ τους και πρέπει να αναλύονται μαζί ως μία οντότητα.

Σχήμα 2.3: Ανάλυση πρότασης σε λεκτικές μονάδες

2.7 Αποκοπή καταλήξεων

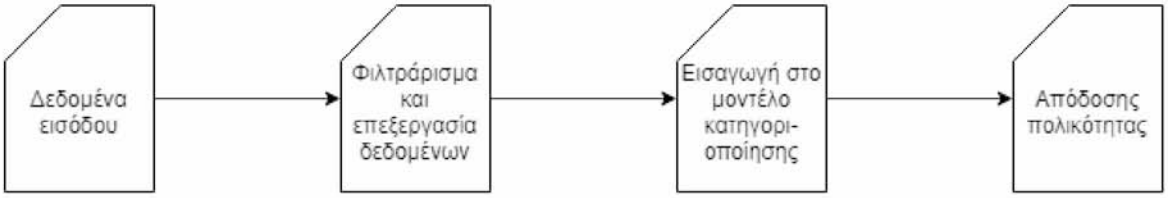
Στην εξόρυξη δεδομένων η αποκοπή καταλήξεων (Stemming) είναι η διαδικασία μετατροπής λέξεων στην λεκτική τους ρίζα. Αυτό βοηθάει στην μείωση του συνόλου των λέξεων προς ανάλυση και στην αυτόματη κατηγοριοποίηση κοινών λέξεων που ουσιαστικά το νόημα τους είναι ίδιο. Από την στιγμή που μας ενδιαφέρει το συναίσθημα που εκφράζουν οι λέξεις, μικρές αλλαγές στις λέξεις μπορεί αν αποβούν μοιραίες

στην τελική τους έννοια. Οι τρεις περισσότερο γνωστοί αλγόριθμοι αποκοπής καταλήξεων είναι ο Porter Stemmer, ο Lancaster Stemmer και ο WordNet Stemmer. Οι δύο πρώτοι δουλεύουν παρόμοια εντοπίζοντας καταλήξεις λέξεων, αποκόβοντάς τες και τέλος πραγματοποιούν μια κανονικοποίηση σε αυτές. Ο τελευταίος διαφέρει στο ότι χρειάζεται πέρα από την λέξη και το μέρος του λόγου στο οποίο ανήκει σαν είσοδο. Αξίζει να αναφερθεί ότι πολλές φορές η διαδικασία του stemming μετατρέπει λέξεις σε μορφή που δεν είναι κατανοητή προς τον άνθρωπο. Κάτι τέτοιο δεν αποτελεί πρόβλημα για έναν κατηγοριοποιητή μηχανικής μάθησης αφού και το corpus από το οποίο θα γίνει η εκπαίδευση του θα περάσει από διαδικασία stemming και αυτό. Όμως ένα μοντέλο που βασίζεται σε λεξικό ίσως να μην αποδώσει τόσο καλά, μειώνοντας φανερά την ακρίβεια του.

2.8 Φιλτράρισμα των tweets

Χρησιμοποιώντας ως πηγή πληροφορίας τα κοινωνικά δίκτυα, τα δεδομένα που αντλούνται δεν ακολουθούν την πρότυπη μορφή γλώσσας που συναντάμε σε επιστημονικά άρθρα. Επίσης είναι πολύ συνηθισμένη η αποκοπή γραμμάτων από λέξεις για συντομογραφία. Έτσι γίνεται κατανοητό, ότι πριν την τροφοδότηση των δεδομένων στον κατηγοριοποιητή για εκμάθηση, χρειάζεται μια διαδικασία καθαρισμού και ομοιογένειας των δεδομένων (Cleansing). Αυτή η διαδικασία συχνά περιλαμβάνει την αφαίρεση links και ονομάτων, των σημείων στίξης, των πιο συχνά εμφανιζόμενων λέξεων, των διπλότυπων, την μετατροπή των γραμμάτων σε πεζά και αν χρειαστεί την μετάφραση. Η διαδικασία φιλτραρίσματος είναι σχεδόν ζωτικής σημασίας. Ο θόρυβος που περιέχουν τα tweets μπορεί να αποβεί μοιραίος στην απόδοση του κατηγοριοποιητή, καθιστώντας τον μη αποδοτικό. Όπως είναι κατανοητό τα δεδομένα εισόδου θα πρέπει να είναι στην απλούστερη μορφή που μπορούν να έρθουν μετά το φιλτράρισμα, ώστε να μειωθεί η γκάμα διαφορετικών λέξεων. Η αφαίρεση των links, των επισημάνσεων άλλων χρηστών, των συχνότερα εμφανιζόμενων λέξεων και των διπλότυπων αγνοείται ολοκληρωτικά, όμως τα σημεία στίξης και τα emoticons άλλες φορές αποδίδουν θετικά στην απόδοση και άλλες αρνητικά.

Σχήμα 2.4: Διαδικασία ανάλυσης συναισθήματος



Κεφάλαιο 3

Αλγοριθμικό υπόβαθρο

3.1 Μηχανική μάθηση

Η ανάλυση συναισθήματος είναι ένα πρόβλημα κατηγοριοποίησης, αφού το κείμενο μπορεί να διακριθεί σε θετικό, αρνητικό ή ουδέτερο. Η συνήθης τεχνική είναι με την χρήση αλγόριθμων μηχανικής μάθησης (Machine Learning). Στο επιστημονικό πεδίο της ανάλυσης δεδομένων η μηχανική μάθηση χρησιμοποιείται για την δημιουργία πολύπλοκων αλγοριθμικών μοντέλων με απώτερο σκοπό την πρόβλεψη, η αλλιώς την κατηγοριοποίηση της εισόδου. Οι αλγόριθμοι μηχανικής μάθησης διακρίνονται σε τρεις μεγάλες κατηγορίες, την επιβλεπόμενη μάθηση (Supervised Learning), την μάθηση χωρίς επίβλεψη (Unsupervised Learning) και τέλος την ενισχυτική μάθηση (Reinforcement Learning).

3.1.1 Μηχανική μάθηση με επίβλεψη

Η επιβλεπόμενη μηχανική μάθηση είναι μια κατηγορία μάθησης με στόχο την κατηγοριοποίηση νέων και άγνωστων δεδομένων ύστερα από εκπαίδευση με γνωστά ήδη κατηγοριοποιημένα δεδομένα (Training Set). Τα κατηγοριοποιημένα δεδομένα είναι ένα σύνολο παραδειγμάτων όπου αφού τροφοδοτηθούν στο μοντέλο το εκπαιδεύουν. Αποτελούνται συνήθως από ένα σύνολο εισόδου και μια επιθυμητή τιμή εξόδου. Το μοντέλο ύστερα, αφού έχει εκπαιδευτεί, είναι σε θέση να κατηγοριοποιήσει τα νέα δεδομένα στις αντίστοιχες κλάσεις με γνώμονα τις πληροφορίες που άντλησε από το σετ παραδειγμάτων.

Μερικοί διαδεδομένοι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιούνται στην ανάλυση συναισθήματος από κείμενο είναι:

- Support Vector Machines
- Naive Bayes
- Maximum Entropy

Support Vector Machines

Οι μηχανές διανυσματικής υποστήριξης (Support Vector Machines) είναι μοντέλα επιτηρούμενης μηχανικής μάθησης που αναλύουν δεδομένα για κατηγοριοποίηση [13]. Προσπαθούν να δημιουργήσουν ένα υπερεπίπεδο (hyperplane), όσον αφορά την απόσταση μεταξύ των δύο κλάσεων (θετικά ή αρνητικά συναισθήματα) με τον καλύτερο δυνατό τρόπο. Το διάνυσμα διαχωρισμού δίνεται από:

$$< \vec{w}\vec{x} > +b = \sum_i y_i a_i < \vec{x_i}\vec{x} > +b = 0$$

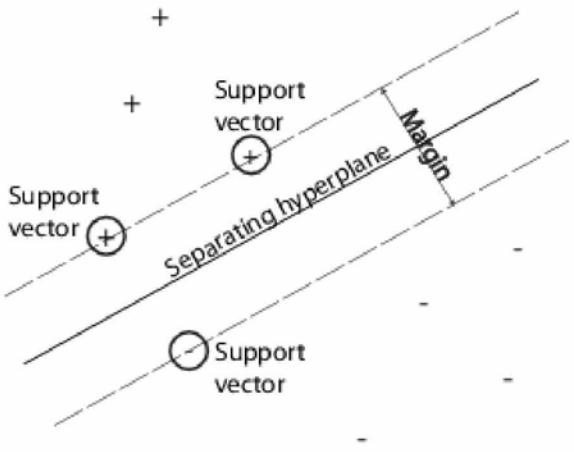
όπου το $\vec{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$ είναι το n-διάστατο διάνυσμα εισόδου, το y_i είναι το αποτέλεσμα του, $\vec{w} = (w_1, w_2, ..., w_n)$ είναι τα αντίστοιχα βάρη που προσδιορίζουν το hyperplane και a_i είναι οι Langarian πολλαπλασιαστές που χρησιμοποιούνται για τον εντοπισμό τοπικών ελαχίστων και μεγίστων μιας συνάρτησης. Από την στιγμή που δημιουργείται το hyperplane από ένα σετ παραδειγμάτων (Training Set) ένα νέο $\vec{x_i}$ δεδομένο μπορεί να προσδιοριστεί ως:

1. Εάν $\vec{w}\vec{x_i} + b \geq 0$ τότε ανήκει στην θετική κλάση.
2. Σε οποιαδήποτε άλλη περίπτωση ανήκει στην αρνητική.

Ένα κλασικό μοντέλο SVM μπορεί να κατηγοριοποιήσει μόνο εάν υπάρχουν το πολύ δύο κλάσεις. Για την αντιμετώπιση προβλημάτων με περισσότερες κλάσεις, χρειάζεται μια διαδικασία μετατροπής του προβλήματος σε πρόβλημα δύο κλάσεων. Αυτό επιτυγχάνεται με δύο τρόπους:

1. Ο πρώτος τρόπος ονομάζεται ένας εναντίων όλων (One vs All) και ο SVM κατηγοριοποιητής χτίζεται για δύο κλάσεις. Έτσι παίρνει την πρώτη κλάση ως θετική και όλες τις υπόλοιπες κλάσεις ως την αρνητική. Έτσι ένα νέο δεδομένο X κατηγοριοποιείται στην κλάση N εάν απορριφθεί από όλες τις υπόλοιπες.
2. Ο δεύτερος τρόπος ονομάζεται ένας εναντίον ενός (One vs One) και ο SVM χτίζεται για ζεύγη κλάσεων. Αφού υπάρχουν $0.5N(N-1)$ πιθανά ζευγάρια για N κλάσεις θα δημιουργηθούν παραπάνω κατηγοριοποιητές από ότι προηγουμένως.

Σχήμα 3.1: Μηχανές Διανυσματικής Υποστήριξης



Όπως φαίνεται στο σχήμα υπάρχουν δύο υπερεπίπεδα για τα οποία αν τα δεδομένα πάρουν την τιμή $+1$ ή -1 θα βρίσκονται επάνω ή κάτω αντίστοιχα από αυτά. Τα οριακά σημεία ονομάζονται support vectors.

Naive Bayes

Ο αλγόριθμος Naive Bayes βασίζεται στο θεώρημα του Bayes με την παραδοχή ότι όλα τα γεγονότα είναι ανεξάρτητα μεταξύ τους. Ο συγκεκριμένος αλγόριθμος κατηγοριοποίησης υποθέτει ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε ένα σύνολο δεδομένων δεν συσχετίζεται με τα υπόλοιπα χαρακτηριστικά. Χρησιμοποιεί τις υποθετικές πιθανότητες, δηλαδή ποια είναι η πιθανότητα να συμβεί ένα γεγονός Α εάν ένα άλλο γεγονός Β έχει ήδη συμβεί. Θεωρείται απλοϊκός και αποδοτικός και χρησιμοποιείται συχνά επειδή δεν απαιτεί μεγάλο training set για ικανοποιητικά αποτελέσματα. Το θεώρημα Bayes ορίστηκε μαθηματικά ως η εξίσωση:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

όπου A και B είναι τα γεγονότα, $P(A)$ και $P(B)$ είναι οι πιθανότητες των A και B όπου είναι ανεξάρτητα μεταξύ τους, $P(A|B)$ είναι η πιθανότητα του A δεδομένου ότι το B είναι αληθής και $P(B|A)$ είναι η πιθανότητα του B δεδομένου ότι το A είναι αληθής.

Αφού γίνουν οι υπολογισμοί πιθανοτήτων για έναν αριθμό υποθέσεων, γίνεται επιλογή της υπόθεσης με την μεγαλύτερη πιθανότητα. Αυτή ονομάζεται maximum a posteriori (MAP) υπόθεση και μπορεί να αποτυπωθεί ως:

$$MAP(A) = max(P(A|B))$$

ή

$$MAP(A) = max(\frac{P(B|A)P(A)}{P(B)})$$

ή

$$MAP(A) = max(P(B|A) * P(A)))$$

Η $P(B)$ είναι ένας όρος κανονικοποίησης που επιτρέπει τον υπολογισμό των πιθανοτήτων. Μπορεί να αγνοηθεί όταν υπάρχει ενδιαφέρον για την πιο πιθανή υπόθεση, αφού παραμένει σταθερά. Στους αλγορίθμους κατηγοριοποίησης εάν ο αριθμός παρατηρήσεων είναι ίσος σε κάθε κλάση (εάν μιλάμε για πρόβλημα ανάλυσης συναισθήματος με τρεις βασικές κατηγορίες τότε: Χ θετικά = Χ αρνητικά = Χ ουδέτερα) τότε η πιθανότητα της κάθε κλάσης είναι ίση με την πιθανότητα των υπολοίπων κλάσεων. Οπότε μπορεί να αναφερθεί ότι:

$$MAP(A) = max(P(B|A))$$

Οι πιθανότητες της κάθε κλάσης (Class Probabilities) είναι το άθροισμα των παρατηρήσεων που ανήκουν σε κάθε κλάση (Κ) διαιρεμένο με το σύνολο των παρατηρήσεων σε όλες τις κλάσεις (Ν). Δηλαδή:

$$P(Class_A) = \frac{K}{N}$$

Σχήμα 3.2: Κατηγοριοποιητής Naive Bayes

Maximum Entropy

Στην επιστήμη των υπολογιστών η εντροπία χρησιμοποιείται ως μονάδα μέτρησης για την απροβλεψία των δεδομένων. Παραδείγματος αν ρίξεις ένα ζάρι έξι φορές και τα έξι αποτελέσματα έχουν την ίδια πιθανότητα 1/6, έχοντας μέγιστη εντροπία ίση με μονάδα. Η ιδέα πίσω από αυτόν τον αλγόριθμο Maximum Entropy είναι ότι θέλουμε έναν αλγόριθμο που είναι αμερόληπτος, δίνοντας την μέγιστη εντροπία. Αντίθετα από τον αλγόριθμο Naive Bayes δεν θεωρεί πως όλα τα χαρακτηριστικά των δεδομένων είναι ανεξάρτητα μεταξύ τους. Εάν c είναι η κλάση και d είναι η λέξη τότε θεωρείται ότι:

$$P_{ME}(c|d) = \frac{exp(\sum \lambda_i f_i(c,d))}{\sum exp(\sum \lambda_i f_i(c,d))}$$

Ένα διαφορετικό όνομα της τεχνικής αυτής είναι πολυεπίπεδη λογιστική παλινδρόμηση (Multinomial Logistic Regression/Softmax Regression). Πρόκειται για ένα μη γραμμικό μοντέλο που αποτελεί γενίκευση της απλής λογιστικής παλινδρόμησης (Logistic Regression) για την περίπτωση όπου η εξαρτημένη μεταβλητή Y παίρνει παραπάνω από δύο τιμές. Σε αυτή την περίπτωση η πιθανότητα μία δοσμένη παρατήρηση X να ανήκει στην κλάση Y είναι:

$$P(y^{(i)} = k|x^{(i)};\theta) = \frac{exp(\theta^{(k)T}x^{(i)})}{\sum_{j=1}^{k} exp(\theta^{(j)T}x^{(i)})}$$

Έτσι η υπόθεση παίρνει την μορφή:

$$h_\theta(x) = \begin{vmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ . \\ . \\ . \\ P(y=K|x;\theta) \end{vmatrix} = \frac{1}{\sum_{j=1}^{k} exp(\theta^{(j)T}x)} \begin{vmatrix} exp(\theta^{(1)T}x) \\ exp(\theta^{(2)T}x) \\ . \\ . \\ . \\ exp(\theta^{(k)T}x) \end{vmatrix}$$

Όπου εδώ τα $\theta^{(1)}$, $\theta^{(2)}$,..., $\theta^{(k)}$ ϵ $\Re^n$ είναι οι παράμετροι του μοντέλου και ο όρος $\frac{1}{\sum_{j=1}^{k} exp(\theta^{(j)T}x)}$ κανονικοποιεί την κατανομή έτσι ώστε να αθροίζεται σε μονάδα.
Έτσι επιλέγεται η κλάση με την μεγαλύτερη πιθανότητα για να κατηγοριοποιηθεί η νέα παρατήρηση. Η συνάρτηση κόστους θα είναι:

$$J(\theta) = -[\sum_{i=1}^{m}\sum_{k=1}^{k} 1(y^{(i)}=k) log \frac{exp(\theta^{(k)T}x^{(i)})}{\sum_{j=1}^{k} exp(\theta^{(j)T})x^{(i)})}]$$

Όπου 1() είναι μία συνάρτηση όπου εάν 1(Αληθής) = 1 και 1(Ψευδής) = 0. Επειδή δεν μπορούμε να ελαχιστοποιήσουμε την $J(\theta)$, θα γίνει χρήση του αλγορίθμου απότομης καθόδου (Gradient Descent):

$$\nabla\theta^{(k)} J(\theta) = -\sum_{i=1}^{m} [x^{(i)}(1(y^{(i)}=k) - P(y^{(i)}=k|x^{(i)};\theta))]$$

3.1.2 Μηχανική μάθηση χωρίς επίβλεψη

Η μηχανική μάθηση χωρίς επίβλεψη είναι άλλη μια κατηγορία μηχανικής μάθησης με σκοπό την κατηγοριοποίηση δεδομένων χωρίς να έχει προηγηθεί εκπαίδευση μέσω παραδειγμάτων. Γίνεται προσπάθεια εύρεσης μοτίβου συσχέτισης μεταξύ των δεδομένων, αυτή είναι και η κύρια διαφορά με την μηχανική μάθηση με επίβλεψη. Από την στιγμή που δεν υπάρχει ήδη κατηγοριοποιημένο σύνολο δεδομένων για εκπαίδευση του μοντέλου δεν δίνεται η δυνατότητα να εκτιμηθεί η ευστοχία του.

3.1.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση είναι η τρίτη μεγάλη κατηγορία του τομέα της μηχανικής μάθησης και διαφέρει σε πολλά σημεία από τις δύο παραπάνω. Βασίζεται στη μάθηση από το περιβάλλον αλληλεπίδρασης και χρησιμοποιείται ευρέως στον τομέα της ρομποτικής. Το μοντέλο κάνει προσπάθεια μεγιστοποίησης μια συνάρτησης επιβράβευσης χωρίς καθοδήγηση από εξωτερικό παράγοντα, προβλέποντας την επόμενη κίνηση του ανάλογα με το τι θα του αποφέρει μεγαλύτερο κέρδος.

3.2 Χρήση λεξικού

Άλλη μια τεχνική ανάλυσης συναισθήματος είναι αυτή με χρήση συναισθηματικού λεξικού. Με την τεχνική αυτή υπολογίζεται η πολικότητα ενός κειμένου με βάση το συναίσθημα των λέξεων που το απαρτίζουν [18]. Αυτό γίνεται δυνατό με τη χρήση ενός λεξικού που εμπεριέχει αντιστοιχίες λέξεων με την πολικότητα τους. Η τεχνική αυτή είναι αρκετά χρήσιμη όταν το κείμενο που γίνεται η ανάλυση είναι ένα κομμάτι λόγου από κοινωνικά δίκτυα, αφού η γλώσσα που χρησιμοποιείται σε αυτά δεν είναι επίσημη, με αποτέλεσμα να καθιστά δύσκολη την ανάλυση της. Υπάρχουν λεξικά που δημιουργήθηκαν συγκεκριμένα για κοινωνικά δίκτυα όπου απαρτίζονται από λέξεις της καθομιλουμένης και ιδιωματισμούς που διευκολύνουν σε μεγάλο βαθμό την συναισθηματική ανάλυση με ικανοποιητικά αποτελέσματα [6]. Ο συναισθηματικός βαθμός ενός κειμένου όπως αναφέρθηκε υπολογίζεται από το άθροισμα των συναισθηματικών βαθμών των επιμέρους λέξεων:

$$polarity = \sum_{j=1}^{K} score(x)$$

Ύστερα από τον υπολογισμό του συναισθήματος των επιμέρους λέξεων της πρότασης γίνεται κανονικοποίηση από το μήκος του κειμένου:

$$S = \sum_{i=1}^{N} \frac{polarity}{T}$$

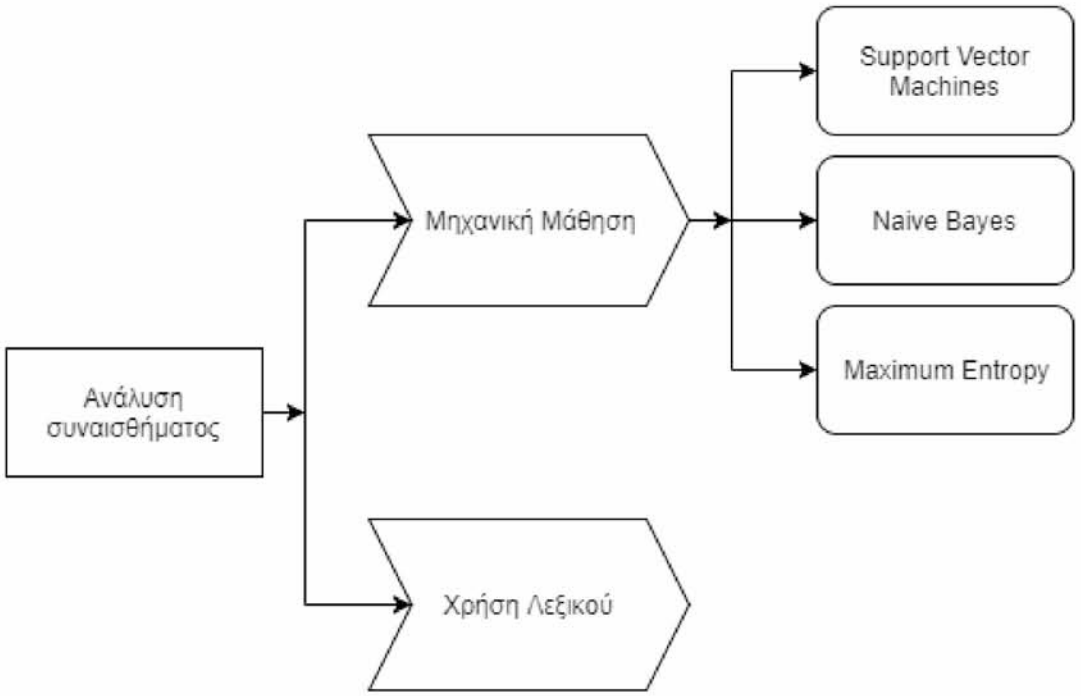
Όπως γίνεται κατανοητό η απόδοση του αλγορίθμου είναι άρρητα εξαρτημένη από το είδος του λεξικού που θα γίνει χρήση. Στα περισσότερα λεξικά γίνεται εκμετάλλευση

της άρνησης πριν από μία λέξη προσδίδοντας αρνητικό βαθμό. Ακόμα δίδονται και ανάλογα βάρη σε μέρη του λόγου όπως επιρρήματα, ρήματα και επίθετα αφού είναι σημαντικότερα για απόφανση του συναισθήματος της πρότασης που γίνεται η ανάλυση.

Η περισσότερο αποδοτική αλλά χρονοβόρα διαδικασία δημιουργίας συναισθηματικού λεξικού απαιτεί ανθρώπινο παράγοντα για την ανάλυση λέξεων και την κατηγοριοποίηση αυτών σύμφωνα με την κρίση του αναλυτή. Παράδειγμα τέτοιου λεξικού είναι των Bin Liu και Minqing Hu όπου το 2004 δημιούργησαν το δικό τους λεξικό αποτελούμενο από 6800 λέξεις της αγγλικής, κατηγοριοποιημένες σε θετικές, αρνητικές ή χωρίς συναίσθημα. Ακόμα ένα λεξικό είναι το The Multi Perspective Question Answering MPQA Opinion Corpus που δημιουργήθηκε από τους Theresa Wilson, Janyce Wiebe, Paul Hoffmann το 2005 που εμπεριέχει 8,222 κατηγοριοποιημένες λέξεις σύμφωνα με το συναίσθημα που δηλώνουν.Τέλος το πιο διαδεδομένο λεξικό είναι το SentiWordNet που δημιουργήθηκε αναλύοντας 155,287 λέξεις από την βάση λέξεων WordNet απο τους Esuli, Sebastiani το 2006.

Ένας άλλος τρόπος δημιουργίας συναισθηματικού λεξικού είναι αυτός που ανήκει στον τομέα της επιτηρούμενης μηχανικής μάθησης, εκπαιδεύοντας ένα μοντέλο ώστε να μπορεί να αποφανθεί για το συναίσθημα μίας λέξης. Στις περισσότερες περιπτώσεις λόγω του σφάλματος που υπάρχει στην εκπαίδευση μηχανικών μοντέλων η απόδοση είναι μικρότερη σε σχέση με τα λεξικά όπου επενέβη ανθρώπινη κρίση.

Σχήμα 3.3: Οι προσεγγίσεις της ανάλυσης συναισθήματος



3.3 Τεχνική Σάκος από Λέξεις

Το μοντέλο σάκος από λέξεις (Bag of Words) είναι μια αναπαράσταση λέξεων που χρησιμοποιείται ευρέως στην επεξεργασία φυσικής γλώσσας. Σε αυτή την αναπαράσταση όλες οι λέξεις ενός κειμένου αναπαρίστανται ως ένα σύνολο, αγνοώντας την γραμματική και την σειρά των λέξεων με την οποία εμφανίστηκαν. Το μοντέλο αυτό χρησιμοποιείται συχνά για κατηγοριοποίηση κειμένων όπου η συχνότητα εμφάνισης των λέξεων τροφοδοτεί έναν αλγόριθμο κατηγοριοποίησης.

Ένα από τα πιο διαδεδομένα παραδείγματα χρήσης αυτής της τεχνικής είναι αυτής του spam filtering σε emails. Στον πρώτο σάκο υπάρχουν λέξεις σχετικές με διαφημίσεις, συνδρομές και προσφορές, ενώ στον δεύτερο αποδεκτές λέξεις. Ύστερα για την κατηγοριοποίηση ενός email ο Naive Bayes αλγόριθμος θεωρεί ότι το email είναι ένα σύνολο λέξεων και χρησιμοποιεί τις Μπαγιεσιανές πιθανότητες για να αποφασίσει σε ποιόν από τους δύο σάκους είναι πιθανότερο να ανήκει.

Ένα τρόπος χρήσης αυτής της αναπαράστασης είναι με την συνάρτηση CountVectorizer της scikit-learn. Είναι απλοϊκή και βασίζεται στην καταμέτρηση των λέξεων μέσα στο σύνολο, δημιουργώντας έναν πίνακα όπου αποθηκεύει την συχνότητα της κάθε λέξης. Πιο αναλυτικά:

Έστω ένα σύνολο δεδομένων που περιλαμβάνει την παρακάτω πρόταση:

- *Η μπαταρία του κινητού Α διαρκεί περισσότερο από την μπαταρία του κινητού Β*

Μετά την διαδικασία μοντελοποίησης της πρότασης ως σάκος από λέξεις, προκύπτει ο ακόλουθος πίνακας διανυσμάτων:

Σχήμα 3.4: Σάκος από λέξεις

κινητού	2
Μπαταρία	2
του	2
Α	1
από	1
Β	1
διαρκεί	1
Η	1
περισσότερο	1
την	1

3.4 Τεχνική Συχνότητας Όρων

Η τεχνική συχνότητας όρων (Term Frequency and Inverse Document Frequency, TF-IDF) χρησιμοποιείται για την ανακάλυψη της σημαντικότητας μιας λέξης σε ένα κείμενο. Συχνά χρησιμοποιείται για να δώσει το ανάλογο βάρος σε μια λέξη για καλύτερο χειρισμό της. Παράδειγμα η αφαίρεση των πιο συχνά εμφανιζόμενων λέξεων σε μια πρόταση μπορεί να επιτευχθεί με αυτή την τεχνική απλοποιώντας κατά πολύ το πρόβλημα. Η χρήση της είναι παρόμοια με αυτή της τεχνικής σάκος από λέξεις έχοντας όμως σημαντικές διαφορές μεταξύ τους. Οι τιμές μπορεί να αυξάνονται αναλογικά με την συχνότητα εμφάνισης των λέξεων, όμως γίνεται κανονικοποίηση ανάλογης της συχνότητας της λέξης σε όλο το σύνολο του κειμένου. Αυτό βοηθάει στην καλύτερη αντιμετώπιση των λέξεων που εμφανίζονται συχνότερα από κάποιες άλλες. Η χρήση της είναι εύκολη αφού παρέχεται από την scikit-learn ως TfidfVectorizer.

A(T) = (Ο αριθμός εμφάνισης του όρου T) / (Συνολικός αριθμός όρων στο κείμενο).

B(T) = log_e(Συνολικός αριθμός κειμένων / Αριθμός κειμένων με τον όρο T μέσα).

Αξίζει να σημειωθεί ότι το 83% του συνόλου των αλγόριθμων που σχετίζονται με ανάλυση κειμένων, χρησιμοποιούν αυτή την τεχνική.

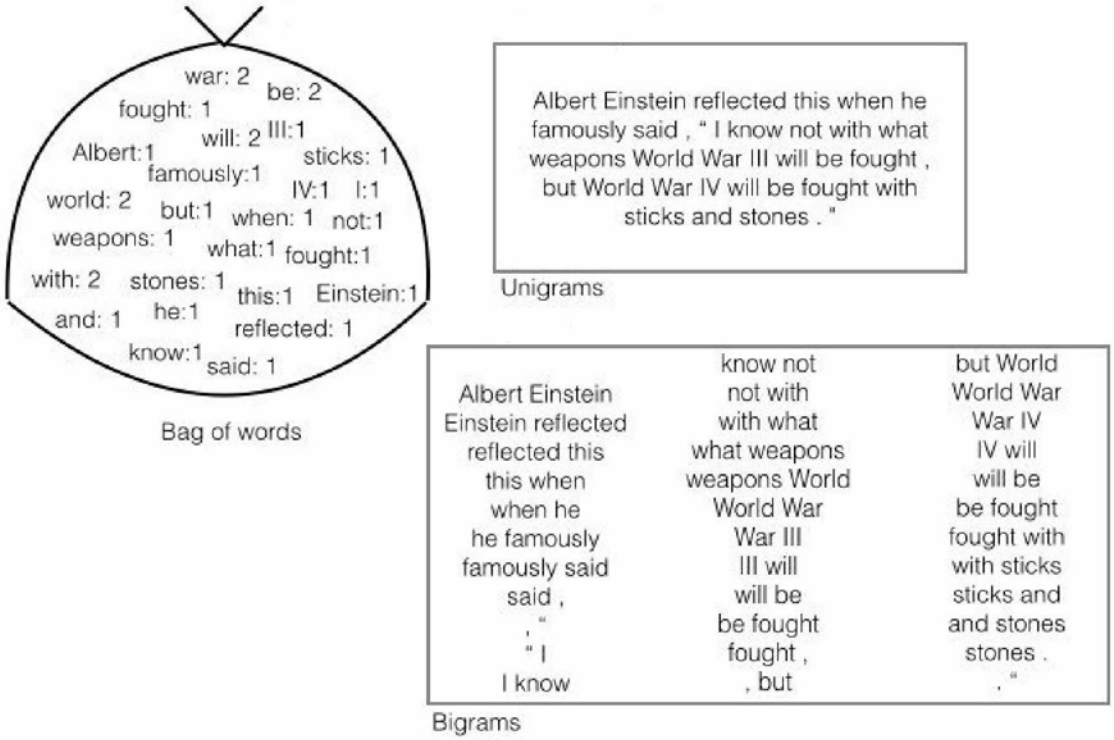
3.5 N-grams

Σε αντίθεση με την τεχνική του σάκου από λέξεις όπου η σειρά των λέξεων σε ένα κείμενο αγνοείται, υπάρχει και η τεχνική των N-λέξεων. Η πρώτη τεχνική υστερεί στο γεγονός ότι δεν λαμβάνει υπόψη την γραμματική μιας γλώσσας, δηλαδή στις περισσότερες περιπτώσεις ένα όνομα ακολουθείται από ένα ρήμα. Με την δεύτερη τεχνική μπορούμε να χρησιμοποιήσουμε ομάδες λέξεων και να παράγουμε καλύτερα αποτελέσματα. Στην περίπτωση της 1-λέξης (unigram) η διαφορά με την τεχνική του σάκου είναι ότι δεν χάνεται η σειρά των λέξεων μέσα στο σύνολο. Το αποτέλεσμα της διαδικασίας της ανάλυσης σε λεκτικές μονάδες που αναφέρθηκε προηγουμένως παράγει unigrams, δηλαδή μεμονωμένες λέξεις και σύμβολα. Εάν συμβολίζουμε X τον αριθμό των λέξεων σε ένα κείμενο θα υπάρχουν:

$$N_{grams} = X - (N - 1)$$

διαφορικές περιπτώσεις από N-λέξεις οι οποίες σαν ομάδες αποδίδουν φυσικά καλύτερα από ότι αν ήταν μεμονωμένες.

Σχήμα 3.5: Μέθοδος N-λέξεων και σάκος από λέξεις



3.6 Αξιολόγηση Μοντέλων

Η αξιολόγηση των μοντέλων επιβλεπόμενης μηχανικής μάθησης είναι αναγκαία εάν θέλουμε να έχουμε μία εικόνα για το πως αντιμετωπίζει ο αλγόριθμος τα δεδομένα. Η αξιολόγηση αυτή γίνεται μέσω των δεδομένων εκμάθησης, όπου είναι ήδη κατηγοριοποιημένα σε κλάσεις. Δύο τρόποι αξιολόγησης είναι το Precision και Recall και ο Confusion Matrix.

3.6.1 Precision και Recall

Πέρα από τη μέτρηση της ακρίβειας ενός κατηγοριοποιητή μετρώντας το ποσοστό των σωστών προβλέψεων προς το σύνολο τους, υπάρχουν και άλλα εργαλεία που δίνουν πολύ σημαντικά στοιχεία. Αυτά είναι το Precision, το Recall, το F-score και το Support.

Στην ουσία τα Precision και Recall χρησιμοποιούνται για την αξιολόγηση ενός κατηγοριοποιητή. Βασίζονται στην μηδενική υπόθεση (Null Hypothesis), δηλαδή σε μία

υπόθεση που ισχύει κατά κανόνα, αλλά διατυπώνεται με σκοπό να αμφισβητηθεί. Αν δεν απορριφθεί τότε θεωρείται ότι τα δεδομένα δεν επαρκούν για την απόρριψη της, ενώ αν απορριφθεί τότε συμπεραίνεται ότι τα δεδομένα δεν επαληθεύουν την μηδενική υπόθεση, αλλά είναι συμβατά με μία άλλη την εναλλακτική υπόθεση.

Σε περίπτωση λάθους απόρριψης της μηδενικής υπόθεσης έχουμε σφάλμα τύπου I (False Positive), ενώ στην περίπτωση λάθους μη απόρριψης της μηδενικής υπόθεσης έχουμε σφάλμα τύπου II (False Negative).

Τώρα το Precision είναι η ικανότητα του κατηγοριοποιητή να μη χαρακτηρίσει θετική μια παρατήρηση που είναι αρνητική. Μαθηματικά διατυπώνεται ως:

$$Precision = \frac{truepositive}{truepositive + falsenegative}$$

δηλαδή ο αριθμός των θετικών προβλέψεων δια τον συνολικό αριθμό των προβλέψεων που θεωρήθηκαν ως θετικά.

Το Recall είναι ένα άλλο μέτρο αξιολόγησης, είναι η ικανότητα του κατηγοριοποιητή να βρει όλες τις θετικές παρατηρήσεις. Δηλαδή:

$$Recall = \frac{truepositive}{truepositive + falsenegative}$$

δηλαδή ο αριθμός των θετικών προβλέψεων δια τον αριθμό του συνόλου των δεδομένων που ήταν πραγματικά θετικά.

F-score ονομάζεται το ζυγισμένο μέσο του Precision και του Recall και κυμαίνεται από 0 έως 1. Συγκεκριμένα:

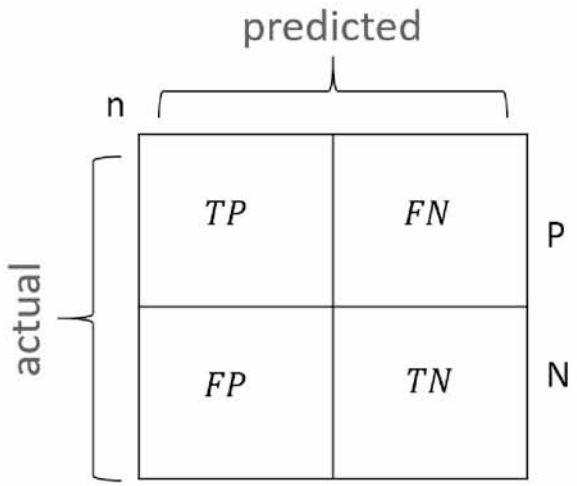
$$F - score = 2 * \frac{precision * recall}{precision + recall}$$

3.6.2 Confusion Matrix

Ο πίνακας λαθών (Confusion Matrix) είναι ένας πίνακας που απεικονίζει την απόδοση ενός κατηγοριοποιητή. Σαν ορισμός διατυπώνεται ότι ένας πίνακας λαθών C είναι κάθε $C_{i,j}$ όπου είναι ίσο με τον αριθμό των παρατηρήσεων που ανήκουν στην κλάση i αλλά προβλέφθηκαν στην κλάση j. Συγκεκριμένα κάθε γραμμή αναπαριστά τις παρατηρήσεις στην κλάση που πραγματικά ανήκουν, ενώ κάθε στήλη τις παρατηρήσεις στην κλάση που έγινε πρόβλεψη.

Στην δυαδική κατηγοριοποίηση ο αριθμός των true negatives είναι $C_{0,0}$, ο αριθμός των false negatives είναι $C_{1,0}$, των true positives είναι $C_{1,1}$ και των false positives είναι $C_{0,1}$.

Σχήμα 3.6: Πίνακας λαθών



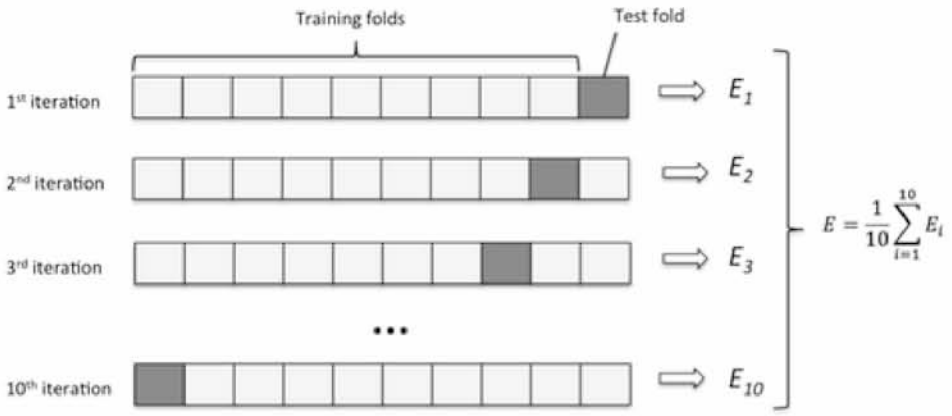
Είναι ένα μέτρο αξιολόγησης που μας επιτρέπει να εντοπίσουμε τις κλάσεις στις οποίες ο αλγόριθμός μας δεν αποδίδει και τόσο καλά, αφού μας δίνονται πληροφορίες για κάθε μια ξεχωριστά. Σε ένα παράδειγμα ανάλυσης συναισθήματος με κλάσεις θετική, αρνητική ή χωρίς συναίσθημα, ο πίνακας θα είναι διαστάσεων τρία επί τρία. Έτσι αντλώντας πληροφορίες από αυτόν μπορούμε να εντοπίσουμε τα σημεία στα οποία υπάρχει το πρόβλημα και να προβούμε σε αντίστοιχες ενέργειες.

3.7 Cross Validation

Τροφοδοτώντας έναν αλγόριθμο με δεδομένα για εκπαίδευση και ύστερα κάνοντας έλεγχο με τα ίδια δεδομένα είναι ένα σημαντικό λάθος που οδηγεί σε overfitting. Το μοντέλο έχει ήδη εκπαιδευτεί από αυτά και θα εμφανίσει τέλειο σκορ, αλλά δεν θα είναι σε θέση να κατηγοριοποιήσει σωστά νέα, άγνωστα προς αυτό δεδομένα. Για να αποφευχθεί αυτό μια συνήθης τεχνική είναι να χωρίσουμε τα δεδομένα σε δύο σετ, ένα για

εκπαίδευση και ένα για έλεγχο (Holdout Method). Με αυτόν τον τρόπο το πρόβλημα αντιμετωπίζεται, όμως μειώνοντας το σύνολο των δεδομένων προς εκπαίδευση αυξάνεται το σφάλμα. Μια άλλη τεχνική ονομάζεται K-Fold Cross Validation. Τα δεδομένα χωρίζονται σε K-σετ και η διαδικασία επαναλαμβάνεται K φορές, κάθε φορά ένα από τα K σετ χρησιμοποιείται για τεστ και τα υπόλοιπα K-1 για εκπαίδευση. Σε κάθε επανάληψη υπολογίζεται η ακρίβεια του αλγορίθμου, δηλαδή το ποσοστό των σωστών προβλέψεων ως προς το σύνολο αυτών. Η τελική απόδοση υπολογίζεται από τον μέσο όρο όλων των επαναλήψεων. Η σωστή επιλογή τιμής για το K είναι πολύ σημαντική. Μεγάλες τιμές συνεπάγονται σε μεγάλη ζήτηση για υπολογιστική ισχύ, ενώ μικρές τιμές αποφέρουν ένα μη αμερόληπτο αποτέλεσμα. Τέλος υπάρχει και η τεχνική Leave P out, LPO. Όπως μαρτυράει το όνομα της, αφήνεται ένα σύνολο από P παρατηρήσεις εκτός, ενώ γίνεται εκπαίδευση με όλες τις υπόλοιπες. Στην περίπτωση όπου $P = 1$ ονομάζεται Leave One Out,LOO όπου αφήνεται μία μοναδική παρατήρηση εκτός.

Σχήμα 3.7: 10-fold Cross Validation



Κεφάλαιο 4

Υλοποίηση

Στο παρόν κεφάλαιο γίνεται υλοποίηση ενός συστήματος όπου αναλύοντας δεδομένα από το Twitter παρέχονται πληροφορίες για συγκεκριμένα προϊόντα και υπηρεσίες, με σκοπό τον καλύτερο μελλοντικό επιχειρησιακό προγραμματισμό της εκάστοτε εταιρείας. Οι κατηγοριοποιητές που χρησιμοποιούνται βασίζονται σε λεξικά αλλά και σε επιβλεπόμενη μηχανική μάθηση, για σύγκριση αποτελεσμάτων μεταξύ τους.

Οι βασικές υπηρεσίες που παρέχει το παρακάτω σύστημα είναι η εξόρυξη tweets που σχετίζονται με το MacBook της Apple και η κατηγοριοποίηση αυτών ανάλογα με το συναίσθημα του εκάστοτε χρήστη. Ύστερα γίνεται αναζήτηση των θεμάτων όπου οι χρήστες είναι αρνητικοί πάνω σε αυτό το προϊόν, δίνοντας την δυνατότητα στην εταιρεία να τα διορθώσει στα επόμενα μοντέλα. Επίσης έγινε ανάλυση διαφόρων επικεφαλίδων από άρθρα της εφημερίδας The Huffington Post σε βάθος χρόνου [15], και ελέγχθηκε ο συσχετισμός του είδους του συναισθήματος κάθε άρθρου σχετικά με την εταιρεία Apple με την τιμή της μετοχής της (AAPL) την ίδια χρονική περίοδο.

4.1 Συλλογή δεδομένων από το Twitter

Η επικοινωνία με το Twitter για την εξόρυξη των tweets έγινε μέσω του API που αναλύθηκε στο Κεφάλαιο 2.1 και παρέχεται από το ίδιο το Twitter. Υπάρχουν δύο API's, το Stream API και το Rest API. Η διαφορά τους είναι ότι στο πρώτο ανοίγει μία σύνδεση και αντλούνται δεδομένα σε πραγματικό χρόνο, έως ότου ο χρήστης να αποφασίσει να το σταματήσει, ενώ στο δεύτερο μπορεί να γίνει επιλογή συγκεκριμένου χρονικού διαστήματος επικοινωνίας έχοντας όμως έναν περιορισμό των 150 tweets κάθε ώρα. Ένας άλλος περιορισμός που υφίσταται είναι η μη διαθεσιμότητα tweets που είναι παλαιότερα της μίας εβδομάδας.

Για την εξόρυξη αυτή χρησιμοποιήθηκε το Tweepy, μία εύκολη στην χρήση βιβλιοθήκη

της Python για πρόσβαση στις υπηρεσίες του Twitter. Επιπρόσθετα, για την συλλογή των δεδομένων χρειάστηκε η δημιουργία ενός λογαριασμού στο Twitter, αποκτώντας έτσι τα απαραίτητα κλειδιά που αποτελούν τα διαπιστευτήρια για την πρόσβαση στις διεπαφές του Twitter, τα οποία είναι το Consumer Key, το Consumer Secret, το Access Token και το Access Token Secret. Μέσω του Tweepy έγινε αναζήτηση για tweets που σχετίζονται με το MacBook της Apple [27]. Επίσης απο αυτά τα tweet αποθηκεύτηκαν μόνο αυτά που αποτελούνταν εξ΄ ολοκλήρου από αγγλικούς και μόνο χαρακτήρες και παράλληλα αντλήθηκε η ημερομηνία δημοσίευσης αλλά και η τοποθεσία του εκάστοτε χρήστη εάν αυτό ήταν δυνατό. Στο σύνολο συλλέχθηκαν 3055 tweets, από 09-10-2017 έως 30-10-2017, όλα σχετικά με το MacBook.

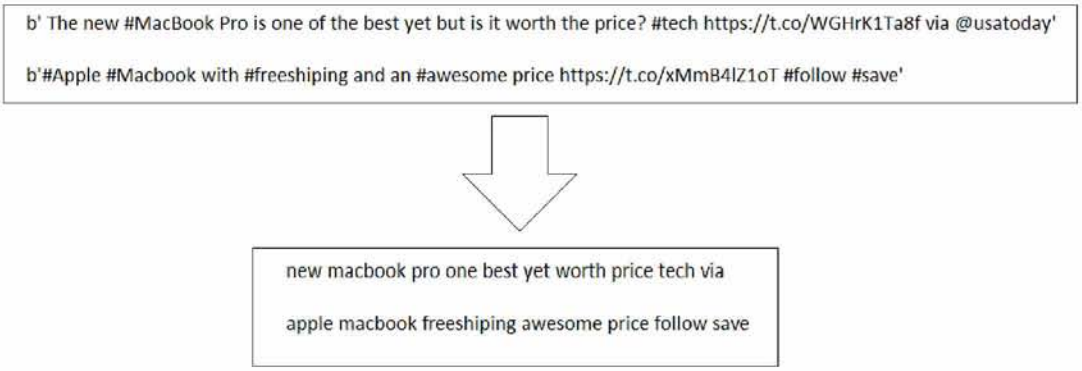
4.2 Φιλτράρισμα των tweets

Τα tweets που συλλέχθηκαν, είναι μία μίξη από urls και από άλλα δεδομένα που δεν έχουν κάποιο συναίσθημα όπως #hashtags, @annotations και 'RT' retweets. Αυτά θα πρέπει να αφαιρεθούν ώστε να μην επιβαρύνουν το σύστημα, αφού δεν έχουν να προσφέρουν καμία πληροφορία στην όλη ανάλυση [25]. Η διαδικασία φιλτραρίσματος των tweets που πραγματοποιήθηκε είναι η παρακάτω:

1. Αφαίρεση των Non-ASCII χαρακτήρων
2. Αφαίρεση των επισημάνσεων (@)
3. Αφαίρεση των URL's (http://)
4. Αφαίρεση των αριθμητικών χαρακτήρων
5. Αφαίρεση συμβόλων (=,%,&)
6. Μετατροπή σε πεζά γράμματα
7. Αφαίρεση των stopwords

Όπως αναλύθηκε στο κεφάλαιο 2.5, είναι ζωτικής σημασίας η μορφή των δεδομένων που θα τροφοδοτηθούν στον κατηγοριοποιητή, γι΄ αυτό δόθηκε μεγάλη έμφαση στο κομμάτι αυτό. Στο παρόν σύστημα όλα τα παραπάνω έγιναν με χρήση της βιβλιοθήκης *re* - Regular Expressions της Python, όπου με την χρήση κανονικών εκφράσεων δίνεται η δυνατότητα παραμετροποίησης συμβολοσειρών. Έτσι τα tweets είναι έτοιμα να δοθούν σαν είσοδος σε αλγορίθμους κατηγοριοποίησης και να ταξινομηθούν σε κατηγορίες.

Σχήμα 4.1: Παράδειγμα χρήσης κανονικών εκφράσεων για παραμετροποίηση των tweets



4.3 Προσέγγιση με χρήση λεξικού - Vader Lexicon

Το εργαλείο ανάλυσης φυσικής γλώσσας VADER Sentiment (Valence Aware Dictionary and sEntiment Reasoner) είναι λεξικό και ταυτόχρονα αναλυτής συναισθήματος για δεδομένα κειμένου. Χρησιμοποιείται περισσότερο για ανάλυση σε κοινωνικά δίκτυα αλλά δουλεύει ικανοποιητικά και σε άλλες μορφές κειμένων. Έχει εκτιμηθεί ότι η απόδοση του (F score, αναλύθηκε στο κεφάλαιο 3.7) είναι περίπου ίση με έναν πραγματικό κριτή, vader: F score = 0.96, human rater: F score = 0.84 [21], στην αντικειμενική κρίση του συναισθήματος για tweets και ταινίες.

Σχήμα 4.2: Παράδειγμα του λεξικού Vader

Word	Sentiment rating
tragedy	-3.4
rejoiced	2.0
insane	-1.7
disaster	-3.1
great	3.1

Όπως αναφέρθηκε στο κεφάλαιο 3.2, για να είναι ακριβής τα αποτελέσματα σε ένα λεξικό, προηγείται εκτίμηση του συναισθήματος των λέξεων από ανθρώπινο παράγοντα. Το συγκεκριμένο λεξικό αναλύει ένα κομμάτι λόγου και αντιστοιχεί τις λέξεις που το απαρτίζουν με λέξεις στο λεξικό. Το Vader παράγει τρεις βαθμίδες συναισθηματικής μέτρησης, θετικό, ουδέτερο ή αρνητικό. Συγκεκριμένα η βαθμωτή κλίμακα κυμαίνεται από το -4 (εξαιρετικά αρνητικό) έως το 4 (εξαιρετικά θετικό). Εκτός από την συ-

ναισθηματική πολικότητα των προτάσεων το λεξικό αυτό παράγει και μία παραπάνω πληροφορία, του κανονικοποιημένου συναισθήματος (compound) όπου ουσιαστικά είναι μια αθροιστική κανονικοποίηση, με εύρος από -1 έως 1.

Σχήμα 4.3: Συναισθηματικές βαθμίδες

Sentiment metric	Value
Positive	0.45
Neutral	0.55
Negative	0.00
Compound	0.69

Το γεγονός ότι τα λεξικά χρειάζονται ανθρώπινο παράγοντα για να έχουν μεγάλη ακρίβεια τα καθιστά χρονοβόρα με αποτέλεσμα να μην είναι πάντα ενημερωμένα με νέες λέξεις της καθομιλουμένης. Αξίζει να αναφερθεί ότι το συγκεκριμένο λεξικό περιλαμβάνει μία μεγάλη γκάμα λέξεων της αγγλικής καθημερινής γλώσσας, αλλά και συντομογραφίες και emoticons. Τα αποτελέσματα αυτού του αλγορίθμου με είσοδο τα tweets σχετικά με το MacBook ήταν 611 θετικά, 237 αρνητικά και 2207 ουδέτερα.

4.4 Μέθοδος μηχανικής μάθησης - Naive Bayes

Όπως αναφέρθηκε στο κεφάλαιο 3.1, ο συγκεκριμένος αλγόριθμος βασίζεται στην θεωρία του Bayes και χρησιμοποιεί πιθανοτικούς όρους για να λύσει το πρόβλημα της κατηγοριοποίησης. Πριν γίνει τροφοδότηση των tweets στο μοντέλο κατηγοριοποίησης απαιτείται μια διαδικασία. Αρχικά απαιτείται η διάσπαση των λέξεων μέσα στην πρόταση (tokenization). Στη συνέχεια αυτές οι λέξεις θα πάρουν τη μορφή διανυσμάτων προκαθορισμένου μεγέθους πριν δοθούν ως είσοδος σε έναν αλγόριθμο μηχανικής μάθησης. Η διαδικασία μετατροπής που χρησιμοποιείται είναι η τεχνική σάκος από λέξεις (Κεφάλαιο 3.3). Η βιβλιοθήκη scikit learn παρέχει εργαλεία που καθιστούν αυτήν την διαδικασία ευκολότερη.

Ο τρόπος διανυσματοποίησης που χρησιμοποιήθηκε είναι μέσω της CountVectorizer μίας συνάρτησης της scikit learn, αρχικά για διάσπαση των λέξεων όλων των tweets και ύστερα για την δημιουργία ενός λεξικού γνωστών λέξεων ώστε να μπορεί να γίνει ανάλυση νέων και άγνωστων. Χρησιμοποιήθηκε ως εξής:

1. Δημιουργία της κλάσης CountVectorizer
2. Κάλεσμα της συνάρτησης fit() ώστε να γίνει εκμάθηση ενός λεξικού

3. Χρήση της transform() σε ένα σύνολο λέξεων ώστε να γίνει κωδικοποίηση αυτών ως διάνυσμα αριθμών.

Τα κωδικοποιημένα διανύσματα επιστρέφονται με το μέγεθος του λεξικού που δημιουργήθηκε και έναν ακέραιο αριθμό για την συχνότητα των εμφανίσεων κάθε λέξης μέσα στο σύνολο. Η παραπάνω διαδικασία θα γίνει αρχικά για το corpus με το οποίο εκπαιδεύτηκε το μοντέλο και ύστερα για τα νέα tweets που πρέπει να κατηγοριοποιηθούν. Το corpus που χρησιμοποιήθηκε αποτελείται από ένα σύνολο 11517 χειροκίνητα κατηγοριοποιημένων tweets σύμφωνα με το συναίσθημα που δηλώνουν. Ονομάζεται Sentiment140 και είναι πολύ δημοφιλές σε αναλύσεις κοινωνικών δικτύων.

Σχήμα 4.4: Sentiment140

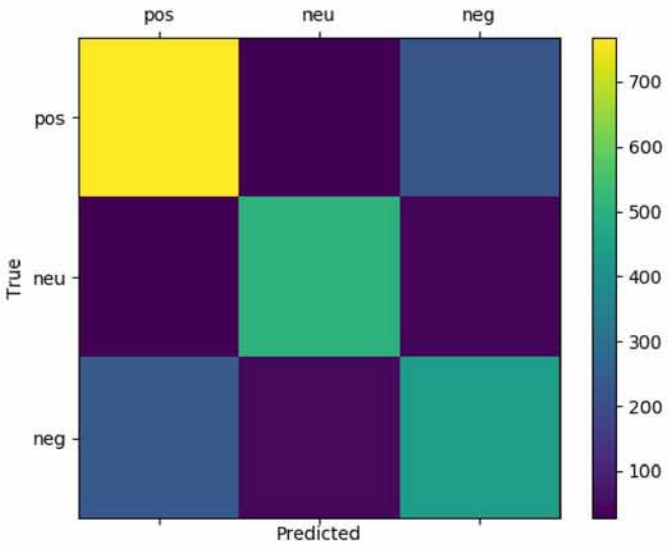
is so sad for my APL friend...	neg
Sunny Again, work tomorrow :-! TV tonight	neg
goodbye exams. HELLO ALCOHOL TONIGHT	pos
No Sat off... Need to work 6 days a week	neg
thrilled about being at work this morning	pos
Goodmorning #twitter!	neu
Shout out to all my followers! #twitter	neu

Τα δεδομένα είναι έτοιμα για είσοδο σε αλγόριθμο κατηγοριοποίησης και έτσι καλείται ο Multinomial Naive Bayes κατηγοριοποιητής από την βιβλιοθήκη scikit learn με εκ των προτέρων πιθανότητες 0.31 για την θετική κλάση, 0.25 για την ουδέτερη κλάση και 0.44 για την αρνητική κλάση. Οι πιθανότητες αυτές εξαρτώνται από τον αριθμό των παρατηρήσεων σε κάθε κλάση που έχουμε στο αρχικό corpus εκμάθησης και υπολογίζονται ακριβώς από:

$$class\ prior = \frac{Number\ of\ samples\ in\ the\ class}{Total\ number\ of\ samples}$$

Το corpus που χρησιμοποιήθηκε για εκμάθηση (Sentiment140) διασπάστηκε σε 20% για τεστ και το 80% για εκπαίδευση του μοντέλου. Φυσικά αυτό είναι ένα τίμημα που παίρνει όποιος θέλει να ελέγξει την απόδοση του μοντέλου του, γιατί εκπαιδεύοντας μόνο με το 80% αντί με όλο το corpus προφανώς θα είναι μειωμένη η ευστοχία του. Έτσι ο αλγόριθμος εκπαιδεύτηκε σε 9212 παρατηρήσεις και έγινε τεστ στις υπόλοιπες 2304, με το αποτέλεσμα ευστοχίας να κυμαίνεται στο 75%. Όπως σημειώθηκε και στα κεφάλαια 3.7 και 3.8, η ευστοχία δεν είναι ο μόνος τρόπος αξιολόγησης ενός μοντέλου κατηγοριοποίησης.

Σχήμα 4.5: Πίνακας λαθών του μοντέλου



$$\begin{bmatrix} 759 & 23 & 250 \\ 30 & 500 & 45 \\ 218 & 29 & 450 \end{bmatrix}$$

Ο πίνακας λαθών του μοντέλου όπως φαίνεται δείχνει να έχει πολύ καλά αποτελέσματα στην αντιμετώπιση των θετικών tweets αλλά λιγότερα καλά στα αρνητικά. Στην διαγώνιο του πίνακα φαίνονται οι σωστές προβλέψεις και συγκεκριμένα κατηγοριοποιήθηκαν σωστά 759 θετικά tweet, 500 ουδέτερα και 450 αρνητικά. Τα περισσότερα λάθη έγιναν στην πρόβλεψη θετικών tweets ως αρνητικά και αρνητικών ως θετικά.

Το Precision και το Recall του συγκεκριμένου μοντέλου για κάθε κλάση ξεχωριστά είναι αντίστοιχα:

Precision: $\begin{bmatrix} 0.75372393 & 0.9057971 & 0.60402685 \end{bmatrix}$

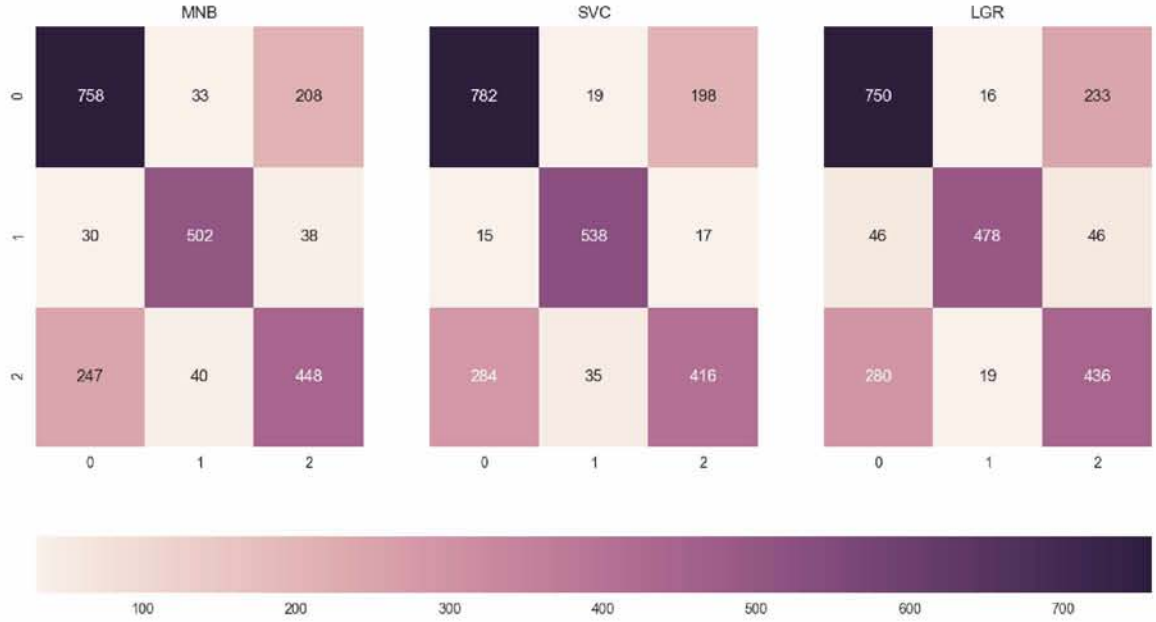
Recall: $\begin{bmatrix} 0.73546512 & 0.86956522 & 0.6456241 \end{bmatrix}$

Fscore: $\begin{bmatrix} 0.74448259 & 0.88731145 & 0.62413315 \end{bmatrix}$

Η παραπάνω διαδικασία ακολουθήθηκε για την εκπαίδευση και ακόμα δύο μοντέλων μηχανικής μάθησης, αυτά των μηχανών διανυσματικής υποστήριξης και λογιστικής

παλινδρόμησης. Και οι δύο αυτοί αλγόριθμοι αναλύθηκαν περαιτέρω στο Κεφάλαιο 3.

Σχήμα 4.6: Σύγκριση πινάκων λαθών



Τα Fscores των δύο νέων αλγόριθμων για κάθε κλάση ξεχωριστά είναι αντίστοιχα:

Fscore SVC: $\begin{bmatrix} 0.75192308 & 0.92598967 & 0.6090776 \end{bmatrix}$

Fscore LGR: $\begin{bmatrix} 0.72289157 & 0.88273315 & 0.60137931 \end{bmatrix}$

Όπως φαίνεται ο κατηγοριοποιητής βασισμένος στις μηχανές διανυσματικής υποστήριξης ίσως δουλεύει λίγο καλύτερα όσον αναφορά τα θετικά και τα ουδέτερα tweets αλλά όχι στα αρνητικά. Στην παρούσα εργασία επειδή θα γίνει ανάλυση των αρνητικών tweets μας ενδιαφέρει να υπάρχει καλή αντιμετώπιση αυτών, οπότε στην υλοποίηση παρακάτω επιλέχθηκε ο κατηγοριοποιητής Naive Bayes ως βασικός.

Γίνεται αντιληπτό ότι γενικά οι κατηγοριοποιητές τα καταφέρνουν ικανοποιητικά όσον αναφορά θετικά και ουδέτερα σε συναίσθημα tweets, όμως δεν ισχύει το ίδιο για τα αρνητικά. Μπορούμε να υποψιαστούμε ότι δεν μπορούν να διακρίνουν εύκολα λέξεις που δηλώνουν άρνηση και να αντιστρέψουν την πολικότητα της πρότασης, με αποτέλεσμα να την κατηγοριοποιούν ως θετική. Τώρα με χρήση του Naive Bayes μοντέλου αφού έγινε εκπαίδευση και αξιολόγηση της απόδοσής του, ήρθε η ώρα να δοκιμαστεί σε νέα, άγνωστα προς αυτό δεδομένα. Αφού λοιπόν τα tweets πέρασαν από διαδικασία φιλτραρίσματος και έγιναν διανύσματα από αριθμούς όπως αναφέρθηκε παραπάνω,

μεταφέρθηκαν ως είσοδος στο εκπαιδευμένο μοντέλο. Τα αποτελέσματα είναι 772 θετικά tweets, 805 αρνητικά και τα υπόλοιπα 1478 δεν παρουσίαζαν κάποιο συναίσθημα. Κάποια παραδείγματα κατηγοριοποιημένων tweets είναι:

Ως θετικό tweet: → [love macbook awesome]

Όπως είναι λογικό η συγκεκριμένη πρόταση περιέχει δύο λέξεις που δηλώνουν θετικό συναίσθημα σε μεγάλο βαθμό, οπότε γι' αυτό και κατηγοριοποιήθηκε ως θετικό.

Ως αρνητικό tweet: → [really need new macbook damn prices]

Λογικά η χρήση της λέξης 'damn' έδωσε αρνητική σημασία και άλλαξε την πολικότητα αυτής της πρότασης.

Ως ουδέτερο tweet: → [remove virus apple macos news]

Ένα tweet που δεν παρουσιάζει κάποιο συναίσθημα και μοιάζει περισσότερο σε μία διαφήμιση.

Φυσικά όπως φαίνεται παραπάνω στην αξιολόγηση του κατηγοριοποιητή, δεν είναι τέλειος. Υπάρχουν λανθασμένα κατηγοριοποιημένα tweets τα οποία έπρεπε να ανήκουν σε διαφορετικές κλάσεις, όπως:

× [bought new macbook days half keys stop working]

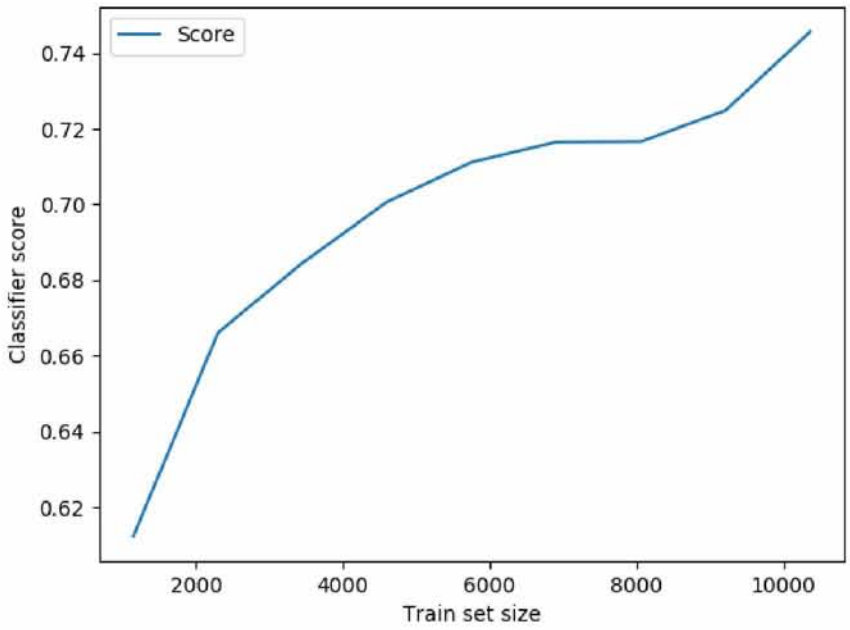
το οποίο κατηγοριοποιήθηκε ως θετικό, ενώ στην πραγματικότητα μάλλον κρύβει μια δυσαρέσκεια ως προς το πληκτρολόγιο του MacBook.

Το μοντέλο δοκιμάστηκε με διαφόρου μεγέθους training sets, ξεκινώντας από το 90% του συνολικού corpus μέχρι το 10%. Τα αποτελέσματα φαίνονται στην παρακάτω εικόνα. Είναι ξεκάθαρο ότι όσο μεγαλύτερο είναι το training set που τροφοδοτείται σε ένα μοντέλο για εκπαίδευση, τόσο περισσότερα θα μάθει και θα είναι σε θέση να αναγνωρίσει και να κατηγοριοποιήσει καλύτερα άγνωστα δεδομένα.

4.5 Αποτελέσματα ανάλυσης

Ένα μικρό ποσοστό του συνόλου των δεδομένων, είναι αρνητικό. Αυτό όμως από μόνο του δεν μπορεί να δώσει παραπάνω πληροφορίες. Έτσι με την χρήση της βιβλιοθήκης που αναφέρθηκε στο κεφάλαιο 2.4, Natural Language Toolkit, έγινε προσπάθεια εύρεσης των ουσιαστικών μέσα στο σύνολο των αρνητικών tweets και ύστερα αποθήκευση

Σχήμα 4.7: Απόδοση μοντέλου σε σχέση με το ποσοστό του corpus που χρησιμοποιήθηκε για εκπαίδευση



μόνο των πιο συχνά εμφανιζόμενων. Πέρα όμως από τα ουσιαστικά άξιζε η αποθήκευση και άλλων μερών του λόγου, που βοηθούσαν αρκετά στην ανάλυση [23].

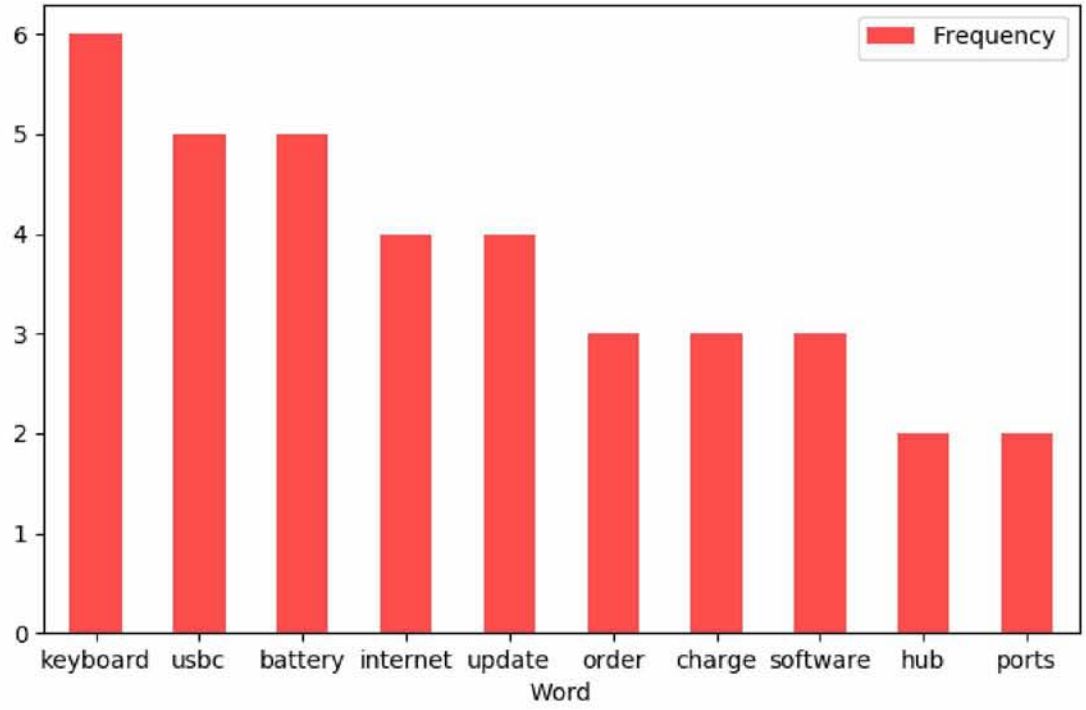
Όπως φαίνεται παρακάτω στο σχήμα 4.8, οι περισσότεροι χρήστες του MacBook αντιμετωπίζουν προβλήματα με το πληκτρολόγιο, ενώ ακολουθούν το usb-c η μπαταρία, η συνδεσιμότητα με το internet, τα updates της Apple, προβλήματα με τις παραγγελίες, την φόρτιση, το λογισμικό και τέλος με τις θύρες.

Αυτές οι πληροφορίες είναι χρήσιμες σε μία εταιρεία η οποία με βάση αυτών θα προσπαθήσει να διορθώσει τα τυχόν προβλήματα σε νέες εκδόσεις των προϊόντων της. Φυσικά θα βοηθούσε αρκετά η εύρεση των tweets που αναφέρουν αυτά τα προβλήματα για καλύτερη κατανόηση του προβλήματος, οπότε αναζητώντας σε όλα τα tweets με αρνητικό συναίσθημα λέξεις κλειδιά όπως πληκτρολόγιο, usb-c, μπαταρία και τα λοιπά εμφανίζονται σημαντικές πληροφορίες. Μερικά παραδείγματα αρνητικών tweets με θέμα το πληκτρολόγιο του MacBook είναι:

1.In the midst of these #MacBook keyboard complaints, I will say that I got a @surface book 2 weeks ago and have never liked a computer more

6.@ZaBlanc ... I already own a #macbook (a year old) and the keyboard is something I don't like at all ... https://t.co/9S8c2sUqKI

Σχήμα 4.8: Τα συχνότερα παράπονα των καταναλωτών



13.Bought new #MacBook online after all. Was kicked out of living room since keyboard makes annoying sounds while typing #EmbarassingQuality

65.#macbook #touchbar should come with tactile feedback and replace the entire keyboard. Wait, isn't that a double folded #ipad?

Με θέμα την μπαταρία:

29.My #Macbook pro (early 2013) has reached battery cycle count 1003. Seeing battery issue. At 35% laptop auto powers-off. Is it replaceable?

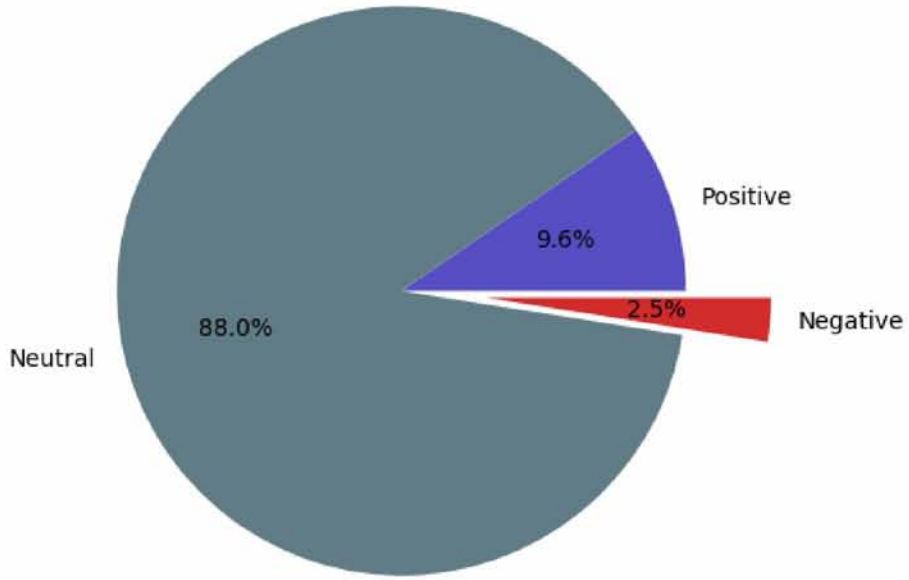
70.#iphone not the only #apple product with battery problems #macbook #battery #fails rapidly their answer "we don't care, we don't have to"

Με θέμα την παραγγελία:

Ordered #MacBook from @amazon they cancelled order but charge still on my card. After a month no laptop no refund. Nothing they can do wtf

@AmazonHelp Ordered a #macbook from @amazonIN was promised to be delivered yesterday, havent received yet. https://t.co/lxDwhII32C

Σχήμα 4.9: Αποτελέσματα ανάλυσης στο σύνολο των tweets

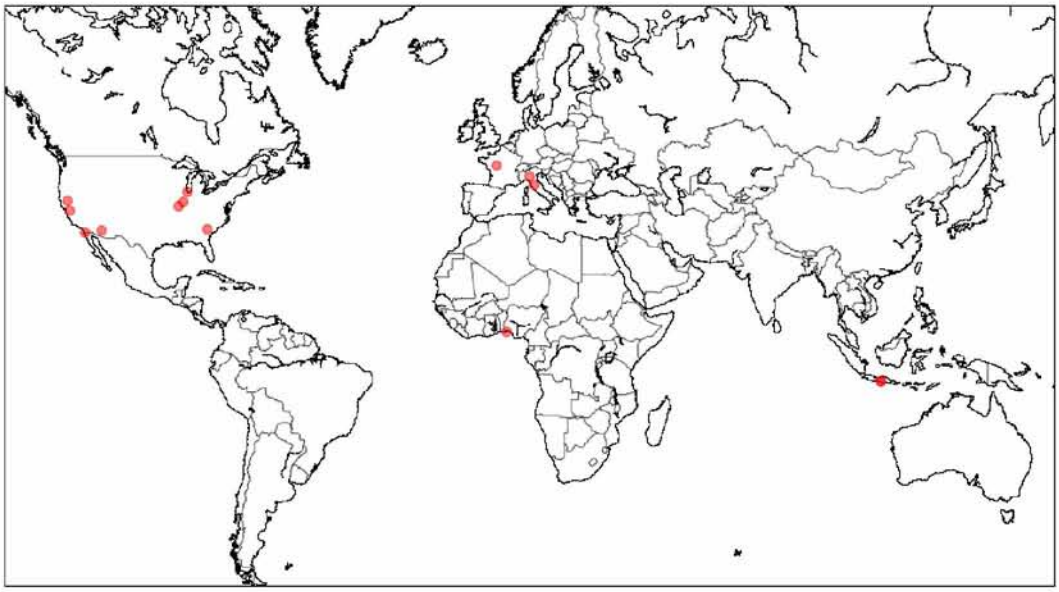


4.6 Συσχέτιση μετοχής και δημόσιας γνώμης

Εκτός από διάφορες δημοσιεύσεις στα κοινωνικά δίκτυα, γνώμες για προϊόντα και υπηρεσίες μπορούν να βρεθούν σε άρθρα και εφημερίδες. Φυσικά η γνώμη θα είναι πιο αμερόληπτη σε σχέση με το Twitter από τη στιγμή που ο συγγραφέας δεν είναι ο καθημερινός καταναλωτής, αλλά επαγγελματίας αρθρογράφος που γνωρίζει ότι αυτό που γράφει μπορεί να επηρεάσει την αγοραστική γνώμη. Έτσι θεωρώντας ότι τα άρθρα ειδήσεων έχουν αντίκτυπο στο χρηματιστήριο, έγινε μια απόπειρα μελέτης της σχέσης αυτής. Επενδυτές και αναλυτές προσπαθούν να μελετήσουν την συμπεριφορά και το πλάνο των μετοχών που έχουν στην κατοχή τους. Από την στιγμή που το χρηματιστήριο παράγει μεγάλες ποσότητες δεδομένων καθημερινά, είναι δύσκολο για ένα άτομο να λάβει υπόψιν του όλες αυτές τις πληροφορίες και να κάνει πρόβλεψη για το μέλλον μιας μετοχής. Υπάρχουν δύο είδη μεθόδων για την πρόβλεψη κίνησης μετοχών, η ανάλυση του ιστορικού της μετοχής με στόχο την πρόβλεψη και η ανάλυση των οικονομικών στοιχείων της εταιρείας για λήψη αποφάσεων.

Η συγκεκριμένη υλοποίηση έχει στόχο την ανάλυση διάφορων οικονομικών άρθρων από την ηλεκτρονική εφημερίδα The Huffington Post με σκοπό την κατηγοριοποίηση

Σχήμα 4.10: Τοποθεσίες χρηστών



των επικεφαλίδων από τα άρθρα ως θετικά, αρνητικά ή χωρίς κάποιο συναίσθημα. Εάν υπάρχουν άρθρα με θετική προδιάθεση τότε υπάρχουν περισσότερες πιθανότητες η αντίστοιχη μετοχή να αυξηθεί στο μέλλον, ενώ αν έχουν αρνητική τότε να μειωθεί.

Στο παρελθόν έχουν γίνει διάφορες απόπειρες για ανάλυση της κοινής γνώμης:

Οι Nagar και Hahsler [15] παρουσίασαν ένα μοντέλο ανάλυσης ειδήσεων από διάφορες πηγές και δημιούργησαν ένα corpus. Παράλληλα πραγματοποίησαν συναισθηματική ανάλυση σε αυτά τα άρθρα μετρώντας τις θετικές και αρνητικές λέξεις.

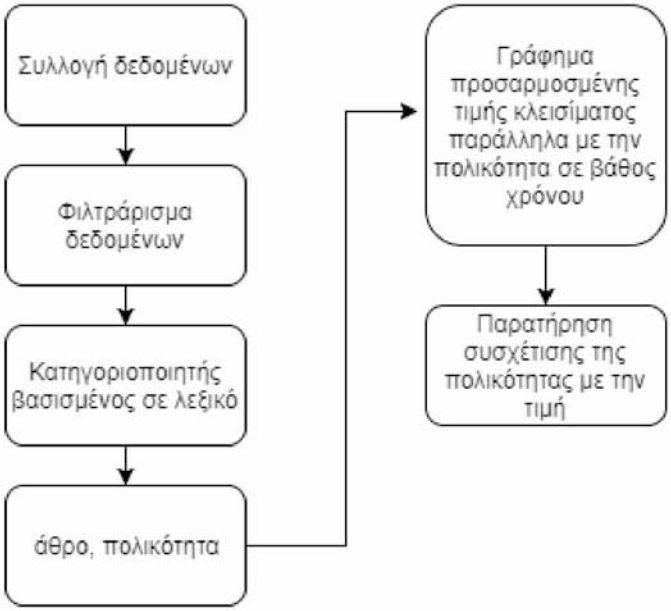
Ο Yu et al [29] παρουσίασε ένα μοντέλο εξόρυξης δεδομένων για την απόφανση του συναισθήματος από άρθρα εφημερίδων και την συσχέτιση αυτών με την ζήτηση ηλεκτρικής ενέργειας. Η πολικότητα των άρθρων προβλήθηκε σε χρονική σειρά παράλληλα με την ζήτηση της ενέργειας και την τιμή της.

Ο J. Bean [14] χρησιμοποίησε το Twitter για την εξόρυξη γνώμης και μελέτησε την ικανοποίηση του κοινού για διάφορες αεροπορικές εταιρείες.

Στην παρούσα έρευνα γίνεται ανάλυση οικονομικών άρθρων σχετικά με την πορεία της Apple μέσα από Rss Feed της διαδικτυακής ιστοσελίδας The huffington post. Σκοπός είναι να απαντηθεί το ερώτημα εάν υπάρχει κάποια συσχέτιση μεταξύ της συναισθηματικής πολικότητας των άρθρων σε βάθος χρόνου, σε σχέση με την πορεία της μετοχής. Τα δεδομένα που συλλέχθηκαν είναι της χρονικής περιόδου από 05-06-2017 έως 24-10-2017. Αυτά τα δεδομένα περιλαμβάνουν βασικά σημαντικά γεγονότα που αφορούν την Apple και τις τιμές της μετοχής της για την ίδια περίοδο. Οι τιμές αυτές περι-

λαμβάνουν τιμή ανοίγματος, μέγιστη τιμή, ελάχιστη, κλεισίματος, όγκος συναλλαγών και την προσαρμοσμένη τιμή κλεισίματος. Στην παρούσα εργασία η προσαρμοσμένη τιμή κλεισίματος θεωρήθηκε και η καθημερινή τιμή της μετοχής αυτής. Τα δεδομένα συλλέχθηκαν από το finance.yahoo.com.

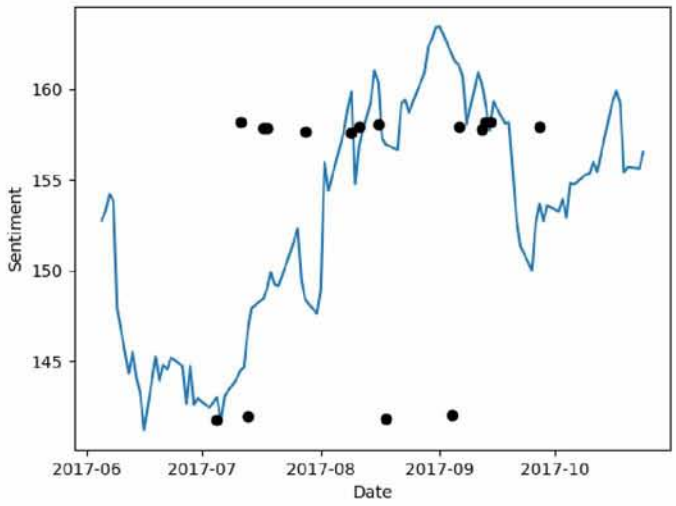
Σχήμα 4.11: Διαδικασία ανάλυσης άρθρων και συσχέτιση με τιμή μετοχής



Όπως φαίνεται από την παραπάνω εικόνα, αρχικά έγινε συλλογή των άρθρων μέσω του Rss Feed της διαδικτυακής εφημερίδας, με μοναδικά κριτήρια την λέξη κλειδί 'AAPL' και την συγκεκριμένη χρονική περίοδο. Τα δεδομένα που συλλέχθηκαν ύστερα περάστηκαν από τον αλγόριθμο συναισθηματικής ανάλυσης Vader που βασίζεται σε λεξικό και οι λέξεις κατηγοριοποιήθηκαν σε θετικές, αρνητικές ή χωρίς συναισθηματική πολικότητα. Τα αποτελέσματα φαίνονται στην παρακάτω εικόνα 4.12.

Η προσπάθεια εύρεσης της μελλοντικής τιμής μιας μετοχής μέσα από ανάλυση οικονομικών άρθρων δεν είναι πάντα καλή ιδέα. Γενικότερα οι τιμές επηρεάζονται από πάρα πολλούς παράγοντες όπου δεν λαμβάνονται υπόψιν. Βέβαια όπως φαίνεται και από τα αποτελέσματα στο γράφημα ίσως υπάρχει μία συσχέτιση μεταξύ αυτών των δύο μεταβλητών χωρίς όμως να είναι απόλυτα έμπιστη. Στην εικόνα, η πολικότητα των άρθρων για να μπορέσει να εκφραστεί σε διάγραμμα παράλληλα με την τιμή της μετοχής κανονικοποιήθηκε ώστε τα θετικά άρθρα να εμφανίζονται στο πάνω μέρος του διαγράμματος και τα αρνητικά στο κάτω. Τα άρθρα χωρίς την παρουσία συναισθήματος αγνοήθηκαν. Παρατηρείται ότι κατά την διάρκεια του Αυγούστου, έναν μήνα δηλαδή πριν η μετοχή της Apple φτάσει στην κορυφαία της τιμή για το συγκεκριμένο εξάμηνο, τα περισσότερα άρθρα είχαν θετική προδιάθεση σύμφωνα με το μοντέλο που χρησιμοποιήσαμε. Φυσικά όπως φαίνεται υπάρχουν και μερικά αρνητικά, αλλά είναι αριθμητικά λιγότερα. Η συσχέτιση αυτή μας προδιαθέτει ότι η Apple έκανε σημαντικές ανακοινώσεις σε ε-

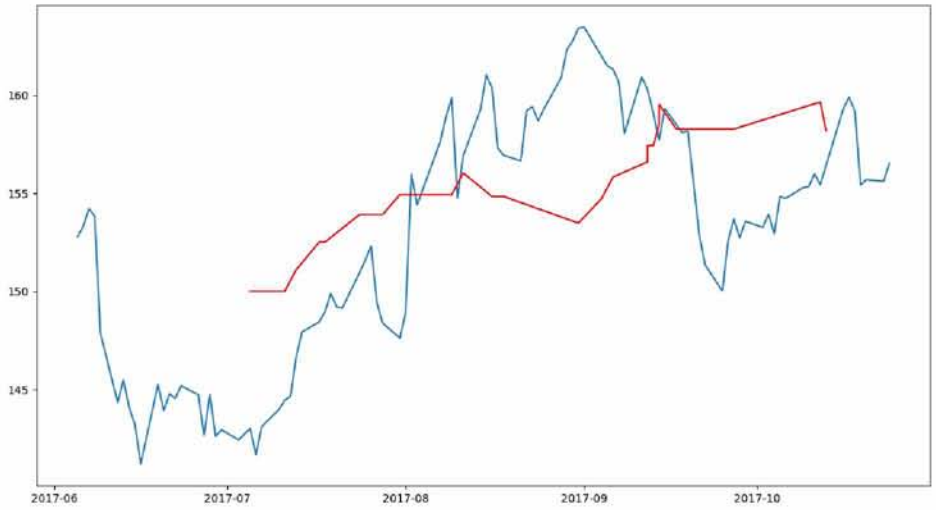
Σχήμα 4.12: Συσχετισμός τιμής με πολικότητα άρθρων



κείνο το διάστημα που ευνόησαν θετικά την τιμή της. Συγκεκριμένα τις πρώτες ημέρες του Σεπτεμβρίου έγινε το Apple Event όπου η ίδια παρουσίασε τα νέα της μοντέλα iPhone X, iPhone 8, iPhone 8 Plus, Apple Watch Series 3 και 4K Apple TV.

Συμπερασματικά μπορούμε να ομολογήσουμε την συσχέτιση της πολικότητας των άρθρων με την πορεία της μετοχής έως έναν βαθμό. Θα είχαμε ακόμα καλύτερη εικόνα και αποτελέσματα, εάν είχαμε επιλέξει περισσότερες πηγές και άρθρα.

Σχήμα 4.13: Συσχέτιση πραγματικής με προβλεφθείσας τιμής



Όπως φαίνεται και από το σχήμα 4.13, η πρόβλεψη της τιμής χρησιμοποιώντας μόνο την πολικότητα των άρθρων μάλλον δεν δίνει και τόσο ικανοποιητικά αποτελέσματα. Αγνοήθηκαν τελείως οι προηγούμενες τιμές της μετοχής και αυτό μας κόστισε.

Κεφάλαιο 5

Συμπεράσματα και μελλοντική έρευνα

Στην παρούσα διπλωματική εργασία έγινε ανάλυση συναισθήματος από κείμενο και συγκεκριμένα από tweets και άρθρα ηλεκτρονικής εφημερίδας. Τα μοντέλα που δημιουργήθηκαν ήταν βασισμένα σε λεξικό αλλά και σε αλγορίθμους μηχανικής μάθησης και η εκπαίδευσή τους έγινε μέσα από δεδομένα που ήταν ήδη κατηγοριοποιημένα χειροκίνητα.

Παρατηρήθηκε ότι ένα μοντέλο βασισμένο στο θεώρημα του Bayes είναι αρκετό ώστε να έχουμε ικανοποιητικά αποτελέσματα όσον αναφορά την κατηγοριοποίηση tweets σύμφωνα με το συναίσθημα που εκφράζουν. Μέσα από αυτή την ανάλυση δίνεται η δυνατότητα σε κάθε εταιρεία να μπορεί να έχει ένα άμεσο feedback από καταναλωτές και να βρίσκεται σε θέση να εντοπίζει τις ατέλειες των προϊόντων της ώστε να τις διορθώσει μελλοντικά. Για να γίνει αυτό εφικτό πρέπει να γίνει αντιμετώπιση προβλημάτων όπως η γλώσσα που χρησιμοποιείται στα κοινωνικά δίκτυα. Γι΄ αυτό παίζει πολύ σημαντικό ρόλο η επιλογή του αρχικού corpus εκπαίδευσης, όπου θα πρέπει αν είναι δυνατόν να έχει μηδενικές αποκλίσεις.

Η δεύτερη υλοποίηση πάνω σε άρθρα της The Huffington Post έδειξε ότι από την μία υπάρχει κατά ένα ποσοστό συσχέτιση μεταξύ της πολικότητας άρθρων και της τιμής μιας μετοχής, όμως γενικότερα οι παράγοντες που επηρεάζουν την τιμή είναι πάρα πολλοί και συχνά απρόβλεπτοι. Η συσχέτιση είναι φυσική, αν αναλογιστούμε ότι πολλοί άνθρωποι διαβάζουν και επηρεάζονται από οικονομικά άρθρα και πολλές φορές βασίζουν την άποψη τους ολοκληρωτικά σε αυτά.

Σαν μελλοντική έρευνα το σύστημα που αναλύθηκε θα μπορούσε να γίνει σε ζωντανό χρόνο. Έτσι θα δίνονται κάθε στιγμή οι πληροφορίες σχετικά με τις αναφορές ενός ή και παραπάνω προϊόντων. Ακόμα μια επέκταση θα ήταν η ανάλυση παραπάνω κοινωνικών δικτύων και όχι μόνο του Twitter που έγινε στην παρούσα φάση ώστε να έχουμε περισσότερο ολοκληρωμένα αποτελέσματα. Με αυτό τον τρόπο τα αποτελέσματα θα

φτάνουν έγκαιρα στον ενδιαφερόμενο και θα είναι σε θέση να προβεί άμεσα σε ενέργειες ώστε να αυξήσει το κέρδος.

Επίσης σχετικά με την δεύτερη υλοποίηση, είναι εφικτή η δημιουργία ενός συστήματος όπου αναλύοντας περισσότερα άρθρα, το ιστορικό των τιμών και απόψεις χρηστών θα μπορούσε να γίνει ακριβής έως έναν βαθμό στην πρόβλεψη της πορείας μετοχών. Παράλληλα, θα μπορούσε να εκπαιδευτεί ώστε να είναι σε θέση να πράττει αγορά ή πώληση ανάλογα με την πρόβλεψη και να αποφέρει το μέγιστο κέρδος.

Βιβλιογραφία

[1] Rachit Verma Dr.Rajesh Bansode Aakash Kamble, Darshan Vakharia. *Prediction Algorithm using Lexicons and Heuristics based Sentiment Analysis.* IOSR Journal of Computer Engineering, 2014.

[2] Finn AArup Nielsen. *A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs.* MSM, 2011.

[3] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau. *Sentiment Analysis of Twitter Data.* International Journal of Computer Applications, 2011.

[4] Alexandra Balahur, Ralf Steinberger, Mijail A. Kabadjov, Vanni Zavarella, Erik Van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. *Sentiment Analysis in the News*, volume abs/1309.6202. LREC, 2010.

[5] Adam Funk, Yaoyong Li, Horacio Saggion, Kalina Bontcheva, and Christian Leibold. *Opinion analysis for business intelligence applications.* ACM New York, NY, USA, 2008.

[6] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. *Large-Scale Sentiment Analysis for News and Blogs.* ICWSM, 2007.

[7] Clayton J. Hutto and Eric Gilbert. *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.* ICWSM, 2014.

[8] Neha S. Joshi and Suhasini A. Itkat. *A Survey on Feature Level Sentiment Analysis.* International Journal of Computer Science and Information Technologies, 2014.

[9] Joshi Kalyani, H. N. Bharathi, and Rao Jyothi. *Stock trend prediction using news sentiment analysis*, volume abs/1607.01958. arXiv:1607.01958, 2016.

[10] Chenghua Lin and Yulan He. *Joint sentiment/topic model for sentiment analysis.* CIKM, 2009.

[11] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. *Learning Word Vectors for Sentiment Analysis.* ACL, 2011.

[12] Prem Melville, Wojciech Gryc, and Richard D. Lawrence. *Sentiment analysis of blogs by combining lexical knowledge with text classification.* KDD, 2009.

[13] Saif Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. *NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets.* SemEval@NAACL-HLT, 2013.

[14] Mohamed M. Mostafa. *More than words: Social networks' text mining for consumer brand sentiments*, volume 40. Expert Systems with Applications, 2013.

[15] Anurag Nagar and Michael Hahsler. *Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams.* International Conference on Computer Technology and Science, 2012.

[16] Tetsuya Nasukawa and Jeonghee Yi. *Sentiment analysis: capturing favorability using natural language processing.* K-CAP, 2003.

[17] Alexander Pak and Patrick Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* LREC, 2010.

[18] Prabu palanisamy, Vineet Yadav, and Harsha Elchuri. *Serendio: Simple and Practical lexicon based approach to Sentiment Analysis.* SemEval@NAACL-HLT, 2013.

[19] Georgios Paltoglou and Mike Thelwall. *A Study of Information Retrieval Weighting Schemes for Sentiment Analysis.* ACL, 2010.

[20] Rudy Prabowo and Mike Thelwall. *Sentiment analysis: A combined approach*, volume 3. J. Informetrics, 2009.

[21] Hassan Saif, Yulan He, and Harith Alani. *Semantic Sentiment Analysis of Twitter.* International Semantic Web Conference, 2012.

[22] Reena Mahe Seema Kolkur, Gayatri Dantal. *Study of Different Levels for Sentiment Analysis.* International Journal of Current Engineering and Technology, April 2015.

[23] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. *Lexicon-Based Methods for Sentiment Analysis*, volume 37. Computational Linguistics, 2011.

[24] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. *Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification.* ACL, 2014.

[25] Vimalkumar B. Vaghela and Bhumika M. Jadav. *Analysis of Various Sentiment Classification Techniques.* International Journal of Computer Applications, April 2016.

[26] G. Vinodhini. *Sentiment Analysis and Opinion Mining: A Survey.* ICWSM, 2012.

[27] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. *Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach.* CIKM, 2011.

[28] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.* HLT/EMNLP, 2005.

[29] Wen-Bin Yu, Bih-Ru Lea, and Balasubramania Guruswamy. *A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting*, volume 5. IJEBM, 2007.

[30] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.* National Conference on "Advanced Technologies in Computing and Networking, 2011.