ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΑΘΗΜΑΤΙΚΑ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΤΩΝ ΑΠΟΦΑΣΕΩΝ»

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ – ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Διπλωματική Εργασία

«Μελέτη και αξιολόγηση τεχνικών εξόρυξης πολιτικής γνώμης σε tweets»

Θηβαίος Γιάννης

Επιβλέπων Καθηγητής

Κωτσιαντής Σωτήριος

Πάτρα, Αύγουστος 2017

Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών - Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Θηβαίος Γιάννης

Τριμελής Επιτροπή

Γράψα Θεοδούλα,

Αναπληρώτρια Καθηγήτρια Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

Καββαδίας Δημήτριος,

Επίκουρος Καθηγητής Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

Κωτσιαντής Σωτήριος,

Λέκτορας Τμήματος Μαθηματικών, Πανεπιστημίου Πατρών

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαιτέρως τον επιβλέποντα καθηγητή μου, κ. Σωτήρη Κωτσιαντή, για την καθοδήγησή του στην εκπόνηση της διπλωματικής μου εργασίας.

Θα ήθελα επίσης να ευχαριστήσω τα μέλη της τριμελούς επιτροπής, κα. Θεοδούλα Γράψα και τον κ. Δημήτρη Καββαδία για τη συνεισφορά τους κατά τη διάρκεια των μεταπτυχιακών σπουδών μου.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου, Νίκο και Ελίζα για την υποστήριξή τους καθ' όλη τη διάρκεια το φοιτητικού μου βίου.

Περίληψη

Τα κοινωνικά μέσα δικτύωσης παράγουν τεράστια ποσά δεδομένων κάθε λεπτό, γεγονός το οποίο οφείλεται στη μεγάλη υιοθέτηση και καθημερινή χρήση τους τα τελευταία χρόνια. Έχει δημιουργηθεί λοιπόν η ανάγκη για ευρύτερη αξιοποίησή τους σε διάφορους τομείς της κοινωνικής επιχειρηματικής ζωής. Γι' αυτό το λόγο δημιουργήθηκαν τεχνικές και αλγόριθμοι για την επεξεργασία κειμένων, όπως η εξόρυξη κειμένου (Text Mining) και η Ανάλυση Συναισθήματος (Sentiment Analysis).

Στην παρούσα εργασία, έχουμε συλλέξει ένα σύνολο δεδομένων από το Twitter, για τους 2 βασικούς υποψηφίους των τελευταίων αμερικανικών εκλογών (Donald Trump, Hillary Clinton). Με χρήση τεχνικών επεξεργασίας των κειμένων, προσπαθούμε να βρούμε «δημοφιλείς» λέξεις για κάθε υποψήφιο και να δημιουργήσουμε ένα μηχανισμό πρόβλεψης, με βάση τον οποίο ένα τυχαίο tweet να κατηγοριοποιείται για έναν από τους δύο υποψηφίους ως θετικό ή αρνητικό.

Δύο αλγόριθμοι επιβλεπόμενης μάθησης, ο 'αφελής' Bayes (Naïve Bayes) και οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) αποτελούν τη βάση για την παραγωγή των ταξινομητών πρόβλεψης και των οποίων την ακρίβεια συγκρίνουμε. Για την προεπεξεργασία και την εφαρμογή των αλγόριθμων χρησιμοποιείται η βιβλιοθήκη sklearn της Python. Επιπλέον, κάνουμε μια προεπισκόπηση στο εργαλείο Orange3 και ακολουθούμε μια παρεμφερής διαδικασία ανάλυσης του συνόλου δεδομένων και αξιολογούμε την ευχρηστία και την απόδοση του συγκεκριμένου εργαλείου.

Λέξεις Κλειδιά: Εξόρυξη Κειμένου, Twitter, Μηχανική Μάθηση, Ανάλυση Συναισθήματος, Πολυωνυμικός «Αφελής» Bayes, Μηχανές Διανυσμάτων Υποστήριξης, Orange3, Python

Abstract

Social networks generate every minute huge amounts of data, which occurs due to their large adoption and daily use in recent years. The need for wider use in various areas of social business life has been created. For this reason, techniques and algorithms for text processing, such as Text Mining and Sentiment Analysis, have been developed.

In this paper, we have collected a set of data from Twitter for the two key candidates in the last US election (Donald Trump, Hillary Clinton). Using text processing algorithms, we try to find "popular" words for each candidate and create a prediction mechanism, according to which a random tweet is categorized as one of the two candidates as positive or negative.

Two supervised learning algorithms, Naïve Bayes and Support Vector Machines (SVM) are the basis for the production of predictive classifiers and the accuracy of them. The Python sklearn library is used to pre-process and apply algorithms. In addition, we preview the Orange3 tool and follow the same process of analyzing the dataset and evaluate the usability and performance of the tool.

Key Words: Text Mining, Twitter, Machine Learning, Sentiment Analysis, Multinomial Naïve Bayes, Support Vector Machines, Orange3, Python

Πίνακας Περιεχομένων

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής. Δηλώνω, συνεπώς, ότι αυτή η Διπλωματική Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

Κατάλογος Πινάκων

Κατάλογος Εικόνων

1. Εξόρυξη γνώσης από κείμενα (Text Mining)

1.1 Εισαγωγή

Το πρόβλημα της εξόρυξης κειμένου έχει αποκτήσει ολοένα και μεγαλύτερη προσοχή τα τελευταία χρόνια λόγω των μεγάλων ποσοτήτων δεδομένων κειμένου που δημιουργούνται σε μια ποικιλία εφαρμογών κοινωνικού δικτύου, ιστού και άλλων εφαρμογών. Τα μη δομημένα δεδομένα είναι η συνηθέστερη μορφή δεδομένων που μπορούν να δημιουργηθούν σε οποιοδήποτε σενάριο εφαρμογής. Ως αποτέλεσμα, υπήρξε τεράστια ανάγκη να σχεδιαστούν μέθοδοι και αλγόριθμοι οι οποίοι να μπορούν να επεξεργάζονται αποτελεσματικά μια μεγάλη ποικιλία εφαρμογών κειμένου. Σε αυτό το κεφάλαιο θα πραγματοποιήσουμε μια επισκόπηση των διαφορετικών μεθόδων και αλγορίθμων που είναι οι πιο διαδεδομένες στον τομέα του κειμένου, με ιδιαίτερη έμφαση στις μεθόδους εξόρυξης.

1.2 Ορισμός

Η εξόρυξη κειμένου είναι η διαδικασία αναζήτησης ή εξαγωγής των χρήσιμων πληροφοριών από τα δεδομένα κειμένου. Είναι ένας συναρπαστικός ερευνητικός χώρος καθώς προσπαθεί να ανακαλύψει τη γνώση από αδόμητα κείμενα. Είναι επίσης γνωστή ως Εξόρυξη δεδομένων κειμένου (TDM) και ανακάλυψη γνώσης σε βάσεις δεδομένων κειμένου (KDT). Το KDT διαδραματίζει έναν ολοένα και σημαντικότερο ρόλο στις αναδυόμενες εφαρμογές, όπως η κατανόηση κειμένου (Text Understanding). [1]Η διαδικασία εξόρυξης κειμένου είναι ίδια με την εξόρυξη δεδομένων, εκτός από τα εργαλεία εξόρυξης δεδομένων που έχουν σχεδιαστεί για να χειρίζονται τα δομημένα δεδομένα, ενώ η εξόρυξη κειμένου μπορεί να χειριστεί μη δομημένα ή ημι-

[1]Vallikannu Ramanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.

δομημένα σύνολα δεδομένων όπως HTML αρχεία, emails, έγγραφα πλήρους κειμένου κλπ [1]. Η Εξόρυξη κειμένου χρησιμοποιείται για την εύρεση των νέων, παλιότερα άγνωστων πληροφοριών από διαφορετικές πηγές κειμένων.[2]

1.3 Αναπαράσταση Εγγράφου

Επειδή τα κείμενα δεν έχουν μια προκαθορισμένη μορφή, μιλάμε δηλαδή για μια αδόμητη μορφή πληροφορίας, αυτό σημαίνει ότι προσπαθούμε να επεξεργαστούμε ένα σύνολο από λέξεις, το γνωστό και «Σακούλα των λέξεων» (Bag of Words), στην οποία βρίσκονται όλες οι λέξεις του κείμενου. Ο δημοφιλέστερος τρόπος αναπαράστασης κειμένου, είναι η διανυσματική αναπαράσταση (vector representation). Κατά τη διανυσματική αναπαράσταση, κάθε κείμενο απεικονίζεται ως ένα διάνυσμα όρων (term vector), και κάθε όρος συνιστά ένα μοναδικό ανεξάρτητο χαρακτηριστικό (feature). Σε κάθε στοιχείο του διανύσματος αποδίδεται μια τιμή, η οποία εκφράζει / περιγράφει την εμφάνιση του όρου μέσα στο κείμενο.

1.3.1 Μοντέλο Διανυσματικού Χώρου(VSM)

Το Μοντέλο Διανυσματικού Χώρου(Vector Space Model - VSM) αναπαριστά τα κείμενα σε μορφή διανυσμάτων. Η διαδικασία του μοντέλου διανυσματικού χώρου μπορεί να χωριστεί σε τρία στάδια. Το πρώτο στάδιο είναι η δεικτοδότηση εγγράφων(Document Indexing), όπου οι όροι που φέρουν περιεχόμενο ανακτώνται από το κείμενο του εγγράφου. Το δεύτερο στάδιο είναι η στάθμιση των όρων που έχουν ευρεθεί για τη βελτίωση της ανάκτησης του εγγράφου που αφορά τον χρήστη. Το τελευταίο στάδιο κατατάσσει το έγγραφο σε σχέση με ένα ερώτημα σύμφωνα με ένα μέτρο ομοιότητας.[3]

[2] Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.
[3] Raghavan, V. V. and Wong, S. K. M. A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, 1986

1.3.1.1 Δεικτοδότηση Εγγράφων(Document Indexing)

Είναι προφανές ότι πολλές από τις λέξεις σε ένα έγγραφο δεν περιγράφουν το περιεχόμενο, λέξεις όπως το, 'είναι'. Χρησιμοποιώντας αυτόματη ευρετηρίαση εγγράφων, αυτές οι μη σημαντικές λέξεις (λέξεις λειτουργίας) αφαιρούνται από το διάνυσμα εγγράφων, οπότε το έγγραφο θα εκπροσωπείται μόνο από λέξεις που έχουν 'αξία'[4]. Αυτή η ευρετηρίαση μπορεί να βασιστεί στη συχνότητα των όρων, όπου όροι που έχουν υψηλή και χαμηλή συχνότητα μέσα σε ένα έγγραφο θεωρούνται λέξεις λειτουργίας.[5],[6] Αποτελείται από ένα κατάλληλο σύνολο λέξεων-κλειδιών που βασίζεται σε όλο το σώμα του κειμένου και αντιστοιχεί βάρη σε εκείνες τις λέξεις κλειδιά για κάθε συγκεκριμένο έγγραφο, μετατρέποντας έτσι κάθε έγγραφο σε ένα διάνυσμα από βάρη λέξεων-κλειδιών. Το βάρος σχετίζεται με τη συχνότητα εμφάνισης του όρου στο έγγραφο και με τον αριθμό των εγγράφων που χρησιμοποιούν αυτόν τον όρο.

1.3.1.2 Στάθμιση Όρων (Term Weighting)

Η στάθμιση όρων είναι μια σημαντική έννοια που καθορίζει την επιτυχία ή την αποτυχία του συστήματος ταξινόμησης. Δεδομένου ότι διαφορετικοί όροι έχουν διαφορετικό επίπεδο σημαντικότητας σε ένα κείμενο, το βάρος ενός όρου σχετίζεται με κάθε όρος ως ένας σημαντικός δείκτης.[7] Τα τρία βασικά συστατικά που επηρεάζουν τη σημαντικότητα ενός όρου σε ένα έγγραφο είναι:

- Συχνότητα του Όρου(Term Frequency - TF)
- Αντίστροφη συχνότητα του εγγράφου(Inverse Document Frequency - IDF)

[4] Salton, Gerard. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[5] Luhn, H. P. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2 (2), p. 159-165 and 317, April 1958.
[6] van Rijsbergen, C. J. Information retrieval. Butterworths, 1979
[7] Salton, G. and Buckley, C. (1988) "Term weighting approaches In automatic text retrieval, Information Processing and Management",Vol. 24, No.5, Pp. 513 - 523.

- Κανονικοποίηση του μήκους του κειμένου(Document Length Normalization)

Εκφράζει τη σημασία της λέξης στο έγγραφο. Η αντίστροφη συχνότητα εγγράφου κάθε λέξης στη βάση δεδομένων του εγγράφου (IDF) είναι ένα βάρος που εξαρτάται από την κατανομή κάθε λέξης στη βάση δεδομένων του εγγράφου. Εκφράζει τη σημασία της κάθε λέξης στη βάση δεδομένων του εγγράφου[8].

Η TF-IDF είναι μια τεχνική η οποία χρησιμοποιεί τόσο την TF όσο και την IDF για να προσδιοριστεί το βάρος μιας λέξης ενός εγγράφου. Το TF-IDF σχήμα είναι πολύ δημοφιλές στο πεδίο της ταξινόμησης κειμένων και σχεδόν όλα τα άλλα σχήματα είναι παραλλαγές αυτού. Δοθέντος μιας συλλογής κειμένων 'D', μια λέξη w και ένα ανεξάρτητο κείμενο D, το w_d περιγράφεται από την παρακάτω σχέση:

$$W_d = f_{w,d} * \log(|D| / f_{w,D}) \text{, όπου}$$

- $f_{w,d}$ ή TF είναι ο αριθμός που το 'w' εμφανίζεται μέσα στο έγγραφο 'd'
- D είναι το μέγεθος του συνόλου δεδομένων
- $f_{w,D}$ ή IDF είναι ο αριθμός των κειμένων μέσα στα οποία εμφανίζεται η 'w'

Το αποτέλεσμα της μετρικής TF-IDF είναι ένα διάνυσμα με επιμέρους βάρη.

1.3.1.3 Μετρικές Ομοιότητας

Η ομοιότητα στα μοντέλα διανυσματικού χώρου προσδιορίζεται χρησιμοποιώντας σχετικούς συντελεστές που βασίζονται στο εσωτερικό γινόμενο του διανύσματος εγγράφων και του διανύσματος ερωτήματος, όπου η αλληλοεπικάλυψη λέξεων υποδεικνύει ομοιότητα. Το εσωτερικό γινόμενο συνήθως κανονικοποιείται. Το πιο δημοφιλές μέτρο ομοιότητας είναι το

[8] Diao, Q. and Diao, H. (2000) "Three Term Weighting and Classification Algorithms in Text Automatic Classification", The Fourth International Conference on High-Performance Computing in the Asia-Pacific Region,Vol. 2, P.629.

διάνυσμα συνημιτόνου, το οποίο υπολογίζει τη γωνία μεταξύ των δύο διανυσμάτων που αναφέραμε προηγουμένως.

1.4 Τεχνικές του Text Mining

1.4.1 Ανάκτηση Πληροφορίας

Η έννοια της ανάκτησης πληροφοριών (IR) έχει αναπτυχθεί σε σχέση με τα συστήματα βάσεων δεδομένων εδώ και πολλά χρόνια. Η ανάκτηση πληροφοριών αποτελεί η συσχέτιση και η ανάκτηση πληροφοριών από ένα μεγάλο αριθμό κειμένων. [9]Τα συστήματα ανάκτησης πληροφορίας και βάσεων δεδομένων χειρίζονται διάφορα είδη δεδομένων. Ορισμένα προβλήματα στα συστήματα βάσεων δεδομένων δεν εμφανίζονται σε αυτά της ανάκτησης πληροφοριών, όπως ο έλεγχος ταυτόχρονης λειτουργίας, η επαναφορά, η διαχείριση αλλαγών και η ενημέρωση. Επίσης, ορισμένα συνηθισμένα προβλήματα ανάκτησης πληροφοριών συνήθως δεν συναντώνται σε συμβατικά συστήματα βάσεων δεδομένων, όπως μη δομημένα έγγραφα, εκτιμώμενη αναζήτηση βασισμένη σε λέξεις-κλειδιά και η έννοια της συνάφειας. Λόγω της τεράστιας ποσότητας πληροφοριών κειμένου, η ανάκτηση πληροφοριών έχει βρει πολλές εφαρμογές. Υπάρχουν πολλά συστήματα ανάκτησης πληροφοριών, όπως ηλεκτρονικά συστήματα καταλόγου βιβλιοθηκών, ηλεκτρονικά συστήματα διαχείρισης εγγράφων και οι πιο πρόσφατα αναπτυγμένες μηχανές αναζήτησης στο Web.

1.4.2 Εξαγωγή Πληροφορίας

Ο στόχος των μεθόδων εξαγωγής πληροφοριών (IE) είναι η εξαγωγή χρήσιμων πληροφοριών από το κείμενο. Προσδιορίζει την εξαγωγή οντοτήτων,

[9] R.Sagayam, S.Srinivasan, S.Roshini, “A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques”. Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

συμβάντων και σχέσεων από ημι-δομημένο ή αδόμητο κείμενο. Οι πιο χρήσιμες πληροφορίες, όπως το όνομα του ατόμου, η τοποθεσία και ο οργανισμός, εξάγονται χωρίς την κατάλληλη κατανόηση του κειμένου [10]. Η IE ασχολείται με την εξαγωγή σημασιολογικών πληροφοριών από το κείμενο. Η IE μπορεί να περιγραφεί ως η κατασκευή δομημένης εικόνας επιλεγμένων σχετικών πληροφοριών που προέρχονται από κείμενα.

1.4.3 Κατηγοριοποίηση

Η κατηγοριοποίηση των κειμένων είναι ένα είδος «επιβλεπόμενης» μάθησης όπου οι κατηγορίες είναι γνωστές εκ των προτέρων και είναι σταθερές σε όλη την εξέλιξη για κάθε εκπαιδευτικό έγγραφο. Η βασική του προβλεπόμενη χρήση ήταν για την εύρεση επιστημονικής βιβλιογραφίας μέσω ελεγχόμενων λέξεων. Ήταν μόλις η δεκαετία του ’90, όταν το πεδίο της επεξεργασίας κειμένων αναπτύχθηκε πλήρως με τη διαθεσιμότητα συνεχώς αυξανόμενου αριθμού εγγράφων κειμένου σε ψηφιακή μορφή και την απαίτηση να οργανωθούν για ευκολότερη χρήση [11]. Η κατηγοριοποίηση είναι η ανάθεση των κανονικών γλωσσικών εγγράφων σε προκαθορισμένο σύνολο θεμάτων ανάλογα με το περιεχόμενό τους. Πρόκειται για μια συλλογή εγγράφων κειμένου, τη διαδικασία εύρεσης ακριβούς θέματος ή θεμάτων για κάθε έγγραφο. Σήμερα, η αυτοματοποιημένη κατηγοριοποίηση κειμένου εφαρμόζεται σε ποικίλα πλαίσια από την κλασσική αυτόματη ή ημιαυτόματη αναζήτηση κειμένων σε εξατομικευμένες διαφημίσεις, φιλτράρισμα ανεπιθύμητων μηνυμάτων και κατηγοριοποίηση ιστοσελίδας κάτω από ιεραρχικούς καταλόγους, αυτόματη δημιουργία μεταδεδομένων και ανίχνευση είδους κειμένου, παρακολούθηση θεμάτων κ.α [12]. Πρόκειται για ένα καυτό θέμα στη μηχανική μάθηση στον τομέα της έρευνας σήμερα.

[10] Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319- 7242, Vol 3 Issue 5, May 2014, pp.5621-5625
[11] Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.
[12] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.

1.4.4 Ομαδοποίηση

Η ομαδοποίηση είναι ένα από τα πιο ενδιαφέροντα και σημαντικά θέματα στην εξόρυξη κειμένου. Σκοπός του είναι να εντοπίσει εγγενείς δομές πληροφόρησης και να τις οργανώσει σε σημαντικές υποομάδες για περαιτέρω μελέτη και ανάλυση. Είναι μια διαδικασία χωρίς επιτήρηση, μέσω της οποίας τα αντικείμενα ταξινομούνται σε ομάδες που ονομάζονται συστάδες. Το πρόβλημα είναι η ομαδοποίηση της δοσμένης μη κατηγοριοποιημένης συλλογής σε σημαντικές ομάδες χωρίς προηγούμενη πληροφόρηση. Όλες οι ετικέτες που σχετίζονται με αντικείμενα λαμβάνονται αποκλειστικά από τα δεδομένα. Για παράδειγμα, η ομαδοποίηση εγγράφων βοηθά στην ανάκτηση δημιουργώντας συνδέσμους μεταξύ των σχετικών εγγράφων, τα οποία με τη σειρά τους επιτρέπουν την ανάκτηση σχετικών εγγράφων μόλις κάποιο από τα έγγραφα θεωρηθεί σχετικό με ένα ερώτημα.[13] Η ομαδοποίηση είναι χρήσιμη σε πολλούς τομείς εφαρμογής όπως η βιολογία, η εξόρυξη δεδομένων, η αναγνώριση προτύπων, η ανάκτηση εγγράφων, η κατάτμηση της εικόνας, η ταξινόμηση προτύπων, η ασφάλεια, η επιχειρηματική ευφυΐα και η αναζήτηση στο Web. Η ανάλυση της ομαδοποίησης μπορεί να χρησιμοποιηθεί ως ένα αυτόνομο εργαλείο εξόρυξης κειμένου για την επίτευξη της κατανομής δεδομένων ή ως ένα βήμα προεπεξεργασίας για άλλους αλγορίθμους εξόρυξης κειμένου που λειτουργούν στις ανιχνευόμενες συστάδες.

1.4.5 Περιληπτική Παρουσίαση της πληροφορίας(Summarization)

Η σύνοψη είναι η διαδικασία της αυτόματης δημιουργίας μια συμπιεσμένης έκδοσης ενός συγκεκριμένου κειμένου, το οποίο παρέχει χρήσιμες πληροφορίες για το χρήστη. Σε ένα μεγάλο οργανισμό ή εταιρεία, ο ερευνητής δεν έχει το χρόνο να διαβάσει όλα τα έγγραφα, ως εκ τούτου συνοψίζει τα έγγραφα και επισημαίνει τα κύρια σημεία. Η περίληψη παράγεται από ένα ή

[13] Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319- 7242, Vol 3 Issue 5, May 2014, pp.5621-5625

περισσότερα κείμενα και περιέχει μια σημαντική μερίδα των πληροφοριών, που αφενός έχει μειωμένο μήκος αφετέρου προσπαθεί να διατηρήσει τη γενική έννοια, όπως είναι τα πρωτότυπα κείμενα. Η σύνοψη κειμένου περιλαμβάνει διάφορες μεθόδους που αναπτύσσονται στην κατηγοριοποίηση κειμένου, όπως νευρωνικά δίκτυα, δέντρα απόφασης, μοντέλα παλινδρόμησης, δέντρα απόφασης, ασαφής λογική κ.α. Ωστόσο, όλες αυτές οι μέθοδοι έχουν ένα κοινό πρόβλημα, δηλαδή, την ποιότητα της ανάπτυξης των ταξινομητών, η οποία εξαρτάται σε μεγάλο βαθμό από το είδος του κειμένου που συνοψίζεται.

1.4.6 Αναγνώριση γλώσσας (Language identification)

Πρόκειται για ένα εργαλείο, το οποίο προσδιορίζει σε ποια γλώσσα είναι γραμμένο ένα κείμενο ή αν αυτό το κείμενο περιέχει κομμάτια κειμένου σε διάφορες γλώσσες και ποιες είναι αυτές. Ερευνητές έχουν μελετήσει τόσο γλωσσολογικά όσο και στατιστικά μοντέλα για αυτή τη διαδικασία. Υπάρχουν μέθοδοι, όπως η αναγνώριση γλώσσας βάσει της κατανομής των γραμμάτων στο κείμενο(Letter Based Language Identification), αλλά και η n-gram προσέγγιση, με την οποία εξάγεται ένα διάνυσμα χαρακτηριστικών που αναφέρεται στη γλώσσα που είναι γραμμένο το κείμενο.

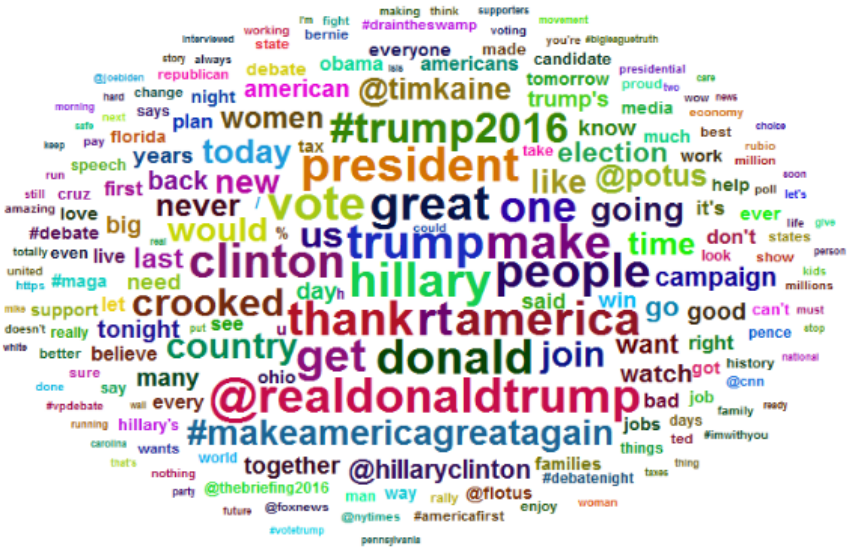
1.4.7 Κανόνες Συσχέτισης (Association Rules)

Στην ανάλυση των κανόνων συσχετίσεων ουσιαστικά μιλάμε για τη σχέση μεταξύ των χαρακτηριστικών που εξάγονται από ένα κείμενο.[14] Υπάρχει λοιπόν μια συνθήκη με βάση ένα πρότυπο κειμένου. Δηλαδή αν μια λέξη μέσα σε ένα κείμενο περιέχεται μέσα στο πρότυπο και βρίσκεται σε μια χ απόσταση από μια άλλη δεδομένη λέξη τότε η συνθήκη αυτή διατηρείται με μεγάλη πιθανότητα.

[14] Dunham Margaret H., “Data Mining Introductory and Andanced Topics”, Pearson Education Inc, (2003).

1.4.8 Οπτικοποίηση (Visualization)

Το εργαλείο αυτό με τη χρήση της εξαγωγής χαρακτηριστικών (feature extraction) μπορεί και κατασκευάζει μια γραφική αναπαράσταση μιας συλλογής εγγράφων. Η προσέγγιση αυτή βοηθάει σημαντικά στον άμεσο εντοπισμό κάποιων σημαντικών στοιχείων που μας ενδιαφέρουν. Πχ αν μια λέξη αποτυπώνεται με μεγάλη συμβολοσειρά, θεωρείται σημαντική(Word Cloud).



Εικόνα 1: Παράδειγμα Word Cloud

1.5 Προεπεξεργασία Δεδομένων

Η φάση της προεπεξεργασίας μετατρέπει τα αρχικά δεδομένα σε μια δομή έτοιμη για εξόρυξη δεδομένων, όπου εντοπίζονται σημαντικά χαρακτηριστικά κειμένου που χρησιμεύουν για τη διαφοροποίηση μεταξύ κειμένων-κατηγοριών. Πρόκειται για τη διαδικασία ενσωμάτωσης ενός νέου εγγράφου σε ένα σύστημα ανάκτησης πληροφορίας.[15] Ένας αποτελεσματικός προεπεξεργαστής αντιπροσωπεύει αποτελεσματικά το έγγραφο τόσο από άποψη χώρου (για την

[15] C. C. Aggarwal and C.-X. Zhai, “Mining Text Data”, New York, NY, USA: Springer, 2012.

αποθήκευση του εγγράφου) όσο και από χρόνο (για επεξεργασία αιτήσεων ανάκτησης) και διατηρεί καλή απόδοση ανάκτησης (ακρίβεια και ανάκληση). Αυτή η φάση είναι η πιο κρίσιμη και ολοκληρωμένη διαδικασία που οδηγεί στην αντιπροσώπευση κάθε εγγράφου από ένα επιλεγμένο σύνολο όρων ευρετηρίου. Ο κύριος στόχος της προεπεξεργασίας είναι η απόκτηση των βασικών χαρακτηριστικών ή όρων κλειδιών από ηλεκτρονικά έγγραφα κειμένων ειδήσεων και η ενίσχυση της σχετικότητας μεταξύ λέξης και εγγράφου και η συσχέτιση λέξης και κατηγορίας.

Ο στόχος πίσω από την προεπεξεργασία είναι η αντιπροσώπευση κάθε εγγράφου ως διανυσματικό στοιχείο, δηλαδή να διαχωρίζει το κείμενο σε μεμονωμένες λέξεις. Η επιλογή της λέξης κλειδιού αποτελεί το κύριο βήμα προεπεξεργασίας που είναι απαραίτητο για την ευρετηρίαση των εγγράφων. Αυτό το βήμα είναι κρίσιμο για τον προσδιορισμό της ποιότητας της επόμενης βαθμίδας, δηλαδή του σταδίου ταξινόμησης. Είναι σημαντικό να επιλέξουμε τις σημαντικές λέξεις-κλειδιά που φέρουν τη σημασία και να απορρίψουμε τις λέξεις που δεν συμβάλλουν στη διάκριση μεταξύ των εγγράφων.

1.5.1 Αφαίρεση των stopwords

Πολλές από τις πιο συχνά χρησιμοποιούμενες λέξεις στα αγγλικά είναι άχρηστες στην ανάκτηση πληροφοριών (IR Retrieval) και την εξόρυξη κειμένου. Αυτές οι λέξεις ονομάζονται λέξεις "Stopwords". Πρόκειται για λέξεις που είναι συγκεκριμένες για κάθε γλώσσα και είναι απαραίτητες για τη σύνταξη της. Είναι συχνές λέξεις που δεν περιέχουν πληροφορίες, δηλαδή αντωνυμίες, προθέσεις, συζεύξεις κ.α. Στην αγγλική γλώσσα, υπάρχουν περίπου 400-500 stopwords. Παραδείγματα τέτοιων λέξεων περιλαμβάνουν τις λέξεις «the», «of», «and», «to». Το πρώτο βήμα κατά την προεπεξεργασία είναι η αφαίρεση αυτών των λέξεων, η οποία έχει αποδειχθεί πολύ σημαντική [16].

[16] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.

1.5.2 Stemming

Οι τεχνικές λημματοποίησης χρησιμοποιούνται για να ανακαλύψουν τη ρίζα μιας λέξης. Η λημματοποίηση μετατρέπει τις λέξεις στις ρίζες τους, η οποία ενσωματώνει πολλές εκδοχές της λέξης αυτής. Η υπόθεση είναι ότι οι λέξεις με την ίδια ρίζα περιγράφουν ως επί το πλείστον ίδιες ή σχετικά στενές έννοιες στο κείμενο και έτσι οι λέξεις μπορούν να συγχωνευθούν με αυτή τη διαδικασία. Για παράδειγμα, οι λέξεις 'user', 'users', 'used', 'using', μπορούν να λημματοποιηθούν στη λέξη 'use'.

1.5.3 Part of Speech Tagging

Με την τεχνική αυτή κάθε λέξη κατηγοριοποιείται στο μέρος του λόγου το οποίο ανήκει (επίθετο, ουσιαστικό, ρήμα κ.α). Το πρόβλημα αυτό λύνεται με τη χρήση γνωστών ταξινομητών. Η ανάγκη χρήσης ταξινομητή προκύπτει από το γεγονός ότι για τις λέξεις ενός κειμένου ως μονάδες δεν μπορεί να αναγνωριστεί μονοσήμαντα το μέρος του λόγου στο οποίο ανήκουν, λόγω πολλών αμφισημιών (π.χ. η λέξη "συντομεύσεις", ανάλογα με τα συμφραζόμενα, θα μπορούσε να λειτουργεί ως ρήμα ή ως ουσιαστικό).[17]

1.5.4 Μείωση διαστασιμότητας (Dimensionality Reduction)

Η συχνότητα εγγράφων (DF) είναι ο αριθμός των εγγράφων στα οποία εμφανίζεται ένας όρος. Το DF κατώφλι είναι η απλούστερη τεχνική για τη μείωση του λεξιλογίου. Η εξάλειψη των stopwords που εξηγήθηκε προηγουμένως, αφαιρεί όλες τις λέξεις υψηλής συχνότητας που δεν σχετίζονται με τη διαδικασία ταξινόμησης, ενώ επίσης απομακρύνει και σπάνιες λέξεις. Δηλαδή, όλες οι λέξεις που εμφανίζονται σε έγγραφα μικρότερα από το 'X' της συλλογής κειμένου δεν θεωρούνται χαρακτηριστικά, όπου το 'X' είναι ένα

[17] Porter, M. (1980) "An algorithm for suffix stripping, Program",Vol. 14, No. 3, Pp. 130–137

προκαθορισμένο όριο. Το DF κατώφλι βασίζεται στην υπόθεση ότι οι σπάνιες λέξεις δεν παρέχουν πληροφορία για την πρόβλεψη κατηγοριών.

1.6 Μηχανική Μάθηση

Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη[18]. Το 1959, ο Arthur Samuel ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μάθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί[19]. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα[20] και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.[21]

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.[22]

1.6.1 Ορισμός

Ο Tom M. Mitchell πρότεινε έναν πιο επίσημο ορισμό που χρησιμοποιείται ευρέως: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε

[18] http://www.britannica.com/EBchecked/topic/1116194/machine-learning
[19] Phil Simon (March 18, 2013). Too Big to Ignore: The Business Case for Big Data. Wiley, σελ. 89. ISBN 978-1-118-63817-0
[20] Ron Kohavi; Foster Provost (1998). «Glossary of terms». Machine Learning 30: 271–274.
[21] Machine learning and pattern recognition "can be viewed as two facets of the same field."
[22] «Machine Learning: What it is and why it matters». www.sas.com

εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E».[23] Αυτός ο ορισμός είναι σημαντικός για τον καθορισμό της μηχανικής μάθησης σε βασικό λειτουργικό πλαίσιο παρά με γνωστικούς όρους, ακολουθώντας έτσι την πρόταση του Alan Turing στην εργασία του «Υπολογιστικές μηχανές και Νοημοσύνη», ότι το ερώτημα αν μπορούν οι μηχανές να σκεφτούν, μπορεί να αντικατασταθεί με το ερώτημα αν μπορούν οι μηχανές να κάνουν αυτό που εμείς (ως σκεπτόμενες οντότητες) μπορούμε να κάνουμε.[24]

1.6.2 Είδη Μηχανικής Μάθησης

Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δυο είδη:

- *Επιβλεπόμενη Μάθηση (Supervised Learning):* Το υπολογιστικό πρόγραμμα δέχεται τις παραδειγματικές εισόδους καθώς και τα επιθυμητά αποτελέσματα από έναν «δάσκαλο», και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.

- *Μη*-Επιβλεπόμενη Μάθηση (Unsupervised Learning): Χωρίς να παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, πρέπει να βρεί την δομή των δεδομένων εισόδου. Η Μημεπιτηρούμενη μάθηση μπορεί να είναι αυτοσκοπός (ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή μέσο για ένα τέλος (χαρακτηριστικό της μάθησης).

Μια άλλη κατηγοριοποίηση των προβλημάτων μηχανικής μάθησης προκύπτει όταν κάποιος θεωρήσει το επιθυμητό αποτέλεσμα του συστήματος μηχανικής μάθησης.[15]

[23] Mitchell, T. (1997). Machine Learning, McGraw Hill, Machine Learning, McGraw Hill, p.2
[24] Harnad, Stevan (2008), «The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence», στο: Epstein, Robert; Peters, Grace, επιμ., The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer, Kluwer

Στην ταξινόμηση, τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Αυτό συνήθως εμπίπτει στην επιτηρούμενη μάθηση. Τα φίλτρα Spam είναι ένα παράδειγμα ταξινόμησης, όπου οι είσοδοι είναι τα emails ή άλλα μηνύματα και οι κλάσεις είναι "spam" και "όχι spam".[25]

Στην παλινδρόμηση, επίσης πρόβλημα επιτηρούμενης μάθησης, τα αποτελέσματα είναι συνεχή και όχι διακριτά.

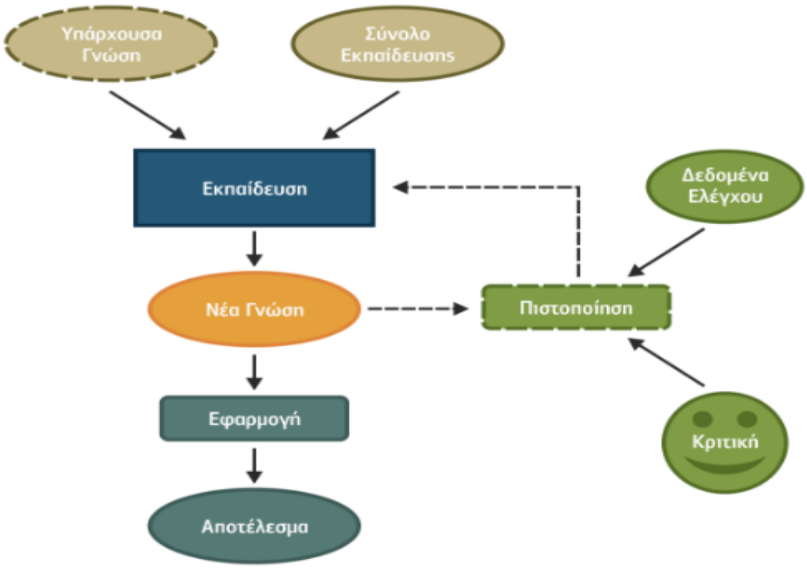
Στην συσταδοποίηση, ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτόν τον διαχωρισμό τυπική εργασία μη επιτηρούμενης μάθησης.

Στην εκτίμηση πυκνότητας βρίσκει την κατανομή των δεδομένων εισόδου σε κάποιο χώρο.

Σε προβλήματα μείωσης διαστασιμότητας (dimensionality reduction), τα δεδομένα απλοποιούνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων. Το στατιστικό μοντέλο θεμάτων (Topic modeling) είναι ένα σχετικό πρόβλημα, όπου η μηχανή καλείται να βρει έγγραφα που καλύπτουν παρόμοια θέματα από ένα σύνολο εγγράφων γραμμένων σε φυσική γλώσσα.

Στην εικόνα 2, αποτυπώνεται ο γενικός τρόπος λειτουργίας των αλγορίθμων Μηχανικής Μάθησης. Η βασικότερη φάση κάθε αλγόριθμου είναι η εκπαίδευση, όπου ο αλγόριθμος χρησιμοποιεί ως είσοδο ένα **σύνολο δεδομένων εκπαίδευσης** (training set) προς επίτευξη του σκοπού του, τη δημιουργία νέας γνώσης. Επιπλέον, μπορεί είτε να χρησιμοποιήσει λιγότερο ή περισσότερο την υπάρχουσα γνώση είτε να μην τη χρησιμοποιήσει καθόλου.

[25] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258.



Εικόνα 2: Γενικός τρόπος λειτουργίας Μηχανικής Μάθησης[26]

Την εκπαίδευση ακολουθεί η φάση της πιστοποίησης της παραγόμενης νέας γνώσης. Συνήθως, η πιστοποίηση πραγματοποιείται καταρχάς από τον ίδιο τον αλγόριθμο μέσω διαδικασιών ανάκλησης (recall) με τη βοήθεια **δεδομένων ελέγχου** (test data) και, στη συνέχεια, μέσω κριτικής που κάνει ο χρήστης βάσει των γνώσεων που διαθέτει για το πρόβλημα που επιχειρεί να λύσει ο αλγόριθμος. Τέλος, η νέα γνώση δίνεται προς χρήση σε εφαρμογές στις οποίες είναι απαραίτητη, για να λυθούν πραγματικά προβλήματα.

1.6.3 Αλγόριθμοι Μηχανικής Μάθησης

Γνωστοί αλγόριθμοι για την επιβλεπόμενη μάθηση είναι ενδεικτικά:[27]

- Αφελής Bayes (Naïve Bayes)
- Μέθοδος k- Πλησιέστερων Γειτόνων(k-NN)
- Μηχανές Διανυσμάτων Υποστήριξης(SVM)

[26] https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf
[27] http://blogs.sas.com/content/sascom/2015/08/11/an-introduction-to-machine-learning/

- Δέντρα Απόφασης (DT)
- Νευρωνικά Δίκτυα(NN)
- Κανόνες Κατηγοριοποίησης(Classification Rules)
- K-Means
- Random Forest

Αργότερα, θα αναλύσουμε δύο από αυτούς, Αφελής Bayes και Μηχανές Διανυσμάτων Υποστήριξης, τους οποίους χρησιμοποιούμε και στην υλοποίησή μας.

Στο Κεφάλαιο 2 θα κάνουμε μια εισαγωγή στη σημασιολογική ανάλυση κειμένων και θα αναφερθούμε στο Twitter ως μέσο κοινωνικής δικτύωσης μέσα από το οποίο με κατάλληλη επεξεργασία μπορούμε να αντλήσουμε χρήσιμες πληροφορίες σχετικά με απόψεις, συναισθήματα, προφίλ χρηστών κ.α

2. Εξόρυξη γνώμης και Ανάλυση συναισθήματος

2.1 Εισαγωγή

Οι πληροφορίες για το κείμενο περιλαμβάνουν δύο είδη: πληροφορίες σχετικά με τα γεγονότα και πληροφορίες κοινής γνώμης. Η πραγματική πληροφόρηση είναι η αντικειμενική δήλωση σχετικά με τα αντικείμενα και οι πληροφορίες της γνώμης είναι υποκειμενική δήλωση που εκφράζει την άποψη των ανθρώπων για τα αντικείμενα. Οι περισσότερες έρευνες σχετικά με την επεξεργασία κειμένων πληροφοριών επικεντρώνονται στην εξόρυξη και την ανάκτηση πληροφοριών από γεγονότα. Όμως, όλο και περισσότεροι ερευνητές και επιχειρηματίες αρχίζουν να ενδιαφέρονται για την εξόρυξη πληροφοριών γνώμης.

Η άνοδος του Παγκόσμιου Ιστού μας φέρνει πολλές πληροφορίες που παράγονται από τους χρήστες (π.χ. φόρουμ, blog, κριτικές σε sites), τα οποία περιέχει μεγάλο αριθμό στοιχείων γνώμης. Όταν κάποιος θέλει να δει πόσο καλό είναι ένα προϊόν που θέλει να αγοράσει, δεν είναι απαραίτητο να ρωτήσει άλλους φίλους αν μπορεί να πάρει πληροφορίες σχετικά με το προϊόν στο διαδίκτυο. Πριν από τις πολιτικές εκλογές, η υπολογιστική έρευνα για το τι σκέφτονται οι ψηφοφόροι μπορεί να γίνει με αυτόν τον τρόπο. Ομοίως, οι κατασκευαστές μπορούν να διεξάγουν έρευνα αγοράς μέσω πληροφοριών εξόρυξης γνώμης στο Διαδίκτυο, προκειμένου να γνωρίζουν ποια προϊόντα επιθυμούν οι τρέχοντες πελάτες. Όλοι αυτοί οι λόγοι προωθούν την ανάπτυξη της έρευνας σχετικά με την εξόρυξη γνώμης και τη συναισθηματική ανάλυση.

2.2 Ορισμός

Ο όρος εξόρυξη γνώμης εμφανίζεται στο έγγραφο [28] με τίτλο “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews” από τους Dave et al. Ο ορισμός που δίνεται εκεί για το **Opinion Mining** είναι:

[28] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of WWW, pp. 519–528, 2003.

Πρόκειται για την επεξεργασία ενός συνόλου αποτελεσμάτων αναζήτησης για ένα συγκεκριμένο στοιχείο, δημιουργώντας μια λίστα χαρακτηριστικών του προϊόντος (ποιότητα, χαρακτηριστικά κ.λπ.) και συγκεντρώνοντας τις απόψεις για καθένα από αυτά (κακή, ουδέτερη, καλή).

Ωστόσο, ο όρος έχει ερμηνευτεί πρόσφατα περιέχοντας πολλές ευρύτερες διαφορετικές πτυχές της ανάλυσης στο κείμενο αξιολόγησης.

Αντίστοιχα, ο όρος **Sentiment Analysis** έχει παρόμοια έννοια και χρήση με το Opinion Mining. Στις εργασίες των Das&Chen[29] και Tong [30] εμφανίζει τον όρο συναίσθημα, που χρησιμοποιείται σε σχέση με την αυτόματη ανάλυση του κειμένου αξιολόγησης και την παρακολούθηση των προγνωστικών κρίσεων. Σε πολλές μελέτες, ο όρος «ανάλυση συναισθημάτων» επικεντρώνεται στην ειδική εφαρμογή της ταξινόμησης κριτικών/απόψεων (θετικών ή αρνητικών). Έτσι, μερικοί άνθρωποι προτείνουν ότι ο όρος πρέπει να αναφέρεται ειδικά σε αυτό τον στενό επιστημονικό τομέα. Ωστόσο, πολλοί εξακολουθούν να εξηγούν τον όρο ευρύτερα για να δώσουν νόημα στην υπολογιστική αντιμετώπιση της γνώμης, του συναισθήματος και της υποκειμενικότητας μέσα σε ένα κείμενο.

Ως εκ τούτου, όταν εφαρμόζεται ευρεία ερμηνεία, η εξόρυξη γνώμης και η ανάλυση συναισθημάτων υποδηλώνουν το ίδιο πεδίο μελέτης. Στη συνέχεια, θα παρουσιάσουμε πιο συγκεκριμένους ορισμούς ορισμένων στοιχείων που περιέχονται στη μελέτη, όπως ο κάτοχος γνώμης, το χαρακτηριστικό γνώρισμα και ο σημασιολογικός προσανατολισμός της άποψης, αλλά και τα επίπεδα της ανάλυσης συναισθήματος.

2.3 Βασικά στοιχεία μιας γνώμης

Η γνώμη μπορεί να εκφράζεται σε οτιδήποτε όπως προϊόν, ταινία, θέμα, άτομο, οργάνωση ή γεγονός. Ο όρος αντικείμενο χρησιμοποιείται για να δηλώσει την οντότητα στην οποία δίνεται η γνώμη. Ένα αντικείμενο μπορεί να αποσυντεθεί

[29] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," in Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.
[30] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in Proceedings of the Workshop on Operational TextClassification (OTC), 2001.

με το μέρος της σχέσης. Έχει ένα σύνολο στοιχείων (μέρη) και ένα σύνολο χαρακτηριστικών.

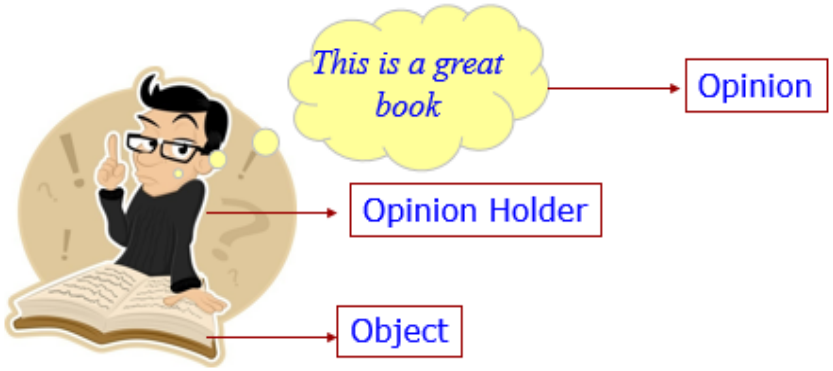
Αντικείμενο (Object): Ένα αντικείμενο είναι μια οντότητα που μπορεί να είναι θέμα, προϊόν, γεγονός, άτομο ή οργάνωση. Συνδέεται με το ζεύγος O: (T, A), όπου T είναι μια ιεραρχία των στοιχείων και των υποστοιχείων του αντικειμένου O. A είναι ένα σύνολο χαρακτηριστικών του αντικειμένου O. Κάθε στοιχείο έχει τα δικά του υποστοιχεία και ένα σύνολο χαρακτηριστικών.

Ωστόσο, απλά, συχνά χρησιμοποιούμε τον όρο 'χαρακτηριστικό' για ν' αναπαραστήσουμε τα στοιχεία και τα χαρακτηριστικά τους. Ένα αντικείμενο είναι επίσης ένα χαρακτηριστικό.

Μπορούμε να ορίσουμε ένα έγγραφο d, το οποίο μπορεί να είναι μια κριτική ταινίας, ένα blog, ένα μήνυμα στο φόρουμ που αξιολογεί ορισμένα αντικείμενα. Ένα έγγραφο d αποτελείται από ορισμένες προτάσεις, έτσι ώστε d = {s1, s2, s3, s4...}.

Κάτοχος Γνώμης (Opinion Holder): Πρόκειται για ένα άτομο ή έναν οργανισμός που δημοσιεύει τη γνώμη σχετικά με ένα αντικείμενο. Για παράδειγμα, δημιουργός ενός μηνύματος σε φόρουμ, σε blog κ.α.

Σημασιολογικός Προσανατολισμός της Γνώμης: Ο σημασιολογικός προσανατολισμός της γνώμης σχετικά με μια άποψη σημαίνει ότι μπορεί να είναι θετική, αρνητική ή ουδέτερη



Εικόνα 3: Βασικά στοιχεία γνώμης[31]

[31] Liao, X., Cao, D., Tan, S., Liu, Y., Ding, G., and Cheng X.Combining Language Model with Sentiment Analysis for Opinion Retrieval of Blog-Post. Online Proceedings of Text Retrieval Conference (TREC) 2006. http://trec.nist.gov/

2.4 Επίπεδα Ανάλυσης Συναισθήματος

Οι προσεγγίσεις του προβλήματος της Ανάλυσης Συναισθήματος, διαφοροποιούνται ως προς το επίπεδο ανάλυσης.

2.4.1 Επίπεδο κειμένου

Η εξόρυξη γνώμης σε επίπεδο κειμένου (Document-Level Opinion Mining) αρχικά επικεντρώνεται στη θεματολογία και το αντικείμενο του κείμενου, δηλαδή αν πρόκειται για γνώμη ή όχι. Στη συνέχεια γίνεται μια κατηγοριοποίηση συναισθήματος με βάση την υποκειμενικότητα του συγγραφέα, θετική, αρνητική, ουδέτερη. Αυτή η μέθοδος μας η βοηθάει στην πρόβλεψη χρησιμότητας μιας άποψης, όμως δεν είναι κατάλληλη εάν το κείμενο περιέχει διάφορες απόψεις σχετικά με διαφορετικά θέματα. [32]

2.4.2 Επίπεδο Πρότασης

Η ανάλυση του συναισθήματος σε επίπεδο πρότασης είναι η πιο λεπτομερής ανάλυση του εγγράφου. Σε αυτό, η πολικότητα υπολογίζεται για κάθε πρόταση, καθώς κάθε πρόταση θεωρείται ξεχωριστή μονάδα και κάθε πρόταση μπορεί να έχει διαφορετική άποψη. Για το λόγο αυτό, η ανάλυση σε επίπεδο πρότασης αναφέρεται και ως υποκειμενική κατηγοριοποίηση (subjectivity classification).

Μια πρόταση μπορεί να είναι είτε υποκειμενική είτε αντικειμενική. Η αντικειμενική πρόταση περιέχει γεγονότα. Δεν υπάρχει άποψη σχετικά με το αντικείμενο. Για παράδειγμα, έχουμε την πρόταση «Ο τουρισμός αποτελεί τη βαριά βιομηχανία της ελληνικής οικονομίας. Η Ελλάδα είναι ένα πανέμορφο μέρος να επισκεφτεί κανείς.». Η πρώτη πρόταση είναι αντικειμενική και δεν εκφράζει κάποια άποψη/συναίσθημα σχετικά με την Ελλάδα. Επομένως, αυτό δεν πρέπει να παίξει κανένα ρόλο στη λήψη αποφάσεων σχετικά με την πολικότητα της άποψης και πρέπει να φιλτραριστεί. Η πρόταση μπορεί να

[32] Peter D. Turney, (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424

ταξινομηθεί ως θετική, αρνητική ή ουδέτερη, ανάλογα με τις λέξεις της γνώμης που υπάρχουν σε αυτήν.[33]

2.4.3 Επίπεδο Χαρακτηριστικών

Η ανάλυση των επιμέρους χαρακτηριστικών είναι ένα εξαιρετικό εργαλείο για την ανάλυση συναισθήματος [34]. Το βασικό βήμα στην ανάλυση συναισθημάτων σε επίπεδο χαρακτηριστικών είναι να προσδιοριστεί το κομμάτι του κειμένου ως χαρακτηριστικό κάποιου προϊόντος. Για παράδειγμα, «Η διάρκεια ζωής της μπαταρίας είναι πολύ μεγάλη.» Σε αυτήν την άποψη, η 'μπαταρία' είναι το χαρακτηριστικό του προϊόντος(ουσιαστικό) και το «πολύ μεγάλη διάρκεια ζωής» αποτελεί την άποψη που προσδιορίζει το αντικείμενο. Άρα ουσιαστικά μιλάμε για μια πρόταση, η οποία αποτελείται από δύο μέρη. Την οντότητα και το συναίσθημα που εκφράζεται για αυτήν.

2.5 Εφαρμογές Ανάλυσης Συναισθήματος

Κατά τα τελευταία έτη, τα ποσά των δεδομένων που παράγονται από τις υπηρεσίες Διαδικτύου έχουν αυξηθεί σημαντικά. Οι καινοτομίες στον τομέα της Πληροφορικής και των Τηλεπικοινωνιών επέτρεψαν νέες επιχειρηματικές ευκαιρίες για τη δημιουργία υπηρεσιών ικανών να χειρίζονται μεγάλους όγκους δεδομένων. Η τεχνολογία έφτασε στο επίπεδο όπου οι άνθρωποι μπορούν και συνδέονται καθημερινά με τα κοινωνικά μέσα και μοιράζονται τη ζωή τους μέσω των κοινωνικών δικτύων.

Η κοινωνική δικτύωση μέσω του Διαδικτύου έχει γίνει δημοφιλής τα τελευταία χρόνια, γεγονός που δικαιολογείται και από τους αυξημένους όγκους δεδομένων. Παρουσιάστηκαν νέες προκλήσεις σε σχέση με τις αρχιτεκτονικές αποθήκευσης δεδομένων με χαρακτηριστικά κλιμάκωσης και

[33] V. S. Jagtap, Karishma Pawar, (2013), Analysis of different approaches to Sentence-Level Sentiment Classification, International Journal of Scientific Engineering and Technology, PP : 164-170
[34] S. ChandraKala and C. Sindhu, (2012), Opinion Mining And Sentiment Classification: A Survey, ICTACT Journal on Soft Computing, Vol- 03, ISSUE: 01, ISSN: 2229-6956

αποτελεσματικούς αλγόριθμους επεξεργασίας. Η ανάλυση εξόρυξης δεδομένων έχει μεγάλες δυνατότητες για την εξεύρεση ουσιαστικών στοιχείων μέσα στα δεδομένα των κοινωνικών δικτύων. Το κοινωνικό δίκτυο του Twitter είναι μια υπηρεσία που αναπτύσσεται προκειμένου να καταστεί δυνατή η επικοινωνία μεταξύ των ανθρώπων στέλνοντας σύντομα μηνύματα.[35]

Σύμφωνα με έρευνα[36], το Twitter ως η δεύτερη μεγαλύτερη πλατφόρμα κοινωνικών μέσων, ακριβώς πίσω από το Facebook, δημιουργεί περίπου 350.000 tweets κάθε λεπτό ή 21 εκατομμύρια ανά ώρα. Αυτοί οι όγκοι δεδομένων παρουσιάζουν προκλήσεις για τους μηχανικούς να αναπτύξουν καινοτόμες λύσεις για αποτελεσματική αρχιτεκτονική δεδομένων και δυνατότητες επεξεργασίας για την εφαρμογή της εξόρυξης δεδομένων. Η σημασία της υλοποίησης μεγάλων δεδομένων σε επιχειρήσεις σε διάφορους τομείς, όπως η βιομηχανία της υγείας, το λιανικό εμπόριο, τα τηλεπικοινωνιακά δίκτυα ή τα κοινωνικά δίκτυα, διαδραματίζει κρίσιμο σενάριο για τη βελτιστοποίηση των επιχειρηματικών διαδικασιών και τη δημιουργία νέων προτάσεων αξίας για τις ροές εσόδων.

Το Twitter έχει μεγάλη επίδραση στην εξόρυξη δεδομένων, καθώς οι χρήστες παράγουν Big Data που μπορεί να επεξεργαστεί. Επιπλέον, υπάρχουν απαιτήσεις αρχιτεκτονική ανάπτυξης που μπορεί να εξελίξει για συνεχή νέα-ροή tweets και επίσης δυνατότητα να ενσωματώσει με προχωρημένους αλγόριθμους μηχανικής μάθησης. Γνωρίζοντας τι σκέφτονται οι χρήστες ή πώς αισθάνονται για προϊόντα είναι πολύτιμη πρόταση για τις εταιρείες.

Η σημασιολογική ανάλυση είναι μέρος της εξόρυξης δεδομένων, η οποία παρακολουθεί τις αντιλήψεις του κοινού σχετικά με διάφορα θέματα. Αυτό μπορεί να αναλύσει τι σκέφτονται οι άνθρωποι για τα επιχειρηματικά προϊόντα και την ποιότητά τους, τα επώνυμα προϊόντα, τις στρατηγικές τιμολόγησης ή παγκόσμιες τάσεις. Επιπλέον, μπορεί να προσδιορίσει τις επιχειρηματικές

[35] "New user FAQs," Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/13920?lang=en. [Accessed: 14-Feb-2016].
[36] "How Much Data Is Generated Every Minute On Social Media?," WeRSM | We Are Social Media. [Online]. Available: http://wersm.com/how-much-data-is-generatedevery-minute-on-social-media/. [Accessed: 14-Feb-2016].

ευκαιρίες και έτσι να γίνει ένας αποτελεσματικός παράγοντας για τις εταιρείες ώστε να καινοτομούν τις υπηρεσίες τους.[37]

Το Twitter ως πλατφόρμα που υποστηρίζεται από τους ενεργούς χρήστες δημιουργεί ευκαιρίες για εξόρυξη δεδομένων και πιο συγκεκριμένες σημασιολογικές αναλύσεις με βάση τα tweets. Οι χρήστες Twitter συχνά εκφράζουν τις απόψεις τους σχετικά με διάφορα θέματα μέσα στα δημοσιευμένα tweets τους. Και συνεπώς, εφαρμόζοντας τεχνική επεξεργασία κειμένων, η τεχνική εξόρυξης δεδομένων μπορεί να εξυπηρετήσει τις εταιρείες μέσω ανατροφοδότηση για την καλύτερη διαχείριση της επωνυμίας της εταιρίας.

Από την άλλη πλευρά, δεδομένου ότι το Twitter παράγει τεράστιους όγκους δεδομένων κάθε μέρα, οι σημασιολογικές αναλύσεις μπορούν να βοηθήσουν στις εκστρατείες που σχετίζονται με το μάρκετινγκ, για να διερευνήσουν τις απόψεις του κοινού σχετικά με το νεοεμφανιζόμενο προϊόν. Για παράδειγμα για μια κινηματογραφική ταινία να αναλύσουν το συναίσθημα για την ικανοποίηση των χρηστών.

2.5.1 Twitter

Το Twitter επιτρέπει στους χρήστες να επικοινωνούν με σύντομα μηνύματα που ονομάζονται tweets. Κάθε tweet μπορεί να περιέχει 140 χαρακτήρες το πολύ. Σύμφωνα με την ιστοσελίδα του Twitter [25], 140 χαρακτήρες παρουσιάζουν το τέλειο μήκος για την αποστολή ενημερώσεων κατάστασης μέσω μηνυμάτων κειμένου. Επιπλέον, 20 άλλοι χαρακτήρες προορίζονται για τα ονόματα των ανθρώπων. Μόλις οι χρήστες συνδεθούν σε υπηρεσία και εγγραφούν για δωρεάν λογαριασμό, τα μέλη μπορούν να στείλουν tweets ή να ακολουθήσουν άλλα μέλη για να ενημερωθούν για τα τελευταία νέα. Αυτά τα σύντομα μηνύματα δημοσιεύονται στο προφίλ των χρηστών. Επιπλέον, μπορούν να σταλούν στους οπαδούς και να μπορούν να αναζητηθούν στο Twitter. [25]

[37] B. Pang and Lillian Lee, "Opinion mining and sentiment analysis," [Online]. Available: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf. [Accessed: 21-Apr- 2016]. [Accessed: 15-Feb-2016].

Η υπηρεσία κοινωνικής δικτύωσης δεν περιορίζεται μόνο στην πρόσβαση στον ιστότοπο, αλλά οι χρήστες μπορούν να εμπλέκονται και ν' αλληλεπιδρούν μέσω εφαρμογών που έχουν αναπτυχθεί για έξυπνες συσκευές(smartphones, tablets). Στην πραγματικότητα, σύμφωνα με τα στοιχεία χρήσης του Twitter [38], υπάρχει το 80% των ενεργών χρηστών στο κινητό. Η δημοτικότητα σε σχέση με τα micro blogging κέρδισε επιτυχία σε ολόκληρο τον κόσμο, γεγονός που υποστηρίζεται από το ότι το Twitter έχει περίπου 320 εκατομμύρια ενεργούς χρήστες που ασχολούνται με την υπηρεσία σε μηνιαία βάση.

Το περιεχόμενο που δημιουργείται από το χρήστη μπορεί να δημιουργήσει πολλές ευκαιρίες για μάρκετινγκ και διαφήμιση, στις οποίες χρησιμοποιούνται οι τεχνικές εξόρυξης δεδομένων. Το Twitter είναι χρήσιμο για την ανάγνωση και εύρεση ενδιαφερόντων θεμάτων που προσελκύουν την προσοχή του χρήστη. Οι άνθρωποι μπορούν να ανακαλύψουν νέα σε πραγματικό χρόνο σχετικά με το τι συμβαίνει στον κόσμο ή να μένουν σε επαφή με φίλους.

Από την άλλη πλευρά, πολλές εταιρείες χρησιμοποιούν το Twitter για να ενημερώσουν τους πελάτες σχετικά με τις προσφορές τους. Το περιεχόμενο των tweets σχετίζεται με δύο επιπλέον μεταδεδομένα που διακρίνονται σε οντότητες και μέρη. Οι οντότητες Tweet είναι αναφορές χρηστών, οι οποίες αντιπροσωπεύουν τον τρόπο αναφοράς των άλλων χρηστών στα δικά τους tweets, συμπεριλαμβάνοντας το @ σύμβολο, ακολουθούμενο από το όνομα χρήστη τους. Επιπλέον, οι οντότητες tweet ενδέχεται να περιέχουν επίσης hashtags και διευθύνσεις URL. Αντίθετα, οι θέσεις tweet αντιπροσωπεύουν τοποθεσίες πραγματικού κόσμου που μπορούν να ενσωματωθούν σε ένα tweet. [39]

Η ορολογία είναι ένα σημαντικό μέρος του Twitter, διότι διδάσκει στους χρήστες τη λειτουργικότητα και τις λειτουργίες της υπηρεσίας. Επιπλέον, ορίζει τις πτυχές του Twitter και διάφορες δυνατότητες για τη χρήση του. Για να κατανοήσουμε την ορολογία του Twitter, παρουσιάζεται μια σύντομη επισκόπηση. Στο Twitter χρησιμοποιείται το (@) για να καλέσετε κάποιον

[38] "Company | About," Twitter About. [Online]. Available: https://about.twitter.com/company. [Accessed: 17-Feb-2016].
[39] M. A. Russell, "Mining the Social Web, Second Edition," Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472 in 2014, ISBN: 978-1-449-36761-9.

χρήστη στο tweet ή για να στείλετε ένα μήνυμα στο χρήστη. Επιπλέον, αυτό το σύμβολο χρησιμοποιείται όποτε ο χρήστης θέλει να δημιουργήσει σύνδεση με άλλους χρήστες και να συνδεθεί με το Twitter προφίλ του. Το όνομα χρήστη προσδιορίζει με μοναδικό τρόπο κάθε χρήστη και γενικά χρησιμοποιείται με το σύμβολο @, για παράδειγμα, ο Andy Murray είναι @andy_murray. [40]

Ένα άλλο δημοφιλές σύμβολο που χρησιμοποιείται στο Twitter ονομάζεται hashtag (#). Βοηθά τους χρήστες να κατηγοριοποιούν τα μηνύματα. Στην πραγματικότητα, η δομή του έρχεται με το (#) σύμβολο ακολουθούμενο από τη σχετική λέξη-κλειδί σε σχέση με το μήνυμα tweet. Ουσιαστικά, βοηθά στην κατηγοριοποίηση των tweets με βάση το περιεχόμενό τους και επιτρέπει καλύτερα αποτελέσματα αναζήτησης από το Twitter Search. [41] Hashtags μπορεί να βρίσκεται οπουδήποτε μέσα στο tweet. Όταν οι χρήστες κάνουν κλικ σε αυτό, θα οδηγηθούν σε κατηγορία που ομαδοποιεί όλα τα tweets από τους χρήστες Twitter στο ίδιο θέμα.

Το twitter επίσης υποστηρίζει αποστολή άμεσων προσωπικών μηνυμάτων μεταξύ των χρηστών, ως εκ τούτου αναπτύσσουν βασικές λειτουργίες ανταλλαγής μηνυμάτων, αλλά με προστιθέμενη αξία της micro blogging υπηρεσία(εμπεριέχονται video, posts, links). Το γεγονός αυτό είναι το κύριο πλεονέκτημα σε σύγκριση με τις απλές υπηρεσίες ανταλλαγής, όπως το WhatsApp ή το Viber.

Η δυνατότητα να εγγραφούν οι χρήστες σε ένα διαφορετικό λογαριασμό Twitter είναι γνωστό υπό όρο «following». Όταν ο χρήστης αποφασίσει να ακολουθήσει άλλο λογαριασμό, το Twitter θα ενημερώσει το προφίλ του ακόλουθου με τα πιο πρόσφατα tweets. Ο χρήστης γίνεται στους όρους του Twitter «follower». Αριθμός των followers εμφανίζεται στην επισκόπηση προφίλ χρήστη.

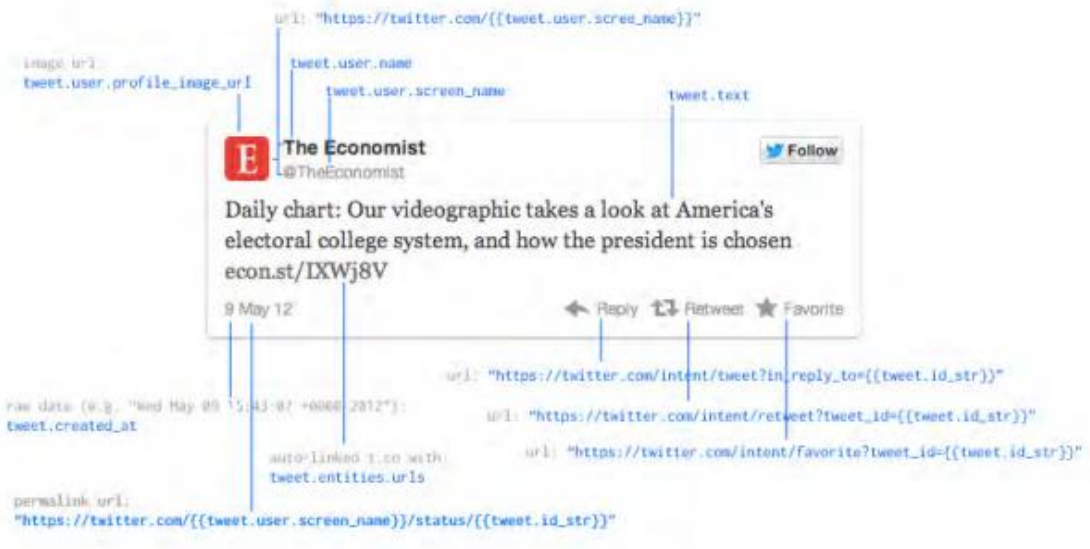
Το Geotag χρησιμοποιείται για να ενημερώσει τους χρήστες σχετικά με την τοποθεσία όπου βρισκόταν ο δημιουργός, όταν καταχώρησε το tweet.

[40] "Twitter", Andy Murray. [Online]. Available: https://twitter.com/andy_murray/with_replies. [Accessed: 25-Feb-2016]
[41] "Using hashtags on Twitter," Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/49309. [Accessed: 23-Feb-2016].

Κουμπί «Like» δείχνει την θετική αντίδραση προς το περιεχόμενο ενός tweet. Το προφίλ χρήστη επιτρέπει στους χρήστες να δημιουργήσουν ειδικές λίστες που θα μπορεί να ομαδοποιεί όλα τα θέματα σε κατηγορίες. Έτσι θα μπορεί να εξασφαλίζει καλύτερη πλοήγηση και συνοχή μεταξύ των θεμάτων ενδιαφέροντος του.

Προτεινόμενοι αλγόριθμοι μπορούν να προτείνουν στους χρήστες «Ποιούς να ακολουθούν». Η πληροφορία βρίσκεται στη σελίδα του προφίλ και συνήθως αποτελείται από προωθούμενους Twitter λογαριασμούς, τάσεις και tweets από διαφήμιση.

Η επιλογή της απάντησης έρχεται με κάθε καταχωρημένο tweet. Δίνει τη δυνατότητα σε κάποιο χρήστη να αντιδρά και να στείλει απάντηση. Επιπλέον, η επιλογή retweet είναι μέρος κάθε μηνύματος. Έχει επίσης τη λειτουργικότητα να μοιράζονται tweets με δικούς του «followers». Αν το tweet αποτελείται από περισσότερα από ένα σύμβολα @, λέγεται «αναφορά».



Εικόνα 4: Δομή ενός tweet[42]

Παρατηρώντας όλα τα «μέρη» ενός tweet, εμείς θα επικεντρώσουμε κατά κύριο λόγο στο tweet.text, που αποτελεί και την κύρια πηγή πληροφοριών για εμάς.

[42] "The Twitter glossary," Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/166337. [Accessed: 20-Feb-2016].

2.5.2 Twitter και Πολιτική

Ιδιαίτερο ενδιαφέρον όμως έχει να αναφερθούμε στην πλέον διαδεδομένη χρήση του twitter από πολιτικούς, ιδίως σε προεκλογικές περιόδους. Οι πολιτικοί επιδιώκουν να μοιραστούν τη δημόσια εικόνα τους, αλλά και να προβάλλουν πτυχές της προσωπικής τους ζωής, προκειμένου να δείξουν πιο 'προσιτοί' στους απλούς ψηφοφόρους.[43]

Επίσης πολύ σημαντικό ρόλο παίζει το γεγονός ότι ένας χρήστης του twitter μπορεί να αποκτήσει πληθώρα από ακόλουθους (followers). Αυτό δίνει τη δυνατότητα σε χρήστες του twitter που έχουν δημόσια πολιτική ζωή να δημιουργούν εμμέσως ψηφοφόρους, αλλά και να «δοκιμάζουν» την επιρροή τους σε αυτούς.

Σύμφωνα με αμερικάνικες μελέτες, έχει αποδειχτεί πώς τα πρόσωπα που διεκδικούν να έχουν μια δημόσια εικόνα στο πολιτικό σκηνικό και επιδιώκουν ν' ασχοληθούν με τα κοινά έχουν περισσότερες πιθανότητες να υιοθετήσουν τη χρήση των μέσων κοινωνικής δικτύωσης.[44] Μπορούμε να πούμε ότι υπάρχουν τρία βασικά κίνητρα για τη χρήση του twitter για πολιτικούς λόγους.

Κατ' αρχήν, για λόγους μάρκετινγκ, δηλαδή οι υποψήφιοι χρησιμοποιούν το twitter για να αυξήσουν την προβολή του προφίλ τους, αλλά και του κόμματός τους ενδεχομένως. Το πιο σημαντικό όμως είναι ότι μπορούν μέσα από το twitter να έρθουν σε επαφή με κοινωνικές και ηλικιακές κατηγορίες με τις οποίες θα ήταν δύσκολο με άλλο τρόπο να επικοινωνήσουν και να μεταφέρουν τις απόψεις τους(πχ σε νέους 18-24). [45]

Επιπλέον, το twitter βοηθά στην αποτύπωση της δημόσιας εικόνας του υποψηφίου. Αφενός μπορεί ο υποψήφιος ν' από τυπώνει ανά πάσα στιγμή τις

[43] Enli, G. S. & Thumin, N. (2012) 'Socializing and self-represantation online: exploring Facebook', Observatorio (OBS) Journal.
[44] Enli, G. S. & Skogerbo, E. (2013). Personalized Campaigns In Party-Centred Politics. Information, Communication & Society.
[45] Karlsen, R. (2010b) 'Fear of the political consultant', Party Politics, vol.16, pp. 193-214

παρεμβάσεις του, αφετέρου μπορεί να καλεί και μέσω του twitter υποψήφιους ψηφοφόρους του σε κοινωνικές και πολιτικές εκδηλώσεις. [46]

Τέλος, δίνεται η ευκαιρία στους υποψηφίους για άμεση επικοινωνία με τους ψηφοφόρους του μέσω διαλόγου. Μπορούν λοιπόν μέσα από μια αλληλεπιδραστική επαφή να συλλέξουν πληροφορίες με βάση τις επιθυμίες των ψηφοφόρων και να προσαρμόσουν ενδεχομένως τη ρητορική τους, την καμπάνια τους και τις θέσεις τους.

Στο Κεφάλαιο 3 κάνουμε μια δήλωση του προβλήματος που έχουμε, αναφερόμαστε πιο συγκεκριμένα στον τρόπο με τον οποίο θα κάνουμε μια κατηγοριοποίηση στα tweets και παρουσιάζουμε εκτενώς τους αλγόριθμους με βάση τους οποίους θα φτιάξουμε το μοντέλο πρόβλεψης που θέλουμε.

[46] Nielsen, R.K. (2011) 'Mundaneinternettools, mobilizingpractices, and thecoproduction of citizenship in political campaigns', NewMedia&Society, vol. 13, pp 755-771

3. Διατύπωση και Προσέγγιση του Προβλήματος

3.1 Διατύπωση του Προβλήματος

Ο σκοπός αυτής της εργασίας είναι να επικεντρώσει στην εξόρυξη πληροφορίας από κείμενο tweet με συγκεκριμένους αλγόριθμους, που αναλύουμε πιο κάτω προκειμένου να ανακαλύψουμε πληροφορία μέσα από το περιεχόμενο των tweets και να βγάλουμε χρήσιμα συμπεράσματα από τα αποτελέσματα που προκύπτουν.[47] Έχουμε επικεντρώσει στις αμερικανικές εκλογές και στην επίδραση των tweets στους χρήστες σχετικά με τους δύο υποψηφίους, Donald Trump και τη Hilary Clinton.

Έχουμε λοιπόν ένα σύνολο δεδομένων από tweets χρηστών τα οποία αναφέρονται είτε στον Trump είτε στην Clinton. Αφού φορτώσουμε το σύνολο δεδομένων, προχωράμε στην απαραίτητη γλωσσολογική προεπεξεργασία των κειμένων των tweets. Στη συνέχεια, με βάση συγκεκριμένες λέξεις με θετική και αρνητική γνώμη αντίστοιχα, κάνουμε μια κατηγοριοποίηση συναισθήματος στα tweets σε θετικά, αρνητικά και ουδέτερα. Παρατηρούμε επίσης δημοφιλείς λέξεις για κάθε υποψήφιο και βλέπουμε την επίδραση αυτών των λέξεων στα tweets(πόσες φορές εμφανίζονται πχ). Έπειτα, δημιουργούμε ένα μηχανισμό πρόβλεψης με τη χρήση συγκεκριμένων αλγόριθμων. Δηλαδή, εκπαιδεύοντας το συγκεκριμένο dataset, μπορούμε να προβλέψουμε ποσοστιαία για ένα νέο tweets, αν ταιριάζει περισσότερο(και πόσο) στο @realDonaldTrump ή @HillaryClinton και αντίστοιχα αν έχει θετική, αρνητική ή ουδέτερη γνώμη. Τέλος, μετράμε την απόδοση των αλγόριθμων για όλα τα υποσύνολα δεδομένων και ελέγχου που «τρέχουμε».

Επιπρόσθετα, Θα χρησιμοποιήσουμε το εργαλείο **Orange3** της Python, στο οποίο θα κάνουμε μια γενική παρουσίαση και στη συνέχεια θα κάνουμε μια υλοποίηση επεξεργασίας κειμένου. Θα ακολουθήσουμε μια αντίστοιχη διαδικασία προεπεξεργασίας και ανάλυσης και θα προσπαθήσουμε να συγκρίνουμε τα αποτελέσματα της απόδοσης των αλγόριθμων με την

[47] Janardhana, Ravikiran. "How to Build a Twitter Sentiment Analyzer." Ravikiranj.net.

αντίστοιχη απόδοσή τους στην προηγούμενη υλοποίηση. Με αυτόν τον τρόπο θα μπορέσουμε να αξιολογήσουμε τις δυνατότητες, αλλά και την ευχρηστία του εργαλείου αυτού.

3.2 Περιγραφή Αλγόριθμου

- Εισαγωγή και φόρτωση του dataset
- Προεπεξεργασία δεδομένων: Φιλτράρουμε τα tweets προκειμένου να μετρήσουμε τη συχνότητα εμφάνισης των λέξεων και να υπολογίσουμε το «βάρος» κάθε λέξης, αλλά και να μετατρέψουμε τις λέξεις σε λήμματα.
- Κάνουμε κατηγοριοποίηση συναισθήματος(sentiment classification) για τα tweets σε θετικά, αρνητικά και ουδέτερα.
- Δημιουργούμε μοντέλο εκπαίδευσης και ελέγχου.
- Εντοπισμός λέξεων: Εντοπίζουμε δημοφιλείς λέξεις μέσα στα tweets και τις καταγράφουμε.
- Υπολογισμός απόδοσης αλγορίθμων και ταξινόμηση: Κάθε νέο tweet, που δημιουργείται ταξινομείται σε έναν από τους δύο υποψηφίους με ένα συγκεκριμένο ποσοστό επιτυχίας και αξιολογείται ποσοστιαία θετικά, αρνητικά ή ουδέτερα.
- Σύγκριση αλγόριθμων

3.3 Ταξινόμηση των tweets – Αλγόριθμοι

Η ταξινόμηση κειμένου έγινε ένα πεδίο σημαντικής μελέτης επέκτασης των μέσων ενημέρωσης, όπως τα κοινωνικά δίκτυα. Προκλήσεις ως προς την ταξινόμηση συνδέονται με ταξινομητές και τις δυνατότητες απόδοσής τους, που εξαρτώνται από το συγκεκριμένο μοντέλο.[48] Η ταξινόμηση απαιτείται στις μέρες μας σε διάφορες περιπτώσεις χρήσης, όπως η ανίχνευση spam, η ανάλυση συναισθήματος, προβλέψεις σε σχεδόν πραγματικό χρόνο.

[48] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the ACM SIGIR 2010 Posters and Demos. ACM press

3.3.1 Naïve Bayes

Ο Naïve Bayes αλγόριθμος βασίζεται στο θεώρημα πιθανοτήτων Bayes[49], που στόχο έχει να προβλέψει αποτελέσματα από μη-επισημασμένα δεδομένα. Στο θεώρημα Bayes, η μέθοδος ταξινόμησης προϋποθέτει την ανεξαρτησία μεταξύ των χαρακτηριστικών(features) και της κλάσης κατηγοριοποίησης(data).

Τα Naïve Bayes μοντέλα χωρίζονται σε διάφορους τύπους ανάλογα με το χειρισμό των χαρακτηριστικών. Στο Bernoulli μοντέλο, οι τιμές των χαρακτηριστικών πρέπει να είναι δυαδικές(0-1, T-F κ.α). Επιπλέον, σε αυτό το μοντέλο δε μας ενδιαφέρει η συχνότητα εμφάνισης κάθε λέξης μες στο κείμενο. Από τη άλλη, στο πολυωνυμικό (multinomial) μοντέλο μας ενδιαφέρει η συχνότητα των λέξεων μες στο κείμενο.

Ο Naïve Bayes αλγόριθμος ταξινόμησης είναι χρήσιμος για να χαρακτηρίσει ακόμα και σύνολα δεδομένων με υψηλό όγκο πληροφοριών, καθώς εκτελείται αποτελεσματικά και είναι εύκολο να εφαρμοστεί.

$$P(c|X) = [P(X|c) * P(c)] / P(X)$$

- P(c|x) = εκ των υστέρων πιθανότητα
- P(x|c) = Δεσμευμένη πιθανότητα
- P(c) = εκ των προτέρων πιθανότητα κλάσης
- P(X) = εκ των προτέρων πιθανότητα ταξινομητή

$$P\ (c \mid X) = P\ (x1 \mid c)\ x\ P\ (x2 \mid c)\ x\ \ldots\ x\ P\ (xn \mid c)\ x\ P\ (c)$$

Η εκ των υστέρων πιθανότητα υπολογίζει την πιθανότητα του αποτελέσματος που προκύπτει από μια νέα πληροφορία. Στο P(c|x), το c αναπαριστά την κλάση που ταξινομούνται τα δεδομένα και το x τον ταξινομητή. Η δεσμευμένη πιθανότητα είναι η πιθανότητα να βρίσκεται ο ταξινομητής(χαρακτηριστικό) μέσα στην κλάση.

[49] Heckerman, D.(1999). Learning in graphical models, chapter A tutorial on learning with Bayesian networks, pages 301–354. MIT Press

Βάσει λοιπόν των πιο πάνω, μπορούμε να υπολογίσουμε τη μέγιστη a Posteriori πιθανότητα P(X) των δεδομένων.

$$c_{MAP} = \arg\max_{c \in C} \quad P(c \mid X)$$

$$= \arg\max_{c \in C} \quad \frac{P(X|c)*P(c)}{P(X)}$$ P(X) σταθερό => Μπορεί να παραλειφθεί

$$= \arg\max_{c \in C} \quad P(X \mid c) * P(c)$$

Υποθέτοντας ότι όλες οι πιθανότητες c είναι το ίδιο πιθανές, δηλαδή $P(c_i) = P(c_j)$, μπορούμε να παραλείψουμε το και το P(c). Άρα:

$$c_{MAP} = \arg\max_{c \in C} \quad P(X \mid c)$$

Η πιο πάνω υπόθεση ονομάζεται **Maximum Likelihood Hypothesis**.[50]

3.3.1.1 Multinomial Naïve Bayes

Τα έγγραφα του training set περιγράφονται από συνδέσεις των τιμών των χαρακτηριστικών τους, δηλαδή των λέξεων που τα αποτελούν, δηλαδή παίρνουν διακριτές τιμές, με πιθανότητες εμφάνισης στην κατηγορία C, $\{P_{c1}, P_{c2}, ..., P_{cn}\}$, τότε υποθέτουμε πολυωνυμική κατανομή. Συνεπώς, υπάρχει ένα πολυώνυμο πιθανοτήτων $\{P_1, P_2, ..., P_n\}$ για κάθε κατηγορία, που εκφράζει την εμφάνιση των γεγονότων {1,2,…,n} σε αυτή την κατηγορία, με συχνότητες $\{X_1, X_2, ..., X_n\}$. Με άλλα λόγια $D = \{x_1, x_2, ..., x_n\}$. Θέλουμε να το αναθέσουμε σε μια κλάση $c_j \in C$.

$$c_{MAP} = \arg\max_{c \in C} P(c_j \mid x_1, x_2, ..., x_n)$$

[50] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment classification using Machine learning Techniques". In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp. 79–86

$$= \arg\max_{c \in C} \frac{P(x_1,x_2,...,x_n|c_j)*P(c_j)}{P(x_1,x_2,...,x_n)}$$

$$= \arg\max_{c \in C} P(x_1, x_2, ..., x_n \mid c_j) * P(c_j)$$

Λόγω του ότι είναι δύσκολο να υπολογιστεί το $P(x_1, x_2, ..., x_n \mid c_j)$ θα κάνουμε ακόμη μια υπόθεση ότι τα χαρακτηριστικά είναι στατιστικώς ανεξάρτητα, δηλαδή αν γνωρίζουμε την τιμή ενός χαρακτηριστικού, δεν μπορούμε να πούμε τίποτα για την τιμή κάποιου άλλου χαρακτηριστικού. Αυτό λέγεται **Υπόθεση υπό συνθήκης ανεξαρτησίας**. Έτσι:

$$P(x_1, x_2, ..., x_n \mid c_j) = P(x_1 \mid c_j) * P(x_2 \mid c_j) * ... * P(x_n \mid c_j) = \prod_i^n P(x_i \mid c_j)$$

Το αποτέλεσμα είναι ο Naïve Bayes Classifier:

$$c_{NB} = \arg\max_{c \in C} P(c_j) \prod_i^n P(x_i \mid c_j)$$

Για να αποφύγουμε να έχουμε αποτέλεσμα 0 σε περιπτώσεις όπου κάποια κλάση δεν περιέχει κάποιο χαρακτηριστικό το οποίο περιέχεται στο έγγραφο που προσπαθούμε να ταξινομήσουμε, δηλαδή $P(x_i \mid c_j) = 0$, θα ορίσουμε το $P(x_i \mid c_j)$ ως

$$P(x_i \mid c_j) = \frac{n_c + 1}{n + k}$$

Όπου, k: αριθμός των τιμών που μπορεί να πάρει το x_i.

Στη συνέχεια της εργασίας, για την κατηγοριοποίηση συναισθήματος, θα χρησιμοποιηθεί το Multinomial Naive Bayes μοντέλο.

3.3.1.2 Gaussian Naïve Bayes[51]

Για τον υπολογισμό των υπό συνθήκη πιθανοτήτων $P(x_i \mid c)$ θα πρέπει να γνωρίζουμε την κατανομή των πιθανοτήτων των X_i. Αν τα χαρακτηριστικά των

[51] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment classification using Machine learning Techniques". In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp. 79-86

στιγμιοτύπων παίρνουν συνεχείς τιμές, τότε υποθέτουμε κανονική κατανομή για τις πιθανότητες των Xi δεδομένης της κατηγορίας C, $P(x_i \mid c)$. Κατά την εκπαίδευση, αρχικά επιλέγουμε για την κατηγορία C τα αντίστοιχα στιγμιότυπα που έχουν ταξινομηθεί χειροκίνητα σε αυτή. Εν συνεχεία, από το σύνολο αυτό, για κάθε χαρακτηριστικό, υπολογίζουμε το μέσο όρο των τιμών του, μC και τη διασπορά σ_c^2 και βάσει αυτών προσδιορίζουμε την κανονική κατανομή του χαρακτηριστικού για την κατηγορία C. Στο τέλος, για κάθε χαρακτηριστικό και για κάθε κατηγορία, έχουμε μία κανονική κατανομή, η οποία καθορίζει την κατανομή του χαρακτηριστικού για την κατηγορία αυτή. Η πιθανοφάνεια των χαρακτηριστικών κατά τον έλεγχο υπολογίζεται ως εξής:

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

3.3.2 Support Vector Machines (SVM)

Πρόκειται για μια δυαδική μηχανή μάθησης. Η κεντρική ιδέα είναι ότι για ένα δείγμα εκπαίδευσης, η svm κατασκευάζει ενα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Η ιδέα αυτή βασίζεται στη θεωρία του πυρήνα εσωτερικού γινομένου μεταξύ ενός διανύσματος υποστήριξης Xi και ενός διανύσματος X το οποίο αντλείται από το χώρο δεδομένων εισόδου. Στη μέθοδο πυρήνα ο αλγόριθμος εξάγει ένα μικρό υποσύνολο σημείων δεδομένων από το ίδιο το δείγμα εκπαίδευσης. Η μέθοδος πυρήνα είναι βέλτιστη.[52]

Η SVM χρησιμοποιείται για:

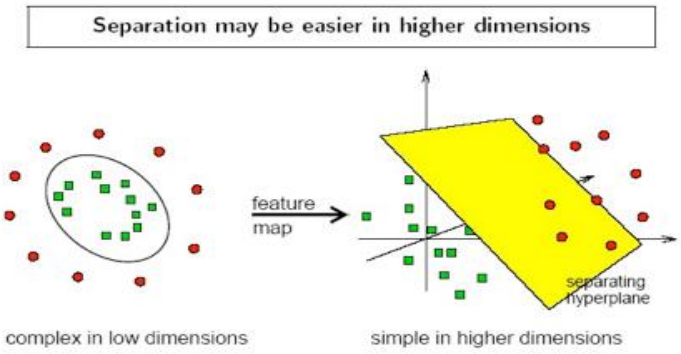
- *Ταξινόμηση προτύπων*
- *Προβλήματα μη-γραμμικής παλινδρόμησης*

[52] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

Οι ταξινομητές SVMs μπορεί να χρησιμοποιηθούν τόσο σε γραμμικά όσο και σε μη γραμμικά δεδομένα.

Μια μηχανή υποστήριξης διανυσμάτων χρησιμοποιεί μια μη γραμμική αντιστοίχιση για να μετασχηματίσει τα αρχικά δεδομένα εκπαίδευσης σε δεδομένα υψηλότερης διάστασης. Με μια κατάλληλη μη γραμμική αντιστοίχιση σε μια επαρκώς υψηλότερη διάσταση, τα δεδομένα πλέον από τις δύο κλάσεις μπορούν να διαχωριστούν από ένα υπερεπίπεδο. Με βάση λοιπόν τη νέα διάσταση, ο ταξινομητής ψάχνει για γραμμικώς διαχωριζόμενα υπερεπίπεδα (όρια απόφασης).[53]

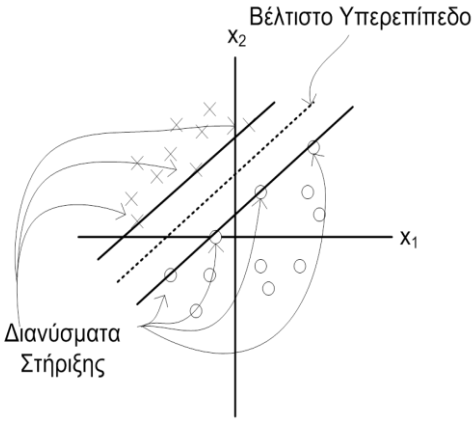
Η μέθοδος SVM βρίσκει αυτά τα υπερεπίπεδα χρησιμοποιώντας τα διανύσματα υποστήριξης (support vectors), που αποτελούν τα σημαντικότερα στιγμιότυπα εκπαίδευσης και τα περιθώρια (margins), που καθορίζονται από τα διανύσματα υποστήριξης. Διαισθητικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο το οποίο έχει τη μεγαλύτερη απόσταση από τα κοντινότερα σημεία εκπαίδευσης των κλάσεων ταξινόμησης (αποκαλούμενο περιθώριο συνάρτησης). Γενικά όσο μεγαλύτερο το περιθώριο συνάρτησης τόσο χαμηλότερο είναι το παραγόμενο λάθος από τη χρήση του ταξινομητή



Εικόνα 5: Διαχωρισμός σε υψηλότερη διάσταση[53]

Το βέλτιστο υπερεπίπεδο για γραμμικά διαχωρίσιμα πρότυπα

[53] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.



Εικόνα 6: Γραμμικά διαχωρίσιμα πρότυπα[54]

3.3.2.1 Γραμμική κατηγοριοποίηση

Η εξίσωση που περιγράφει το υπερεπίπεδο είναι:

$$w^T x_i + b \geq 0 \text{ για } d_i = +1$$

$$w^T x_i + b < 0 \text{ για } d_i = -1$$

x είναι το διάνυσμα εισόδου,w το διάνυσμα βαρών και b η πόλωση

Βέλτιστο υπερεπίπεδο: $w_0^T x + b_0 = 0$

Μπορούμε κάθε φορά να επανακλιμακώσουμε τα w_0

και τα b_0 έτσι ώστε να ισχύει:

$$w_0^T x + b_0 \geq 0 \text{ για } d_i = +1$$

$$w_0^T x + b_0 < 0 \text{ για } d_i = -1$$

Για δεδομένο διάνυσμα βαρών w και πόλωσης b, η απόσταση του πιο κοντινού σημείου στην γραμμή η οποία ορίζεται από το βέλτιστο υπερεπίπεδο ονομάζεται περιθώριο διαχωρισμού και συμβολίζεται ως ρ. Ο στόχος των SV μηχανών είναι να βρουν το υπερεπίπεδο όπου το ρ μεγιστοποιείται.[54]

Η διακρίνουσα συνάρτηση $g(x) = W_0^T x + b_0$, η οποία δίνει ένα αλγεβρικό μέτρο για το πόσο μακριά είναι το X από την γραμμή του βέλτιστου υπερεπιπέδου.

[54] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

Το X μπορεί να γραφεί ως $X = X_\rho + r * \frac{W_0}{|W_0|}$,όπου X_ρ είναι η προβολή του X στη γραμμή του βέλτιστου υπερεπιπέδου και r είναι η επιθυμητή απόσταση $r = \frac{g(x)}{\|W_0\|}$ και ισοδύναμα $r = \frac{W^T.x}{|W|} - (-\frac{b}{|W|}) = \frac{g(x)}{|W|} = \frac{1}{|W|}$

Η απόσταση της $W_0^T x + b_0 = 0$ από την αρχή των αξόνων δίνεται από τον τύπο $\frac{b_0}{\|W_0\|}$ Εάν

- $b_0 > 0$ η αρχή των αξόνων βρίσκεται στην πλευρά της κλάσης +1
- $b_0 < 0$ η αρχή των αξόνων βρίσκεται στην πλευρά της κλάσης -1
- $b_0 = 0$ το $w_0^T x + b_0$ περνά από την αρχή των αξόνων

Υποθέτουμε πως η *ρ* υποδηλώνει την βέλτιστη τιμή του περιθωρίου διαχωρισμού ανάμεσα στις δύο κλάσεις, οι οποίες αποτελούν το σύνολο εκπαίδευσης *T*, τότε ρ = 2r. Από το παραπάνω καταλαβαίνουμε πως για να μεγιστοποιήσουμε το περιθώριο διαχωρισμού μεταξύ των κλάσεων, μπορούμε ισοδύναμα να ελαχιστοποιήσουμε την ευκλείδεια νόρμα του διανύσματος βαρών W. Για την εύρεση του βέλτιστου υπερεπιπέδου χρησιμοποιούμε την τετραγωνική βελτιστοποίηση.[55]

Η δημιουργία ενός υπερεπιπέδου διαχωρισμού είναι σχεδόν αδύνατο να γίνει χωρίς την ύπαρξη σφαλμάτων ταξινόμησης. Στόχος είναι η ελαχιστοποίηση της πιθανότητας σφάλματος ταξινόμησης. Εισάγουμε λοιπόν ένα σύνολο μη-αρνητικών βαθμωτών μεταβλητών ξ, οι οποίες αποκαλούνται μεταβλητές χαλάρωσης και η σχέση που περιγράφει το βέλτιστο υπερεπίπεδο είναι:

$$d_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, ..., N$$

Μετρούν την απόκλιση ενός σημείου δεδομένων από την ιδανική συνθήκη διαχωρισιμότητας προτύπων.

- Για 0<ξ<1, το σημείο εμπίπτει στη σωστή περιοχή διαχωρισμού.

[55] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

- Για ξ>1 εμπίπτει στη λάθος.

Το βέλτιστο υπερεπίπεδο στην περίπτωση αυτή θα δίνεται από την λύση του ακόλουθου προβλήματος βελτιστοποίησης:

$$\Phi(w,\xi)=\frac{1}{2}W*W+C\sum_{i=1}^{N}\xi_i$$

Η ελαχιστοποίηση του πρώτου όρου σχετίζεται με την ελαχιστοποίηση της VC διάστασης του SVM δικτύου και άρα με τον έλεγχο της μάθησης. Όσον αφορά την ελαχιστοποίηση του δεύτερου όρου, αυτή έχει να κάνει με την ελαχιστοποίηση του αριθμού των λαθών. Η παράμετρος C προκαθορίζεται από τον χρήστη και αποτελεί έναν συμβιβασμό μεταξύ πολυπλοκότητας της μηχανής και του αριθμού των μη διαχωρίσιμων σημείων.

3.3.2.2 Μη - Γραμμική κατηγοριοποίηση

Το πρόβλημα βελτιστοποίησης για μη-διαχωρίσιμα πρότυπα περιλαμβάνει το πρόβλημα της βελτιστοποίησης για γραμμικά διαχωρίσιμα πρότυπα ως ειδική περίπτωση. Συγκεκριμένα θέτοντας ξi = 0 για όλα τα i.[56]

Η περίπτωση των μη-διαχωρίσιμων προτύπων διαφέρει από την περίπτωση των διαχωρίσιμων στο ότι ο περιορισμός αi>0 αντικαθίσταται από τον πιο αυστηρό περιορισμό 0<αi<C.

Η παράμετρος C είναι μια καθοριζόμενη από το χρήστη θετική παράμετρος, η οποία ελέγχει το συμβιβασμό μεταξύ της πολυπλοκότητας της μηχανής και του αριθμού των μη-διαχωρίσιμων σημείων

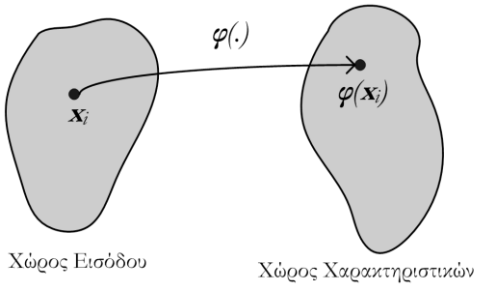
Μεγάλο C σημαίνει ότι υπάρχει μεγάλη εμπιστοσύνη στην ποιότητα του δείγματος εκπαίδευσης

Μικρό C σημαίνει ότι το δείγμα εκπαίδευσης είναι θορυβώδες.

Εν τούτοις, αν ο βέλτιστος τρόπος να διακριθούν τα δεδομένα είναι μη γραμμικά υπερεπίδεδα, τότε κρίνεται καταλληλότερη η χρήση μίας μη γραμμικής

[56] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

διανυσματικής συνάρτησης. Για το σκοπό αυτό, ο χώρος του προβλήματος μετασχηματίζεται, σε έναν άλλο χώρο μεγαλύτερης ή και άπειρης διάστασης μέσω της απεικόνισης



Εικόνα 7:Μετασχηματισμός χώρου[57]

Ο πυρήνας K(x,xi) είναι μια συνάρτηση που υπολογίζει το εσωτερικό γινόμενο των εικόνων που παράγονται από το χώρο χαρακτηριστικών βάσει του Φ δύο σημείων δεδομένων στο χώρο εισόδου[57]

$$K(x, x_i) = \varphi(x) * \varphi(x_i) = \sum_{i=1}^{m_1} \varphi_i(x) * \varphi_i(x_i)$$

Το βέλτιστο υπερεπίπεδο περιγράφεται από την παρακάτω σχέση:

$$\sum_{i=1}^{N} a_i d_i K(x, x_i)$$

Τύπος του SVM δικτύου	***Πυρήνας Εσωτερικού Γινομένου*** $K(\mathbf{x}, \mathbf{x}_i)$	***Παρατηρήσεις***
Πολυωνυμική μηχανή μάθησης	$(\mathbf{x} \cdot \mathbf{x}_i + 1)^p$	Η δύναμη p καθορίζεται εκ των προτέρων
Δίκτυο Ακτινικών Συναρτήσεων Βάσης (RBF)	$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right)$	Το πλάτος σ^2 είναι κοινό για όλους τους πυρήνες και καθορίζεται εκ των προτέρων
Διεπίπεδο Perceptron	$\tanh(\beta_0 \mathbf{x} \cdot \mathbf{x}_i + \beta_1)$	Το θεώρημα του Mercer δεν ικανοποιείται για όλες τις τιμές των β_0 και β_1

Εικόνα 8: Διάφοροι τύποι SVM

[57] Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

3.4 Μετρικές Αξιολόγησης

Προκειμένου να αξιολογηθεί η επίδοση ενός ταξινομητή, έχει προταθεί πληθώρα μετρικών αξιολόγησης. Στη συνέχεια του κεφαλαίου, παρουσιάζονται οι δημοφιλέστερες μετρικές αξιολόγησης αλγορίθμων μηχανικής μάθησης.[58]

$$acc = \frac{a+d}{a+b+c+d}$$

$$prec = \frac{a}{a+c}$$

$$sen = \frac{a}{a+b}$$

$$spec = \frac{d}{c+d}$$

Όπου,

- a (ή TP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1 και ταξινομήθηκαν στην 1
- b (ή FN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 1, αλλά ταξινομήθηκαν στην 2
- c (ή FP) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2, αλλά ταξινομήθηκαν στην 1
- d (ή TN) = όσα παραδείγματα ανήκουν στην κλάση (εξόδου) 2 και ταξινομήθηκαν στην 2

Τέλος, η μετρική F-Measure παρέχει μία συνολική εκτίμηση των μοντέλων, καθώς συνδυάζει δύο άλλες μετρικές, την ανάκληση και την ακρίβεια. Η μετρική F-Measure στην ουσία είναι ο αρμονικός μέσος όρος (harmonic mean) της ανάκλησης και της ακρίβειας, και υπολογίζεται ως εξής:

[58] Gaber, M., Zaslavsky, A., and Krishnaswamy, S. (2007). A survey of classification methods in data streams. In Aggarwal, C., editor, Data Streams, Models and Algorithms, pages 39–59. Springer

$$Fmeasure = \frac{2 * sen * prec}{sen + prec}$$

Ουσιαστικά για κάθε κλάση βλέπουμε το πρόβλημα ως δυαδικό οπού η πρώτη έξοδος είναι η ίδια η κλάση και δεύτερη έξοδος όλες οι υπόλοιπες.[59]

Στο κεφάλαιο 4, περιγράφουμε αναλυτικά τα δεδομένα που θα επεξεργαστούμε και αποτυπώνουμε υπολογιστικά όλα τα βήματα της προεπεξεργασίας που είναι απαραίτητα προκειμένου το σύνολο δεδομένων να έρθει στην κατάλληλη μορφή προκειμένου να εφαρμοστούν οι αλγόριθμοι ταξινόμησης.

[59] Gaber, M., Zaslavsky, A., and Krishnaswamy, S. (2007). A survey of classification methods in data streams. In Aggarwal, C., editor, Data Streams, Models and Algorithms, pages 39–59. Springer

4. Μοντέλο Κατηγοριοποίησης

4.1 Σύνολο Δεδομένων

Έχουμε ένα σύνολο από περίπου 6500 tweets και μεταδεδομένα από τους δύο βασικούς υποψηφίους των Αμερικανικών εκλογών, τη Hillary Clinton και τον Donald Trump.[60]

Name	Type
Id	Numeric
handle	String
text	String
is_retweet	Boolean
original_author	String
time	DateTime
lang	String
retweet_count	Numeric
favorite_count	Numeric
longitude	String
latitude	String
place_full_name	String
source_url	String

Πίνακας 1: Χαρακτηριστικά συνόλου δεδομένων

Στη συνέχεια θα δούμε κι ένα παράδειγμα από tweet έτσι όπως εμφανίζονται μες στο σύνολο δεδομένων.

Name	Type
Id	780925634159796224
handle	HillaryClinton

[60] https://www.kaggle.com/benhamner/clinton-trump-tweets

text	The question in this election: Who can put the plans into action that will make your life better? https://t.co/XreEY9OicG
is_retweet	0
original_author	-
time	2016-09-28 00:22:34
lang	en
retweet_count	218
favorite_count	651
longitude	-
latitude	-
place_full_name	-
source_url	https://studio.twitter.com

Πίνακας 2: Ένα παράδειγμα εγγραφής στο σύνολο δεδομένων

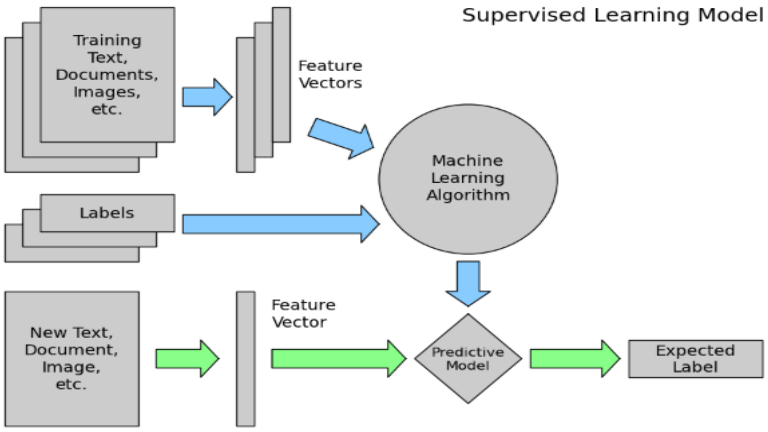
4.2 Περιγραφή μοντέλου πρόβλεψης

Θα προσπαθήσουμε αναλυτικά να μοντελοποιήσουμε το πρόβλημα με υπολογιστικές μεθόδους. Έχουμε να κατηγοριοποιήσουμε ένα σύνολο από tweets $T = \{T_1, \ldots, T_n\}$ σχετικά με το αν υπάρχει θετική ή αρνητική γνώμη σε αυτά. Έχουμε λοιπόν ένα σύνολο λέξεων-κλειδιών τόσο για τη θετική πολικότητα όσο και για την αρνητική. Λέμε λοιπόν πως ένα tweet εκφράζει θετική/αρνητική γνώμη αν εμπεριέχεται κάποια λέξη-κλειδί σε αυτά. Το κατά πόσο ισχυρά θετική/αρνητική η γνώμη εκφράζεται με τη μορφή πιθανότητας (51-100%) και αντίστοιχα γίνεται η κατηγοριοποίηση.

Παρατηρώντας το .csv αρχείο παρατηρούμε πως η στήλη handle είναι η ετικέτα που μας λέει εάν το συγκεκριμένο μήνυμα αναφέρεται στον Donald Trump ή στη Hillary Clinton. Επίσης θα δημιουργήσουμε μια νέα στήλη, με βάση συγκεκριμένες λέξεις κλειδιά, στην οποία θα εξετάζεται η πολικότητα της άποψης(θετική, αρνητική, ουδέτερη).

Το σώμα του κειμένου είναι το επιβλεπόμενο σύνολο εκπαίδευσης. Επιβλεπόμενο γιατί υπάρχουν συγκεκριμένες ετικέτες/κλάσεις. [61]Χρησιμοποιώντας λοιπόν αυτά τα δεδομένα θα εκπαιδεύσουμε ένα μοντέλο εκπαίδευσης, προκειμένου να μπορεί να διαχωρίζει τα tweets ανάμεσα στους δύο βασικούς υποψηφίους αυτόματα, αλλά και να εξετάζει και την άποψη του χρήστη.

Συνεπώς, με αυτό το μοντέλο, θα είμαστε σε θέση να κατηγοριοποιούμε αυτόματα μη-κατηγοριοποιημένα(unlabeled) tweets σε έναν από τους δύο υποψηφίους και σε μια από τις 3 κατηγορίες πολικότητας.



Εικόνα 9: Μοντέλο επιβλεπόμενης μάθησης[62]

4.2.1 Εισαγωγή των δεδομένων

Χρησιμοποιούμε τη βιβλιοθήκη Panda της Python για την εισαγωγή των δεδομένων μου.

```
tweets = pd.read_csv('US_elections.csv', encoding="utf-8")
```

[61] https://www.kaggle.com/benhamner/clinton-trump-tweets
[62] Introduction-to-machine-learning (May 2015), Amit Kumar.http://www.allprogrammingtutorials.com/tutorials/introduction-to-machine-learning.php

Μπορούμε συνολικά να δούμε το σύνολο των tweets, αλλά και να αποτυπώσουμε σ έναν πίνακα κάποια βασικά στατιστικά στοιχεία.[63]

handle		text
HillaryClinton	Count	3226
	Unique	3224
	top	A man who talks about our veterans and militar...
	freq	2
DonaldTrump	Count	3218
	Unique	3210
	top	MAKE AMERICA GREAT AGAIN!
	freq	8

Πίνακας 3: Βασικά στοιχεία για τους 2 υποψηφίους

4.2.2 Επεξεργασία δεδομένων

4.2.2.1 Bag of Words

Το μοντέλο bag-of-words είναι μια απλοποιημένη αναπαράσταση που χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας και ανάκτηση πληροφοριών (IR). Σε αυτό το μοντέλο, ένα κείμενο (όπως μια φράση ή ένα έγγραφο) αναπαρίσταται ως μία «τσάντα», δηλαδή μια πληθώρα από λέξεις, παραβλέποντας γραμματική, ακόμα και τη σειρά των λέξεων, αλλά διατηρώντας την πολλαπλότητα.

Το μοντέλο Bag-of-Words χρησιμοποιείται συνήθως σε μεθόδους ταξινόμησης κειμένων όπου η εμφάνιση (συχνότητα) της κάθε λέξης χρησιμοποιείται ως ένα χαρακτηριστικό γνώρισμα για την εκπαίδευση του ταξινομητή.[64]

[63] http://scikit-learn.org/0.16/tutorial/text_analytics
[64] http://scikit-learn.org/0.16/tutorial/text_analytics

```
def split_into_tokens(message):
    message = message  # convert bytes into proper unicode
    return TextBlob(message).words
messages.message.head()
```

0 [the, question, in, this, election, who, can, ...]

1 [if, we, stand, together, there, 's, nothing, ...]

2 [both, candidates, were, asked, about, how, th...]

3 [join, me, for, a, 3pm, rally, tomorrow, at, t...]

4 [this, election, is, too, important, to, sit, ...]

4.2.2.2 Stemming and Lemmatization[65]

Stemming είναι μια μέθοδος κανονικοποίησης δεδομένων, όπου η κατάληξη της λέξης αφαιρείται από τη βασική δομή ρίζας της, η οποία ονομάζεται stem. Για παράδειγμα, το παιχνίδι λέξεων έχει πολλές μορφές από τη γλωσσική άποψη: playing, played, plays κλπ. Οι αναλύσεις κειμένων δεν απαιτούν διαφοροποίηση μεταξύ των διαφόρων χρονικών περιόδων της λέξης.

Lemmatization είναι παρόμοια με την τεχνική που χρησιμοποιείται για την κανονικοποίηση του κειμένου. Η κύρια διαφορά είναι ότι ενώ το stemming μπορεί να μετατρέψει τη ρίζα της λέξης σε μη υπάρχουσα λέξη, η Lemmatization βασίζεται σε μέρη των κανόνων ομιλίας και μετατρέπει τη ρίζα της λέξης, που λέγεται λήμμα, σε πραγματική υπάρχουσα λέξη που αναζητείται στο λεξικό NLTK. Ουσιαστικά πρόκειται για μια πιο σύνθετη διαδικασία κανονικοποίησης του κειμένου.

[65] http://scikit-learn.org/0.16/tutorial/text_analytics

```
def split_into_lemmas(message):
    message = message.lower()
    words = TextBlob(message).words
    # for each word, take its "base form" =
lemma
    return [word.lemma for word in words]
```

0 [the, question, in, this, election, who, can, ...]

1 [if, we, stand, together, there, 's, nothing, ...]

2 [both, **candidate**, were, **ask**, about, how, th...]

3 [join, me, for, a, 3pm, rally, tomorrow, at, t...]

4 [this, election, is, too, important, to, sit, ...]

4.2.2.3 Μοντέλο διανυσματικού χώρου

Στη συνέχεια πρέπει να μετατρέψουμε κάθε μήνυμα κειμένου, που πιο πάνω το αναπαραστήσαμε διαχωρισμένο σε λήμματα, σε διάνυσμα προκειμένου να γίνει κατανοητό από το μοντέλο μηχανικής μάθησης.

Για να γίνει αυτό, απαιτούνται 3 βήματα, στο μοντέλο bag-of-words:[66]

1. Τη μέτρηση της συχνότητας εμφάνισης μιας λέξης μέσα σ ένα κείμενο.
2. Τον υπολογισμό του «βάρους» αυτής της μέτρησης. Τα λήμματα που εμφανίζονται συχνότερα θα έχουν μεγαλύτερο βάρος.
3. Κανονικοποίηση των διανυσμάτων σε μονάδα μήκους(L2 νόρμα)

Ουσιαστικά με αυτή τη διαδικασία δημιουργείται ένα μεγάλο αραιό μητρώο με συγκεκριμένες γραμμές, στήλες και μηδενικά.

```
messages_bow = bow_transformer.transform(messages['message'])
print('sparse matrix shape:', messages_bow.shape)
print('number of non-zeros:', messages_bow.nnz)
print('sparsity: %.2f%%' % (100.0 * messages_bow.nnz /
(messages_bow.shape[0] * messages_bow.shape[1])))
```

Out: sparse matrix shape: (5722, 9016)
number of non-zeros: 91236
sparsity: 0.18%

4.2.2.4 Αναπαράσταση δεδομένων – TF-IDF

Κατά την αναπαράσταση των κειμένων, ως επί το πλείστον, το βάρος κάθε όρου, ισούται με τη συχνότητα εμφάνισης του όρου, στο αντίστοιχο κείμενο. Η επιλογή της συχνότητας, ως στάθμιση, έχει ως αποτέλεσμα, οι όροι με τη μεγαλύτερη συχνότητα, να θεωρούνται ως οι περισσότερο αντιπροσωπευτικοί όροι του κειμένου, λόγω της βαρύτητάς τους. Μία λύση σε αυτό το πρόβλημα αποτελεί η στάθμιση TF-IDF[67], όπου TF (Term Frequency) η συχνότητα του όρου, ενώ IDF (Inverse Document Frequency) είναι ένα βάρος που δηλώνει τη σημαντικότητα ενός όρου του κείμενου, σε σχέση με το σύνολο των κειμένων. Η στάθμιση υπολογίζεται από τον πολλαπλασιασμό των TF και IDF.
Η στάθμιση TF-IDT, δίνει αρκετά καλά αποτελέσματα, καθώς το βάρος IDF παίρνει μεγάλες τιμές, όταν ένας όρος, υπάρχει σε λίγα κείμενα, ενώ, όταν ο όρος συναντάται σε πολλά από τα κείμενα, τότε το βάρος IDF παίρνει μικρές. τιμές. Με αυτή τη στάθμιση, οι σπάνιοι όροι έχουν υψηλό IDF, και όροι με μεγάλη συχνότητα βαρύνονται με χαμηλότερο IDF. Αυτή η προσέγγιση, έχει ως αποτέλεσμα, τα stopwords να παίρνουν σχετικά μικρό βάρος και να μην αποτελούν πλέον τους πιο αντιπροσωπευτικές όρους στα κείμενα.

```
tfidf_transformer = TfidfTransformer().fit(messages_bow)
messages_tfidf = tfidf_transformer.transform(messages_bow)
print(messages_tfidf)
print(messages_tfidf.shape)
```

[67] http://scikit-learn.org/0.16/tutorial/text_analytics

(0, 8755) 0.217922181616
(0, 8593) 0.161469587626
(0, 8572) 0.195824391455
(0, 7779) 0.183979385543
(0, 7722) 0.196876463794
(0, 7720) 0.162516355764
(0, 6287) 0.321250380345

(5722, 9016)

4.2.3 Επιλογή Αλγόριθμου

Οι αλγόριθμοι που έχουν επιλεγεί, για την υλοποίηση των μοντέλων, προέρχονται από την επιβλεπόμενη μηχανική μάθηση. Ο Multinomial Naive Bayes και ο αλγόριθμός SVM, με γραμμική συνάρτηση πυρήνα, θα αποτελέσουν τη βάση των μοντέλων κατηγοριοποίησης που μελετηθούν στη συνέχεια της εργασίας. Μετά την ολοκλήρωση των μοντέλων, θα πραγματοποιηθεί συγκριτική ανάλυση των αποδόσεών τους, στα τρία σύνολα δεδομένων. Η εισαγωγή των αλγορίθμων γίνεται με τη χρήση της βιβλιοθήκης sklearn.[68]

4.2.4 Αξιολόγηση του μοντέλου

Για την αξιολόγηση των μοντέλων κατηγοριοποίησης, έχουν επιλεγεί οι παρακάτω μετρικές, τις οποίες και αναλύσαμε προηγουμένως:

- Ορθότητα(Accuracy)
- Ανάκληση(Recall)
- Ακρίβεια(Precision)
- F-Measure

[68] http://scikit-learn.org/0.16/tutorial/text_analytics

Επιπρόσθετα, θα εξάγουμε και τους Πίνακες Σύγχυσης (Confusion Matrices) των ταξινομητών, για την περαιτέρω ερμηνεία των αποτελεσμάτων.

4.2.5 Ανάλυση συναισθήματος μέσα από τα tweets

Στο Κεφάλαιο 2 είδαμε αναλυτικά τη διαδικασία εξόρυξη γνώμης και ανάλυσης συναισθημάτων. Εφαρμόζουμε λοιπόν στη συγκεκριμένη υλοποίηση μια κατηγοριοποίηση συναισθήματος (sentiment classification) με βάσει κάποιες χαρακτηριστικές λέξεις-κλειδιά, με θετική και αρνητική χροιά. Χρησιμοποιούμε το λεξικό γνώμης, το οποίο έχει θετικές και αρνητικές λέξεις γνώμης (**Positive.txt, Negative.txt**).[69]

Στο κεφάλαιο 5 παρουσιάζονται αναλυτικά με τη χρήση πινάκων και εικόνων όλα τα πειραματικά αποτελέσματα από τη διαδικασία ταξινόμησης, ενώ γίνεται και σύγκριση μεταξύ των αλγόριθμων. Τέλος χρησιμοποιούμε δειγματικά κάποια γνωστά tweets της προεκλογικής περιόδου, προκειμένου να εφαρμόσουμε το προγνωστικό μοντέλο κατηγοριοποίησης σε τυχαία tweets και να αξιολογήσουμε την απόδοσή του.

[69] Bing Liu. "Sentiment Analysis and Subjectivity." A Chapter in Handbook of Natural Language Processing, Second Edition, (editors: N.Indurkhya and F.J.Damerau), 2010

5. Πειραματικά Αποτελέσματα

Στη συνέχεια παρουσιάζονται αναλυτικά τα πειραματικά αποτελέσματα από την υλοποίησή μας. Αρχικά παρουσιάζονται τα αποτελέσματα με τη χρήση του Multinomial Naive Bayes και στη συνέχεια με τη χρήση των SVM. Εν συνεχεία, γίνεται μια σύγκριση μεταξύ τους για τον προτιμώμενο αλγόριθμο, εξάγουμε τις λέξεις με τη μεγαλύτερη δημοφιλία τόσο στα tweets που αναφέρονται στη Hillary όσο και σε αυτά που αναφέρονται στον Trump και τέλος ταξινομούμε κάποια τυχαία tweets στη βάση όλης της επεξεργασίας που έχουμε κάνει μέχρι τώρα.

5.1 Κατηγοριοποίηση με Multinomial Naïve Bayes

Χωρίζουμε το σύνολο δεδομένων που έχουμε σε σύνολο εκπαίδευσης(training set) και σύνολο ελέγχου(test set). Η συνήθης πρακτική που ακολουθείται είναι να χωρίσουμε το σύνολο εκπαίδευσης(Training_set) σε μικρότερα υποσύνολα, εν προκειμένω σε 10. Συνεπώς εκπαιδεύουμε το μοντέλο σε 9 μέρη και υπολογίζουμε το Accuracy στο τελευταίο μέρος (validation set). Επαναλαμβάνουμε 10 φορές(παίρνοντας κάθε φορά διαφορετικό validation set). Με αυτόν τον τρόπο μπορούμε να έχουμε μια αίσθηση για την αξιοπιστία, τη σταθερότητα του μοντέλου, που δημιουργήσαμε. Εάν βγάλουμε πολύ διαφορετικά νούμερα από τη γενική αξιολόγηση(accuracy) του μοντέλου, τότε σημαίνει ότι πρέπει να επεξεργαστούμε εκ νέου τα δεδομένα μας, καθώς το μοντέλο μας δε δείχνει σταθερό.

CV = 10: Χωρίζουμε το σύνολο δεδομένων τυχαία σε 10 μέρη, 9 για εκπαίδευση και 1 για τα αποτελέσματα

Test_size = 0.1: Το μέγεθος του συνόλου ελέγχου αποτελεί το 10% του συνολικού δείγματος.

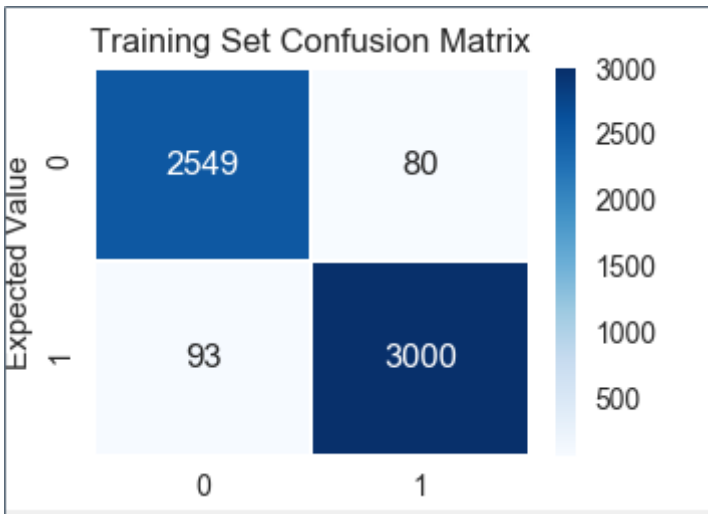
Training_size	Test_size	Data_size
5149	573	5722

Πίνακας 4: Διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο ελέγχου

5.1.1 Κατηγοριοποίηση συνόλου εκπαίδευσης για τον υποψήφιο

Αρχικά θα αξιολογήσουμε το σύνολο εκπαίδευσης(training set) για την **κατηγοριοποίηση των tweets ανά υποψήφιο.**

Accuracy on training set: 96.98%



Εικόνα 10: Confusion Matrix συνόλου εκπαίδευσης - επιλογή υποψηφίου

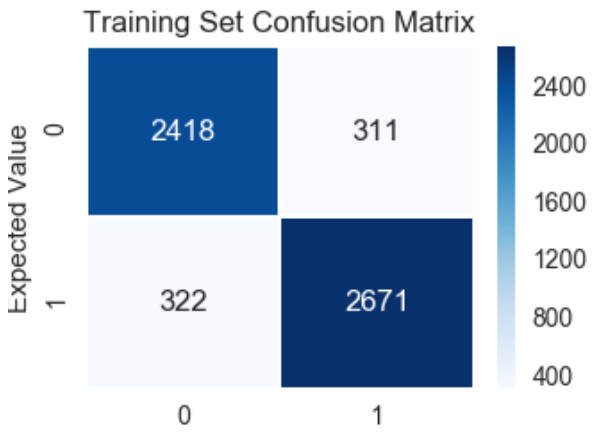
	precision	recall	f1-score	support
Hillary	0.96	0.97	0.97	2629
Trump	0.97	0.97	0.97	3093
avg / total	0.97	0.97	0.97	5722

Πίνακας 5: Μετρικές αξιολόγησης συνόλου εκπαίδευσης για την επιλογή υποψηφίου

5.1.2 Κατηγοριοποίηση συνόλου εκπαίδευσης για την πολικότητα

Στη συνέχεια θα ακολουθήσουμε την ίδια διαδικασία για να αξιολογήσουμε και την **πολικότητα της άποψης για το tweet(θετικό, αρνητικό),** χρησιμοποιώντας τον ίδιο διαχωρισμό σε σύνολο εκπαίδευσης και ελέγχου.

Accuracy on training set: 88.94%



Εικόνα 11: Confusion Matrix συνόλου εκπαίδευσης – πολικότητα

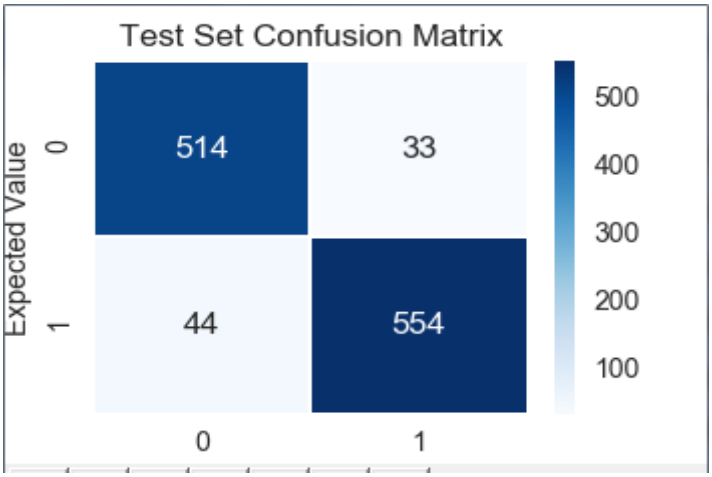
	precision	**recall**	**f1-score**	**support**
Negative	0.88	0.89	0.88	2729
Positive	0.90	0.89	0.89	2993
avg / total	0.97	0.97	0.97	5722

Πίνακας 6: Μετρικές αξιολόγησης συνόλου εκπαίδευσης για την πολικότητα

5.1.3 Κατηγοριοποίηση συνόλου ελέγχου για τον υποψήφιο

Θα αξιολογήσουμε επίσης και το **σύνολο ελέγχου(test set) για την επιλογή υποψηφίου**

Accuracy on test set: 92.96%



Εικόνα 12: Confusion Matrix για το σύνολο ελέγχου - υποψήφιος

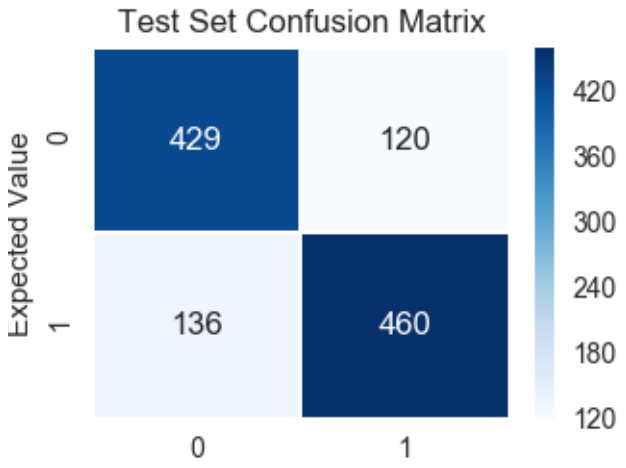
	precision	**recall**	**f1-score**	**support**
Hillary	0.92	0.94	0.93	547
Trump	0.94	0.93	0.94	598
avg / total	0.93	0.93	0.93	1145

Πίνακας 7: Μετρικές αξιολόγησης για το σύνολο ελέγχου - Υποψήφιος

5.1.4 Κατηγοριοποίηση συνόλου ελέγχου για την πολικότητα

Θα αξιολογήσουμε επίσης και την **πολικότητα της άποψης για το tweet(θετικό, αρνητικό), για το σύνολο ελέγχου.**

Accuracy on test set: 77.32%



Εικόνα 13: Confusion Matrix για το σύνολο ελέγχου – πολικότητα

	precision	**recall**	**f1-score**	**support**
Negative	0.76	0.78	0.77	549
Positive	0.79	0.77	0.78	596
avg / total	0.78	0.78	0.78	1145

Πίνακας 8: Μετρικές αξιολόγησης για το σύνολο ελέγχου - πολικότητα

5.1.5 Αφελής Bayes και Παράμετροι

Έχει ένα ενδιαφέρον να εξετάσουμε την επίδραση των μετρικών TF-IDF στην αξιολόγηση του μοντέλου που χρησιμοποιούμε. Υπάρχει όντως επίδραση των διαδικασίων προεπεξεργασίας του κειμένου(Lemmatization πχ) ή όχι?

Παρατηρούμε από τα αποτελέσματα πώς για:

- 'bow__analyzer': <function split_into_tokens>
- 'tfidf__use_idf': True

παίρνουμε τα καλύτερα αποτελέσματα

5.2 Κατηγοριοποίηση με Μηχανές Διανυσμάτων Υποστήριξης

Ακολουθούμε την ίδια διαδικασία διαχωρισμού των δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου.

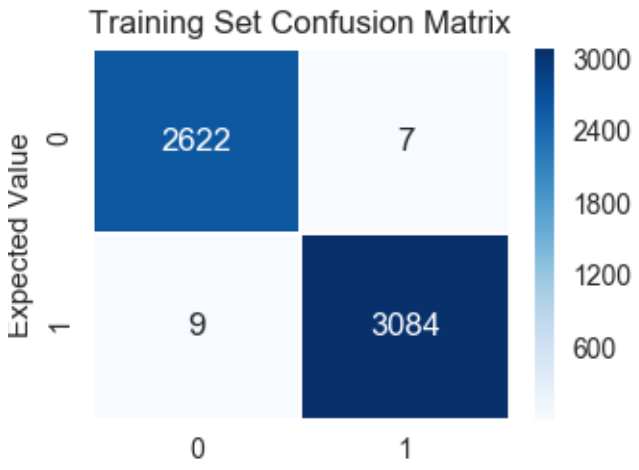
CV = 10: Χωρίζουμε το σύνολο δεδομένων τυχαία σε 10 μέρη, 9 για εκπαίδευση και 1 για τα αποτελέσματα

Test_size = 0.1: Το μέγεθος του συνόλου ελέγχου αποτελεί το 10% του συνολικού δείγματος.

5.2.1 Κατηγοριοποίηση συνόλου εκπαίδευσης για τον υποψήφιο

Αρχικά θα αξιολογήσουμε το σύνολο εκπαίδευσης(training set) για την **κατηγοριοποίηση των tweets ανά υποψήφιο.**

Accuracy on training set: 99.72%



Εικόνα 14: Confusion Matrix για το σύνολο εκπαίδευσης(SVM) - υποψήφιος

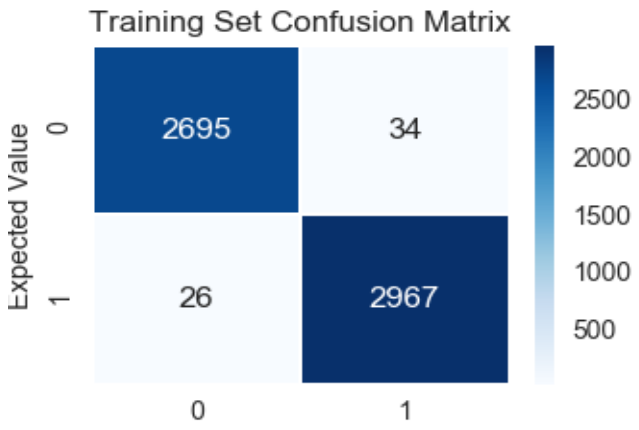
	precision	recall	f1-score	support
Hillary	1.00	1.00	1.00	2629
Trump	1.00	1.00	1.00	3093
avg / total	1.00	1.00	1.00	5722

Πίνακας 9: Μετρικές αξιολόγησης για το σύνολο εκπαίδευσης(SVM) - Υποψήφιος

5.2.2 Κατηγοριοποίηση συνόλου εκπαίδευσης για την πολικότητα

Στη συνέχεια θα αξιολογήσουμε το σύνολο εκπαίδευσης για την **πολικότητα των tweets(θετικό, αρνητικό)**

Accuracy on training set: 98.95%



Εικόνα 15: Confusion Matrix για το σύνολο εκπαίδευσης (SVM) – πολικότητα

	precision	recall	f1-score	support
Negative	0.99	0.99	0.99	2729
Positive	0.99	0.99	0.99	2993
avg / total	0.99	0.99	0.99	5722

Πίνακας 10: Μετρικές αξιολόγησης για το σύνολο εκπαίδευσης(SVM) – Πολικότητα

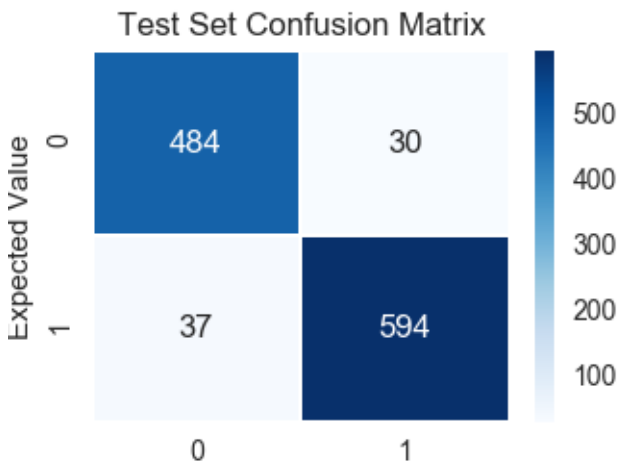
5.2.3 Κατηγοριοποίηση συνόλου ελέγχου για τον υποψήφιο

Θα αξιολογήσουμε επίσης και το σύνολο ελέγχου(test set) για την επιλογή υποψηφίου

Accuracy on test set: 93.47%

	precision	recall	f1-score	support
Hillary	0.93	0.94	0.94	514
Trump	0.95	0.94	0.95	631
avg / total	0.94	0.94	0.94	1145

Πίνακας 11: Μετρικές αξιολόγησης για το σύνολο ελέγχου(SVM) - Υποψήφιος

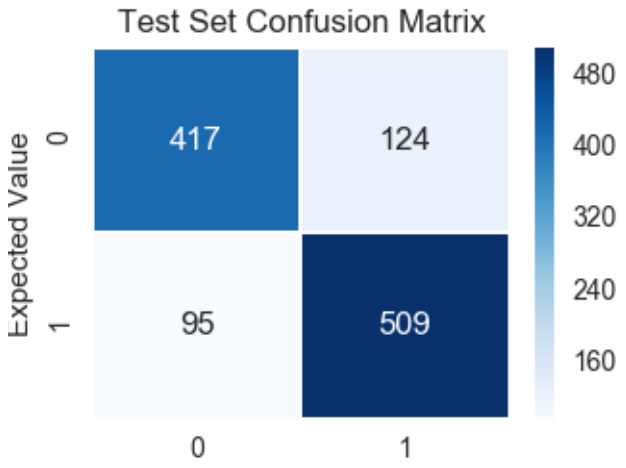


Εικόνα 16: Confusion Matrix για το σύνολο ελέγχου(SVM) - υποψήφιος

5.2.4 Κατηγοριοποίηση συνόλου ελέγχου για την πολικότητα

Τέλος αξιολογούμε και το σύνολο ελέγχου για την πολικότητα του tweet.

Accuracy on test set: 80.27%



Εικόνα 17: Confusion Matrix για το σύνολο ελέγχου(SVM) – πολικότητα

	precision	recall	f1-score	support
Hillary	0.81	0.77	0.79	541
Trump	0.80	0.84	0.82	604
avg / total	0.81	0.81	0.81	1145

Πίνακας 12: Μετρικές αξιολόγησης για το σύνολο ελέγχου(SVM) – Πολικότητα

5.2.5 Γραμμικός SVC και Παράμετροι

Στη γραμμική κατηγοριοποίηση με SVM, εκπαιδεύουμε και τις παραμέτρους του LinearSVC, προκειμένου να επιλέξουμε τον καλύτερο Classifier_kernel και παράμετρο C.

```
# pipeline parameters to automatically explore and tune
param_svm = [
{'classifier__C': [1, 10, 100, 1000], 'classifier__kernel':
['linear']},
  {'classifier__C': [1, 10, 100, 1000], 'classifier__gamma':
[0.001, 0.0001], 'classifier__kernel': ['rbf']},
]
```

Με βάση τα αποτέλεσματα, παρατηρώ πως τα καλύτερα αποτελέσματα τα παίρνω για:

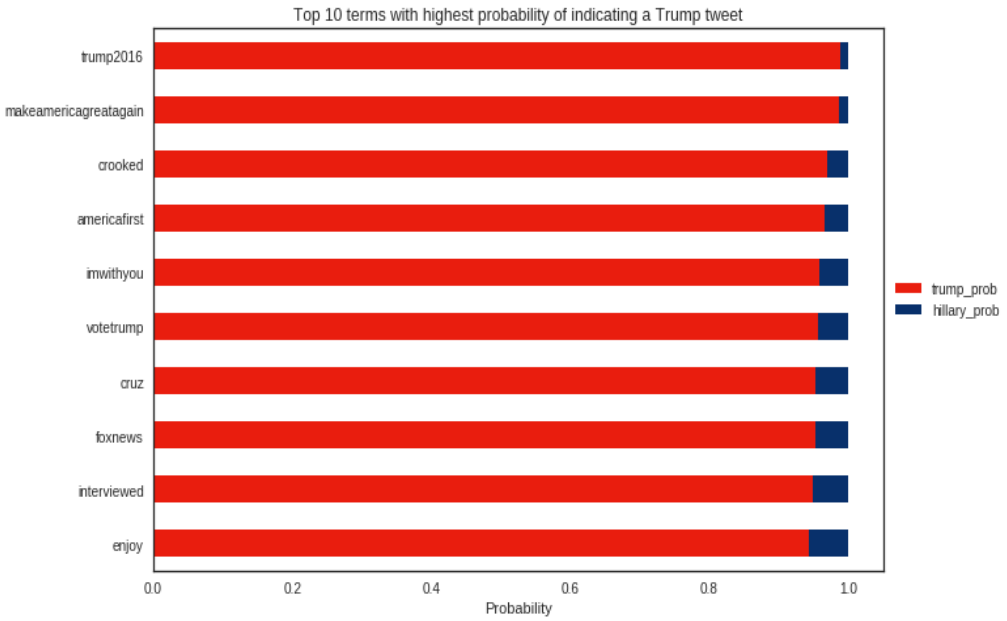
- Linear Kernel
- C = 1

5.3 Σύγκριση Αλγόριθμων Επεξεργασίας

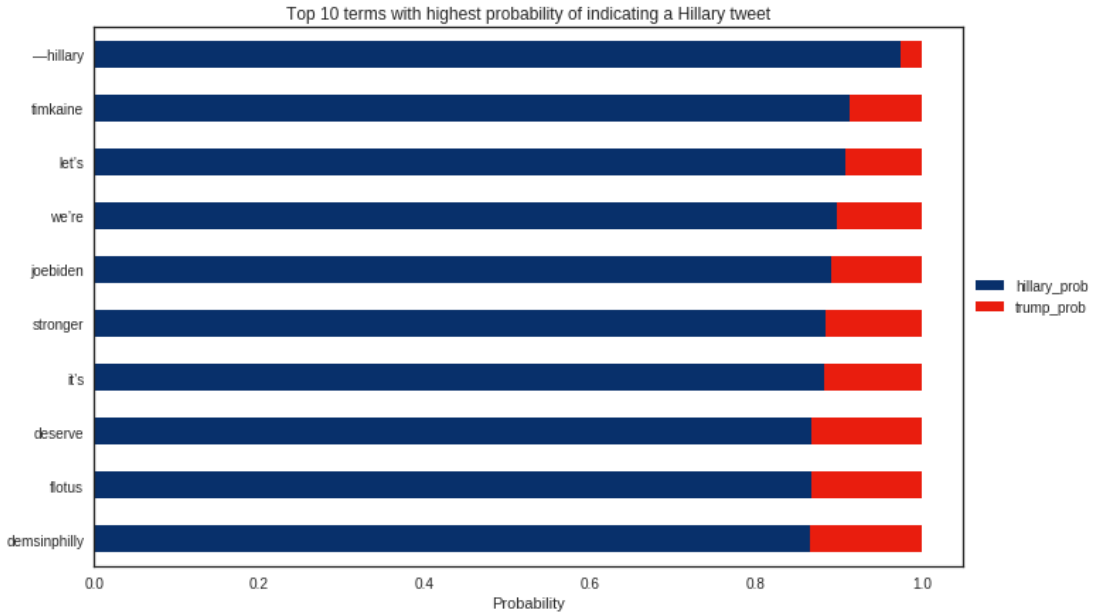
Παρατηρούμε πως με το **LinearSVC** παίρνουμε καλύτερα αποτελέσματα απ' ότι με τον **Naïve Bayes**, όσον αφορά την αποδοτικότητα τις πρόβλεψης κατηγοριοποίησης τυχαίων tweets σε έναν από τους 2 υποψηφίους.

5.4 Εξαγωγή Όρων με μεγάλη δημοφιλία από το σύνολο δεδομένων

Πρόκειται ουσιαστικά για μια διαδικασία στην οποία εξάγουμε λέξεις με υψηλή δημοφιλία(TF-IDF), ξεχωριστά για κάθε υποψήφιο. Μας ενδιαφέρει όχι μόνο η συχνότητα εμφάνισης, αλλά η σπανιότητα ενός όρου. Δηλαδή, πολλοί όροι που χρησιμοποιούνται από έναν υποψήφιο θα χρησιμοποιηθούν επίσης από τον αντίπαλό του, χωρίς να δίνουν ξεκάθαρη εικόνα για τον ‘κάτοχο’ του tweet. Για παράδειγμα, ο όρος "και" χρησιμοποιείται συχνά στο σύνολο δεδομένων, αλλά η επεξεργασία μέσω του μοντέλου θα έχει ως αποτέλεσμα πιθανότητα πρόβλεψης περίπου 50/50 (δεν είναι καθόλου χρήσιμη). Ωστόσο, υπάρχουν όροι που χρησιμοποιούνται πολύ λιγότερο συχνά, αλλά σχεδόν αποκλειστικά από ορισμένους υποψηφίους. Αυτές είναι οι λέξεις (ή ομάδες λέξεων) που αναζητά αυτό το μοντέλο για να κάνουν μια ακριβή πρόβλεψη.



Εικόνα 18: Δημοφιλείς λέξεις για τον Trump



Εικόνα 19: Δημοφιλείς λέξεις για τη Hillary

5.5 Παραδείγματα ταξινόμησης τυχαίων tweets

Θα παραθέσουμε 10 κείμενα από tweets[70] [71] [72] ,τα οποία δημιουργήθηκαν κατά τη διάρκεια της προεκλογικής περιόδου και θα αξιολογήσουμε την υλοποίησή μας, την πρόβλεψη δηλαδή σε ποιόν υποψήφιο από τους δύο αναφέρεται και ποια μπορεί να είναι η γνώμη του χρήστη (θετική, αρνητική).

Αρχικά θα πάρουμε αποτελέσματα εκπαιδεύοντας τα tweets με τη χρήση του **Naïve Bayes**

Με βάση λοιπόν την παραπάνω λογική παραθέτουμε στη συνέχεια 10 ενδεικτικά tweets:

Τα αποτελέσματα που παίρνουμε είναι τα εξής:

[70] http://www.cbsnews.com/news/donald-trump-and-hillary-clintons-most-popular-tweets-of-2016/
[71] http://www.businessinsider.com/most-popular-election-tweets-from-trump-and-clinton-2016-11
[72] http://www.complex.com/life/2016/06/presidential-candidate-tweets

Tweet #1: ' With this election we're simultaneously breaking through the glass ceiling and the rock bottom. We got a really big room now '

I'm about 76% sure this was tweeted by Hillary and the polarity is about 59% Positive

Tweet #2: ' Retweet if you are: -A woman -An immigrant -LGBT+ -Muslim -African American -Latino/Latina-In anywway completely terrified right now '

I'm about 55% sure this was tweeted by Hillary and the polarity is about 60% Negative

Tweet #3: ' Such a beautiful and important evening! The forgotten man and woman will never be forgotten again. We will all come together as never before '

I'm about 64% sure this was tweeted by Trump and the polarity is about 55% Positive

Tweet #4: ' How long did it take your staff of 823 people to think that up--and where are your 33,000 emails that you deleted? '

I'm about 62% sure this was tweeted by Trump and the polarity is about 52% Positive

Tweet #5: ' Women have the power to stop. '

I'm about 61% sure this was tweeted by Hillary and the polarity is about 58% Negative

Tweet #6: ' America needs a leader who treats women with respect '

I'm about 73% sure this was tweeted by Hillary and the polarity is about 50% Positive

Tweet #7: ' No one remembers who came in second '

I'm about 52% sure this was tweeted by Trump and the polarity is about 52% Negative

Tweet #8: ' Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure,it's not your fault '

I'm about 60% sure this was tweeted by Trump and the polarity is about 67% Negative

Tweet #9: ' OH MY GOD VERIZON ELECTION NIGHT IS THE WORST TIME FOR THIS ADVERTISEMENT. '

I'm about 54% sure this was tweeted by Trump and the polarity is about 53% Positive

Tweet #10: ' I hear you, Sanders supporters who plan to vote Trump. One time I asked for Coke but they only had Pepsi, so I set fire to my head. '

I'm about 64% sure this was tweeted by Trump and the polarity is about 55% Positive

Στη συνέχεια θα εκπαιδεύσουμε τα tweets με τον αλγόριθμο **SVM.**

Tweet #1: ' With this election we're simultaneously breaking through the glass ceiling and the rock bottom. We got a really big room now '

I'm about 92% sure this was tweeted by Hillary and the polarity is about 92% Positive

Tweet #2: ' Retweet if you are: -A woman -An immigrant -LGBT+ -Muslim -African American -Latino/Latina-In anywway completely terrified right now '

I'm about 100% sure this was tweeted by Hillary and the polarity is about 86% Negative

Tweet #3: ' Such a beautiful and important evening! The forgotten man and woman will never be forgotten again. We will all come together as never before '

I'm about 87% sure this was tweeted by Trump and the polarity is about 78% Negative

Tweet #4: ' How long did it take your staff of 823 people to think that up--and where are your 33,000 emails that you deleted? '

I'm about 87% sure this was tweeted by Trump and the polarity is about 55% Positive

Tweet #5: ' Women have the power to stop. '

I'm about 57% sure this was tweeted by Trump and the polarity is about 59% Negative

Tweet #6: ' America needs a leader who treats women with respect '

I'm about 95% sure this was tweeted by Hillary and the polarity is about 94% Positive

Tweet #7: ' No one remembers who came in second '

I'm about 64% sure this was tweeted by Trump and the polarity is about 86% Negative

Tweet #8: ' Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure,it's not your fault '

I'm about 75% sure this was tweeted by Trump and the polarity is about 95% Negative

Tweet #9: ' OH MY GOD VERIZON ELECTION NIGHT IS THE WORST TIME FOR THIS ADVERTISEMENT. '

I'm about 96% sure this was tweeted by Trump and the polarity is about 92% Positive

Tweet #10: ' I hear you, Sanders supporters who plan to vote Trump. One time I asked for Coke but they only had Pepsi, so I set fire to my head. '

I'm about 95% sure this was tweeted by Trump and the polarity is about 81% Positive

Τα παραπάνω αποτελέσματα με βάση τόσο τους πίνακες με τις δημοφιλείς λέξεις για τους δύο υποψήφιους που παραθέσαμε προηγουμένως, όσο και δημοφιλή tweets που αντλήσαμε από τη βιβλιογραφία και αναφέρονται σε έναν από τους δύο υποψηφίους, έχουν αρκετά καλή ακρίβεια τόσο για την πρόβλεψη του υποψηφίου όσο και για το συναίσθημα που εκφράζεται μέσα από αυτό.

Στο κεφάλαιο 6 γίνεται μια εισαγωγή στο εργαλείο Orange και δημιουργούμε ένα παράδειγμα υλοποίησης ενός αντίστοιχου μοντέλου πρόβλεψης σε ένα παρεμφερές σύνολο δεδομένων.

6. Εξόρυξη κειμένου με τη χρήση του Orange3 της Python

6.1 Τι είναι το Orange3?

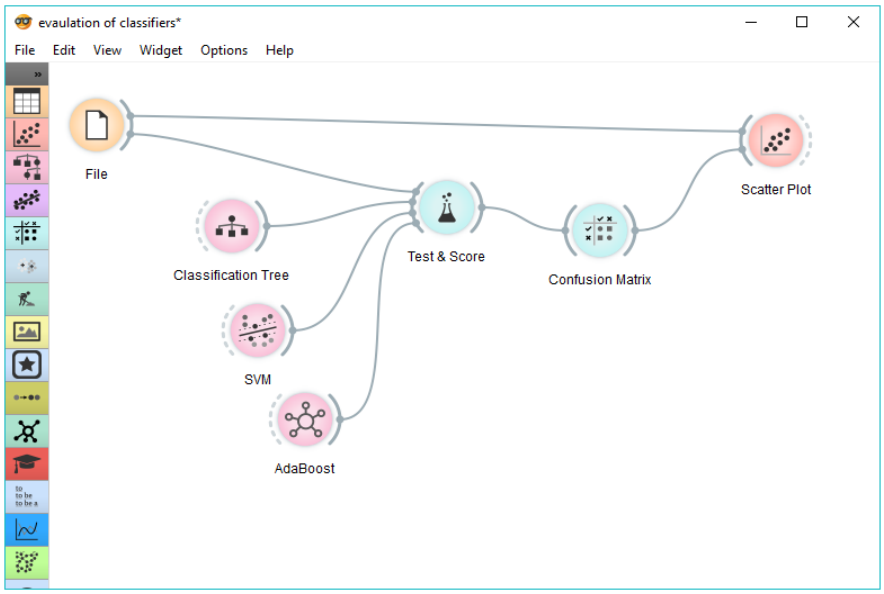
Το Orange είναι μια ανοικτού κώδικα οπτικοποίηση δεδομένων, μηχανικής μάθησης και εργαλείο εξόρυξης δεδομένων. Διαθέτει ένα front-end προγραμματιστικό περιβάλλον για τη διερεύνηση ανάλυσης δεδομένων και τη διαδραστική οπτικοποίησή τους. Επίσης χρησιμοποιείται και σα βιβλιοθήκη της Python.

Το Orange αποτελείται από μικροεφαρμογές(widgets), τα οποία μπορεί να είναι από την απεικόνιση απλών συνόλων δεδομένων, υποσύνολα επιλογής και προεπεξεργασίας μέχρι εμπειρική αξιολόγηση αλγόριθμων μάθησης και προγνωστική μοντελοποίηση. [73]

6.2 Βασικά στοιχεία του Orange

Το Orange αποτελείται από μια επιφάνεια γραφικών, πάνω στην οποία ο χρήστης τοποθετεί τα widgets και δημιουργεί μια ροή ανάλυσης δεδομένων (data analysis workflow). Τα widgets προσφέρουν τις βασικές λειτουργίες, όπως την ανάγνωση των δεδομένων, τη εμφάνιση ενός πίνακα δεδομένων, τα επιλεγμένα χαρακτηριστικά, την επιλογή αλγόριθμων μάθησης, την οπτικοποίηση στοιχείων δεδομένων κ.α. Ο χρήστης μπορεί να αλληλεπιδράσει εύκολα με το περιβάλλον και να προσαρμόζει τα widgets, ανάλογα με το τι επιθυμεί να κάνει. Στην παρακάτω εικόνα φαίνεται ένα παράδειγμα μιας ροής ανάλυσης δεδομένων στο Orange.

[73] Janez Demšar; Tomaž Curk; Aleš Erjavec; Črt Gorup; Tomaž Hočevar; Mitar Milutinovič; Martin Možina; Matija Polajnar; Marko Toplak; Anže Starič; Miha Stajdohar; Lan Umek; Lan Žagar; Jure Žbontar; Marinka Žitnik; Blaž Zupan (2013). "Orange: data mining toolbox in Python" (PDF). JMLR. 14 (1): 2349–2353.



Εικόνα 20: Παράδειγμα workflow στο Orange

- **Canvas:** Γραφική επιφάνεια για front-end ανάλυση δεδομένων

- **Widgets:**

    - **Data:** Widgets για είσοδο δεδομένων, φιλτράρισμα δεδομένων, δειγματοληψία κ.α
    - **Visualize:** widgets για οπτικοποίηση δεδομένων
    - **Classify:** Σύνολο από αλγόριθμους επιβλεπόμενης μηχανικής μάθησης για την κατηγοριοποίηση των δεδομένων
    - **Regression:** Σύνολο από αλγόριθμους επιβλεπόμενης μηχανικής μάθησης για παλινδρόμηση
    - **Evaluate:** cross-validation, διαδικασίες βάσει δειγματοληψίας και βαθμόλογηση μεθόδων πρόβλεψης
    - **Unsupervised:** Μη επιβλεπόμενοι αλγόριθμοι μάθησης, όπως ο k-means για ομαδοποίηση των δεδομένων και τεχνικές προβολής δεδομένων.

- **Add-Ons:** Associate, Bioinformatics**,** Data fusion, Educational, Image Analytics, Network, **Text Mining**, Time Series

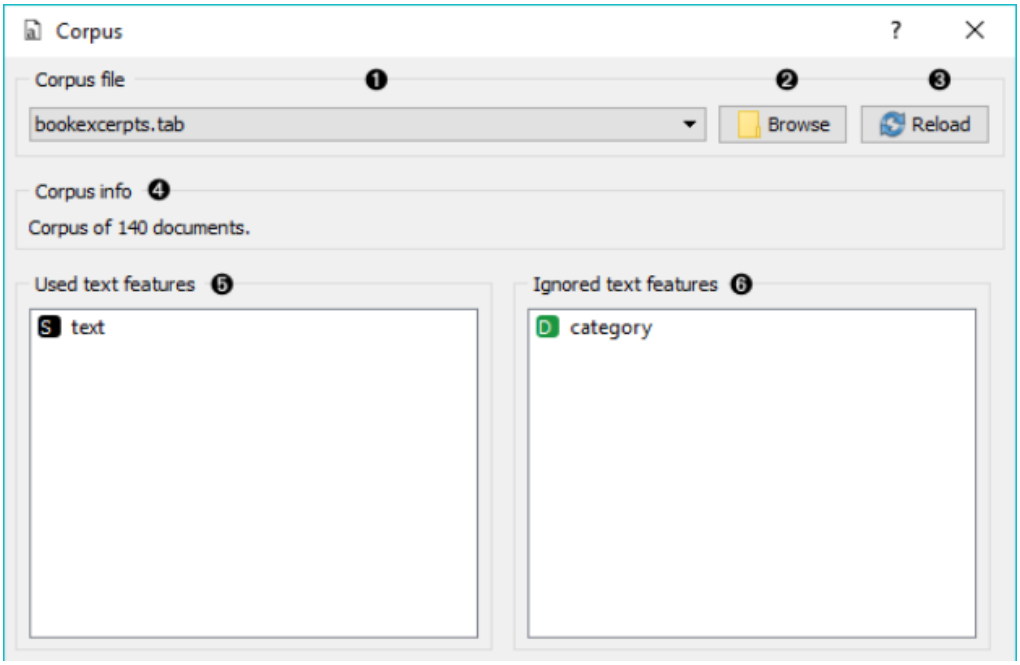
6.3 Επεξεργασία κειμένου με τη χρήση του Orange

Χρησιμοποιούμε το πρόσθετο (add-on) του Orange, Text Mining, προκειμένου να χρησιμοποιηθούν συγκεκριμένες τεχνικές που αφορούν την επεξεργασία κειμένου.

6.3.1 Corpus

Corpus: Φορτώνουμε ένα σώμα από κείμενα κατηγοριοποιημένα με ετικέτες.

Το συγκεκριμένο widget διαβάζει δεδομένα από αρχεία .xlsx ή .csv ή .tab



Εικόνα 21: Φόρτωση συνόλου κειμένων – Corpus στο Orange

Τα βήματα που ακολουθούνται είναι τα εξής:

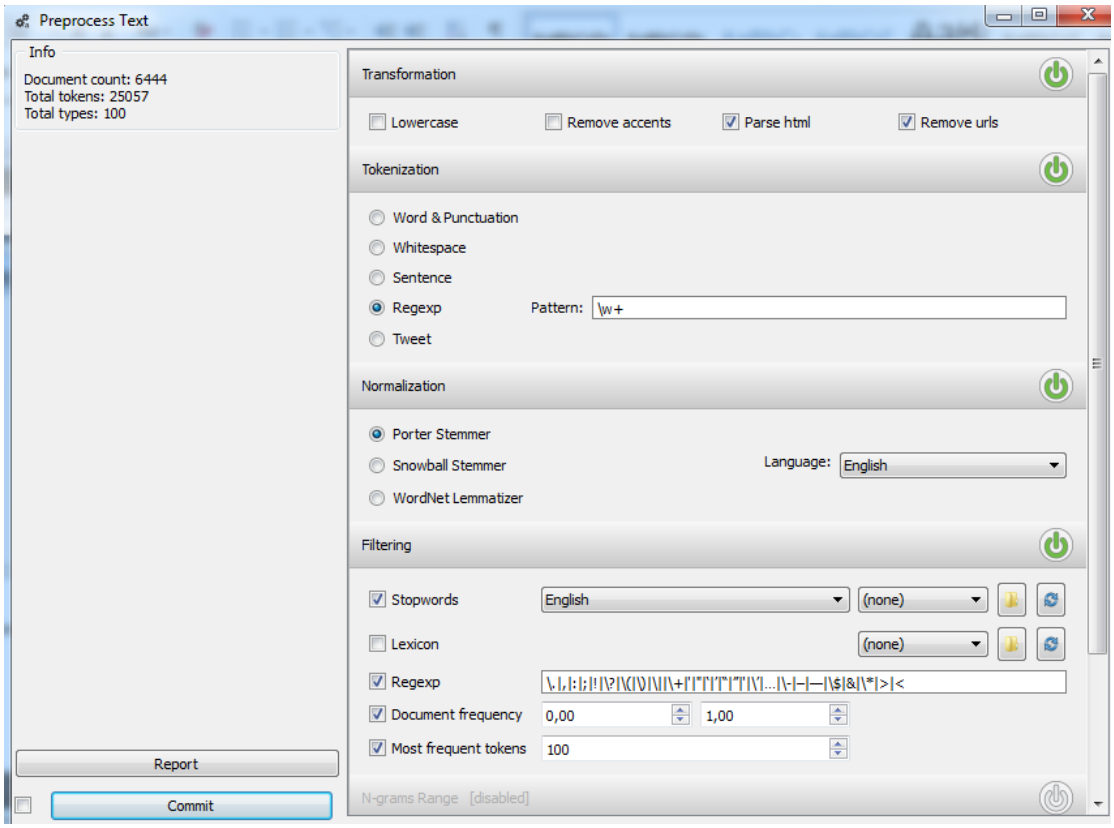
1. Περιήγηση στα αρχεία δεδομένων που έχουμε ήδη ανοίξει ή φορτώνουμε κάποια από τα δειγματικά σύνολα δεδομένων.
2. Περιήγηση από ένα αρχείο δεδομένων

3. Επαναφόρτωση πρόσφατα επιλεγμένων αρχείων δεδομένων
4. Η πληροφορία για το σύνολο δεδομένων
5. Χαρακτηριστικά που θα χρησιμοποιηθούν στην ανάλυση των κειμένων
6. Χαρακτηριστικά που θ' αγνοηθούν στην ανάλυσή μας

Μπορείτε επίσης να μεταφέρετε και να τοποθετήσετε τα χαρακτηριστικά μεταξύ των δύο περιοχών κι επίσης ν' αλλάξετε και τη σειρά εμφάνισης.

6.3.2 Preprocess Text

Το κείμενο προεπεξεργασίας διαχωρίζει το κείμενο σε μικρότερες μονάδες(tokens), τα φιλτράρει, εκτελεί κανονικοποίηση(λημματοποίηση) κ.α. Πιο αναλυτικά τα βήματα παρουσιάζουμε στη συνέχεια:



Εικόνα 22: Προεπεξεργασία δεδομένων στο Orange

1. **Πληροφορίες σχετικά με τα δεδομένα προεπεξεργασίας (Info).** Γίνεται αναφορά στον αριθμό των κειμένων, στο σύνολο των tokens, αλλά και στον αριθμό των μοναδικών tokens, εξαιρώντας διπλογεγγραφές.

2. Μετασχηματισμός δεδομένων εισόδου (Transformation).

- Lowercase, μετατρέπει όλο το κείμενο σε πεζούς χαρακτήρες
- Αφαίρεση τόνων, διαλυτικών
- Διερεύνηση ύπαρξης html tags και αφαίρεση αυτών και διατήρηση μόνο του κειμένου-περιεχομένου.
- Αφαίρεση Urls από το κείμενο

3. Διαχωρισμός του κειμένου σε μικρότερα στοιχεία (λέξεις, προτάσεις, σύνδεσμοι) (Tokenization).

- Word & Punctuation: Διαχωρίζει το κείμενο σε λέξεις και σημεία στίξης
- Whitespace: Χωρίζει το κείμενο με το κενό
- Sentence: Διαχωρίζει το κείμενο μόνο όταν βρει τελείες. Διατηρεί ολόκληρες προτάσεις
- Regexp: Διαχωρίζει το κείμενο με βάση τις κανονικές εκφράσεις που εμείς ορίζουμε
- Tweet: Διαχωρίζει το κείμενο με βάση τη μορφή του tweet πριν επεξεργασθεί, διατηρώντας hashtags, emoticons, κ.α

4. Με την κανονικοποίηση εφαρμόζουμε τη λημματοποίηση στις λέξεις.

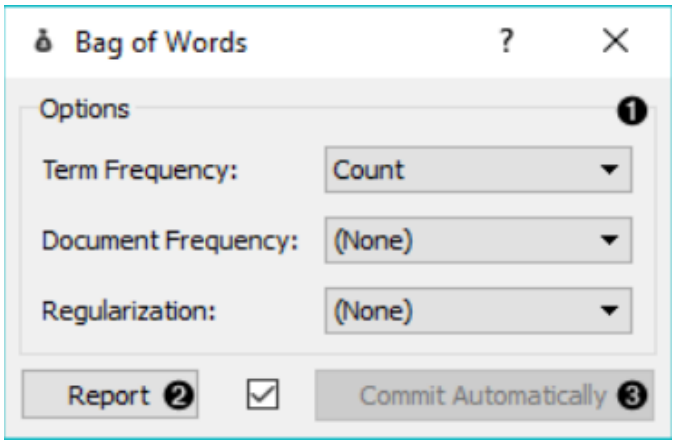
- Porter Stemmer: Εφαρμόζει την κλασική λημματοποίηση
- Snowball Stemmer: Εφαρμόζει μια βελτιωμένη έκδοση του Porter Stemmer και θέτει τη γλώσσα κανονικοποίησης στα Αγγλικά.
- WordNet Lemmatizer: εφαρμόζει ένα δίκτυο από σημαντικά συνώνυμα των λημμάτων βάσει ενός μεγάλου αγγλικού λεξικού

5. **Με το φιλτράρισμα μπορούμε να απομακρύνουμε ή να διατηρήσουμε λέξεις.**

- Stopwords: απομακρύνουν σημεία στίξης (και, ή, …)
- Lexicon: Διατηρεί μόνο τις λέξεις που βρίσκονται στο αρχείο που φορτώνεται.
- Regexp: Απομακρύνει λέξεις που αντιστοιχούν στην κανονική έκφραση που έχουμε γράψει
- Document Frequency: Διατηρεί τα λήμματα που εμφανίζονται τόσες φορές εντός του διαστήματος τιμών, τις οποίες προσδιορίζουμε.
- Most frequent tokens: Διατηρεί ένα συγκεκριμένο αριθμό από τα πιο συχνά λήμματα. Η Default τιμή είναι 100.

6.3.3 Bag of words

Το μοντέλο αυτό δημιουργεί ένα σώμα κειμένου με τον αριθμό εμφάνισης των λέξεων μέσα σε ένα κείμενο. Ο αριθμός εμφάνισης μπορεί να είναι είτε απόλυτος αριθμός, είτε δυαδικός(είναι, δεν είναι), είτε λογάριθμος του αριθμού εμφάνισης. Χρησιμοποιείται ιδιαίτερα σε προβλήματα πρόβλεψης.



Εικόνα 23: Μέθοδος 'Bag of Words' στο Orange

Παράμετροι:

Term Frequency:

- Count: αριθμός εμφανίσεων μιας λέξης σ’ ένα κείμενο
- Binary: εμφάνιση ή μη-εμφάνιση της λέξης μες στο κείμενο
- Sublinear: λογάριθμος του Count

Document Frequency:

- None:
- IDF: είναι ένα βάρος που δηλώνει τη σημαντικότητα ενός όρου του κείμενου, σε σχέση με το σύνολο των κειμένων
- Smmoth IDF: Προσθέτει ένα στις συχνότητες εγγράφων για να αποτρέψει τη μηδενική διαίρεση.

Regularization:

- None:
- L1: Κανονικοποιεί το διάνυσμα μήκους στο άθροισμα των στοιχείων
- L2: Κανονικοποιεί το διάνυσμα μήκους στο άθροισμα τετραγώνων

6.3.4 Word Cloud

Το σύννεφο λέξεων εμφανίζει τα λήμματα του κειμένου και το μέγεθός τους υποδηλώνει τη συχνότητα εμφάνισής τους εντός του κειμένου. Επίσης οι λέξεις εμφανίζονται σε μια λίστα με τον αριθμό εμφάνισής τους.

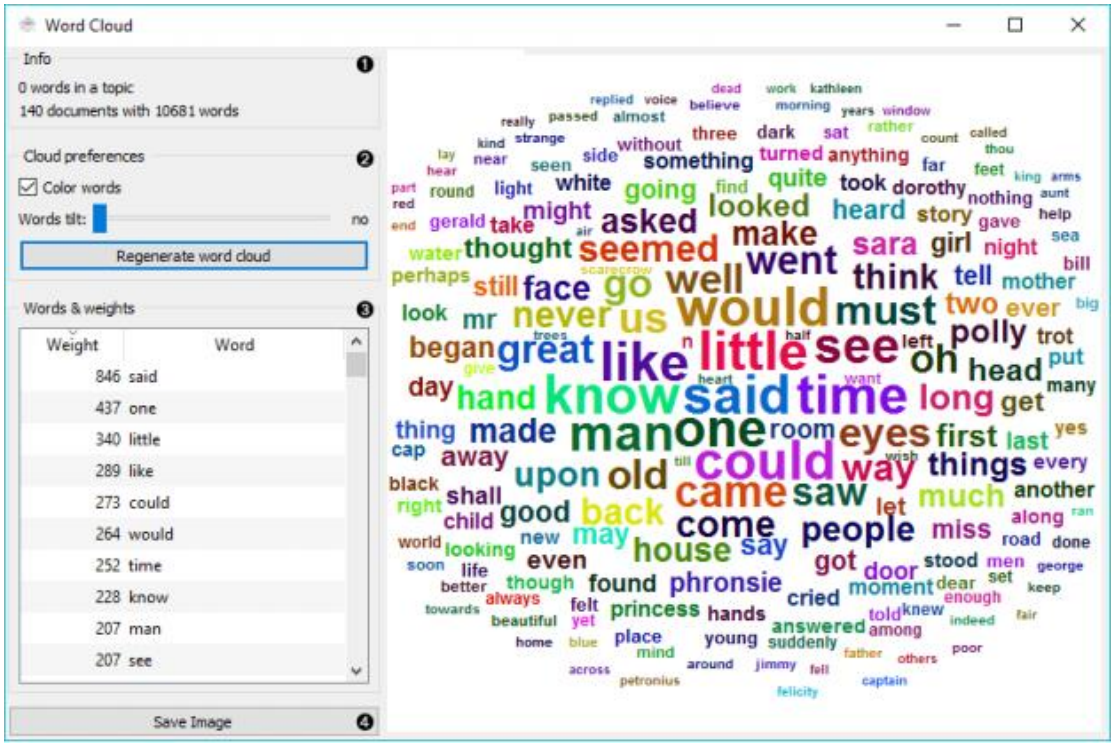
1. Πληροφορία στην είσοδο

- Αριθμός των λημμάτων(tokens)
- Αριθμός των κειμένων και των λημμάτων σε όλο το σώμα κειμένου

2. Διαμόρφωση της εικόνας

- Εάν η επιλογή *Color Words* είναι επιλεγμένη, οι λέξεις θα εμφανιστούν με τυχαία χρώματα. Αν δεν είναι, θα εμφανιστούν ασπρόμαυρα
- Η επιλογή *word tilt* επηρεάζει την κλίση των λέξεων, αλλά συνήθως το αφήνουμε στο 0, που είναι και η default τιμή.
- Πατώντας την επιλογή *Regenerate Word Cloud*, απλά δημιουργούμε εκ νέου ένα σύννεφο λέξεων.

3. Λέξεις και βάρη εμφανίζονται στο πλάι σε μια ταξινομημένη λίστα με βάση τη συχνότητα εμφάνισής τους, σε φθίνουσα σειρά. Κάνοντας κλικ σε μια λέξη αυτή αντιστοιχείται με την ίδια στο σύννεφο λέξεων και μας δίνει σαν έξοδο τα κείμενα στα οποία εμπεριέχεται.



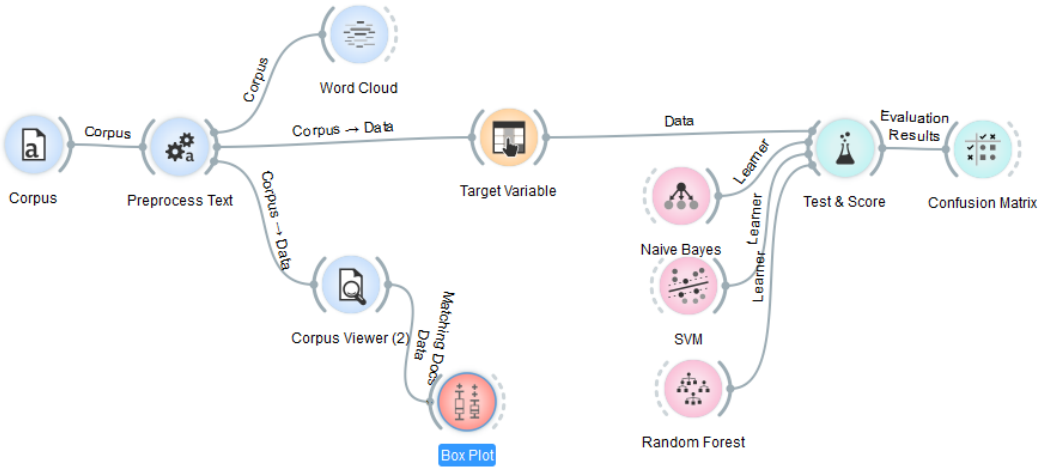
Εικόνα 24: Εφαρμογή Word Cloud από Orange

6.4 Η υλοποίησή μου

Η εξόρυξη πληροφορίας από tweets αποτελεί μια πολύ διαδεδομένη διαδικασία για πολιτικούς αναλυτές, προσπαθώντας να ερμηνεύσουν τις απόψεις των χρηστών σχετικά με υποψηφίους, κόμματα κλπ.

Στο παράδειγμά μας θα χρησιμοποιήσουμε ένα σύνολο δεδομένων, το οποίο έχει δημιουργηθεί μέσα από μια συλλογή χιλιάδων tweets μετά το πρώτο debate μεταξύ των υποψηφίων του ρεπουμπλικανικού κόμματος πριν τις Αμερικανικές εκλογές. Τα tweets έχουν αναφορά σε κάποιον υποψήφιο, ενώ υπάρχει και η άποψη για τον συγκεκριμένο υποψήφιο (θετική, αρνητική, ουδέτερη).

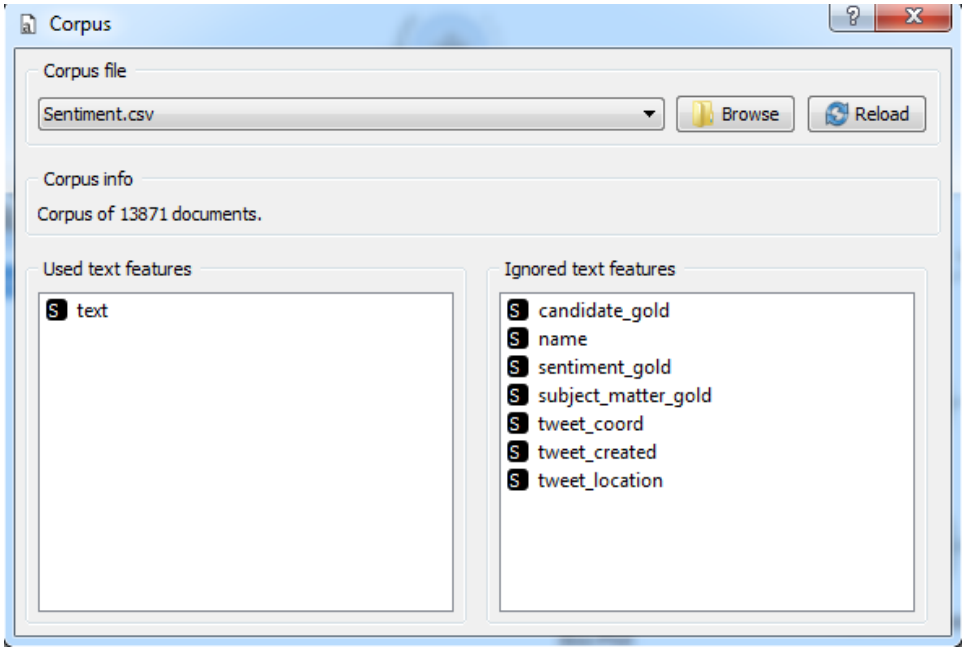
Στην παρακάτω εικόνα φαίνεται όλη η ροή εργασιών, προκειμένου να αποτυπώσουμε ένα προγνωστικό μοντέλο μέσα από συγκεκριμένα βήματα.



Εικόνα 25: Εφαρμογή μοντέλου πρόβλεψης στο Orange

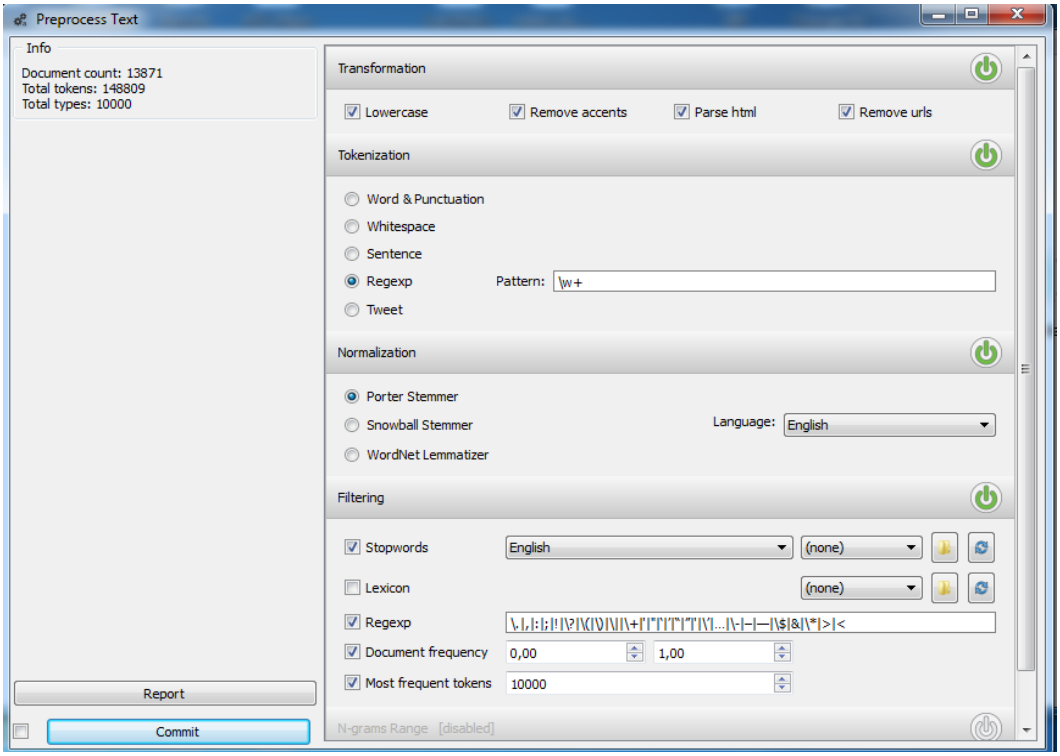
Βήμα 1:

Φορτώνουμε το σύνολο δεδομένων με το Corpus.



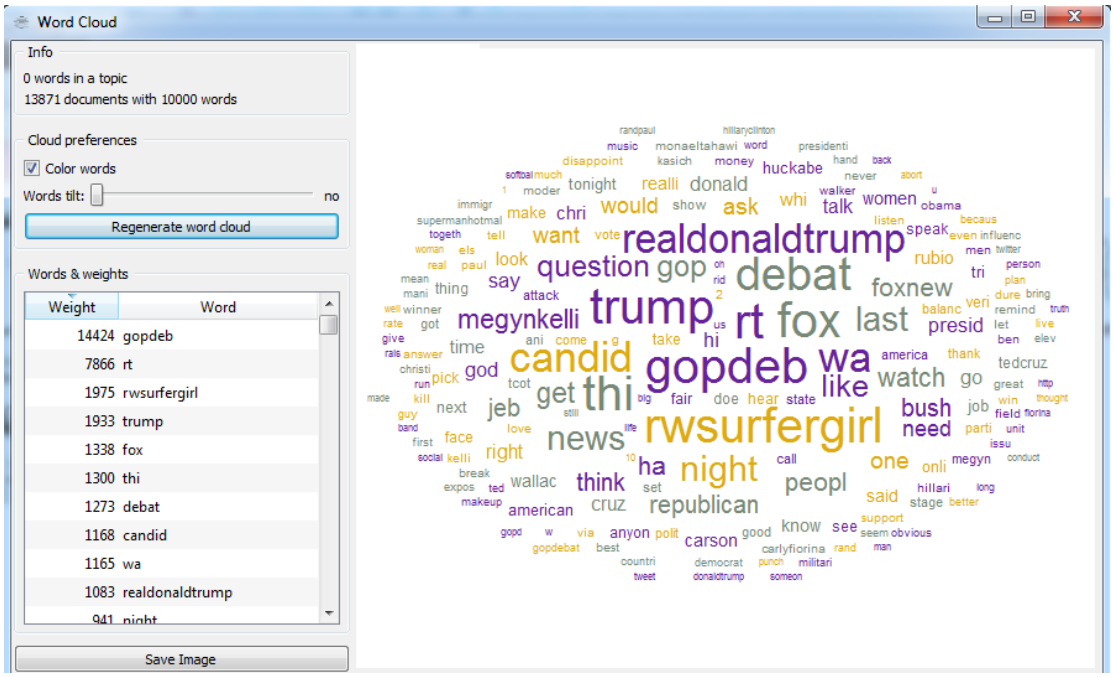
Εικόνα 26: Φόρτωση του συνόλου δεδομένων

Βήμα 2: Περνάμε στο στάδιο της προεπεξεργασίας



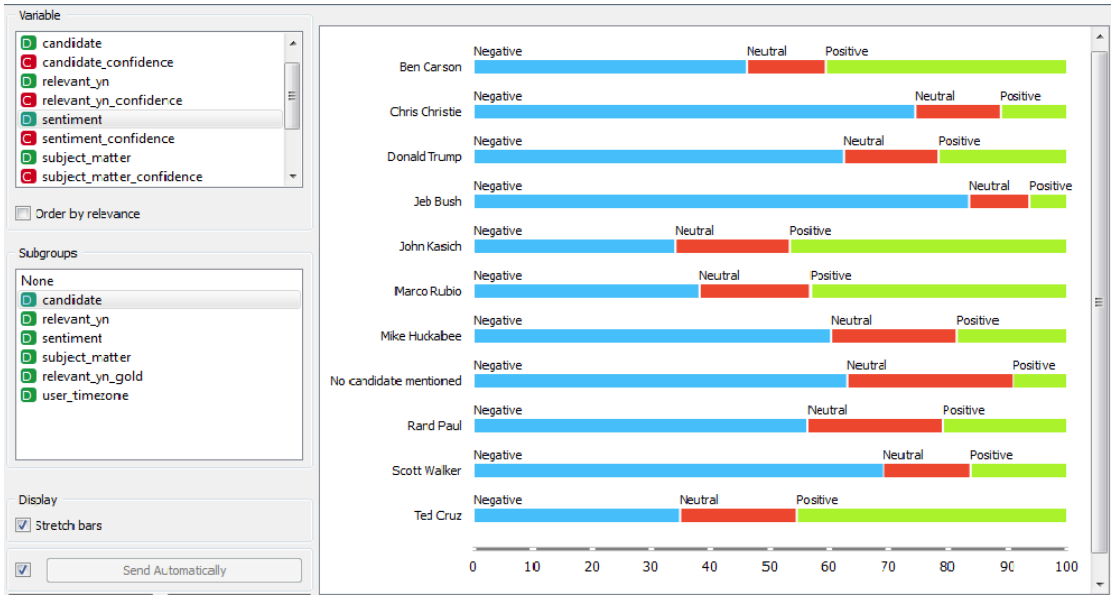
Εικόνα 27: Προεπεξεργασία του συνόλου δεδομένων

Βήμα 3: Αποτυπώνουμε τις λέξεις με τη μεγαλύτερη συχνότητα εμφάνισης με ένα σύννεφο λέξεων.



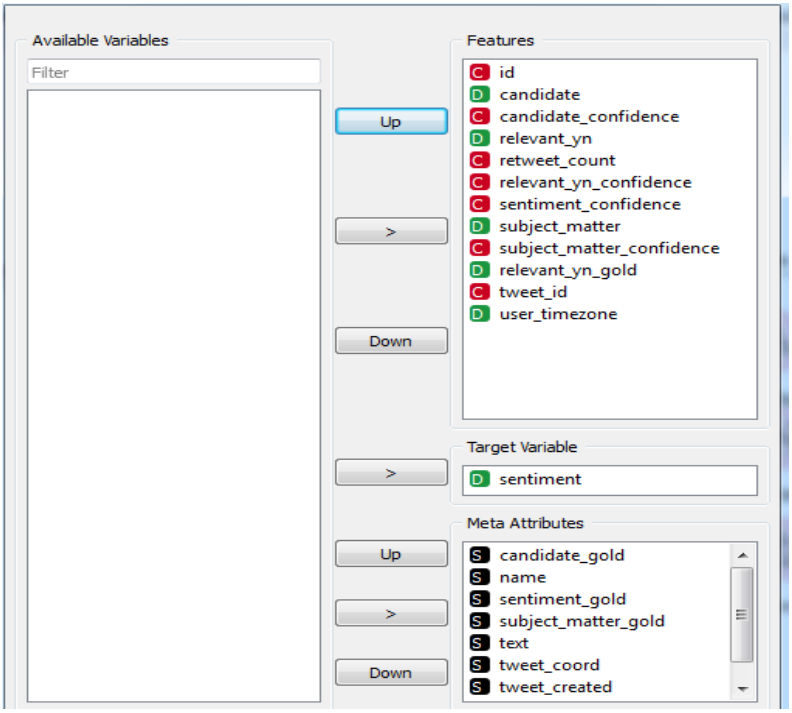
Εικόνα 28: Εφαρμογή Word Cloud στο Σύνολο Δεδομένων

Βήμα 4: Αποτυπώνουμε με τη βοήθεια του Box Plot τη σχέση μεταξύ του χαρακτηριστικού 'sentiment' και του χαρακτηριστικού 'candidate', προκειμένου να δούμε γραφικά την κατανομή σε σχέση με το συναίσθημα(θετικό, αρνητικό, ουδέτερο) των tweets για κάθε υποψήφιο.



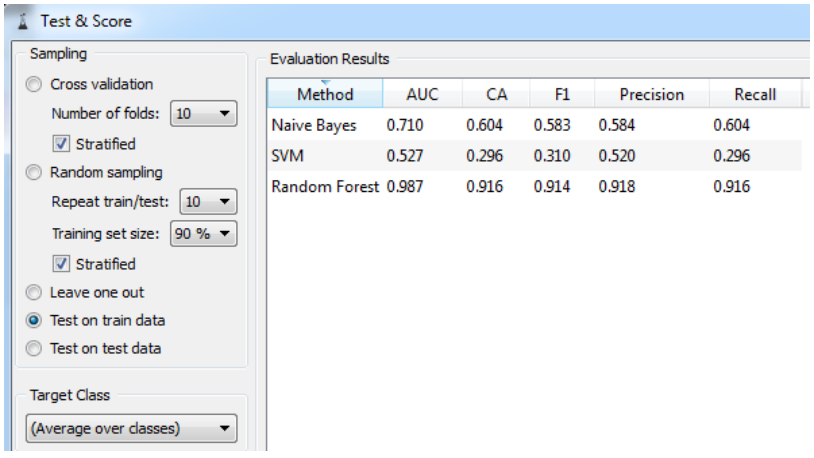
Εικόνα 29: Box Plot για την απεικόνιση του sentiment ανα υποψήφιο

Βήμα 5: Ορίζουμε σα μεταβλητή στόχου το χαρακτηριστικό που θέλουμε, που είναι το sentiment.



Εικόνα 30: Ορισμός μεταβλητής κατηγοριοποίησης

Βήμα 6: Με τη χρήση αλγόριθμων μάθησης (Naïve Bayes, SVM, Random Forest) εκπαιδεύουμε τα δεδομένα μας και φτιάχνουμε ένα μοντέλο πρόβλεψης.



Εικόνα 31: Αποτελέσματα αλγόριθμων μάθησης

Οι παραπάνω εικόνες μας δείχνουν όλη τη διαδικασία που ακολουθήσαμε προκειμένου να πάρουμε τα αποτελέσματα αξιολόγησης των αλγόριθμων για το συγκεκριμένο σύνολο δεδομένων. Παρατηρούμε πως για τον αλγόριθμο Naïve Bayes παίρνουμε σχετικά καλά αποτελέσματα, περίπου στο 70%, ενώ για τον SVM όχι και τόσο, κοντά στο 53%. Εντυπωσιακά είναι τα αποτελέσματα που παίρνουμε βέβαια με τον αλγόριθμο Random Forest με ακρίβεια της τάξης του 98%.

Πολύ σημαντικό είναι να τονίσουμε πως το Orange αποτελεί ένα πολύ γρήγορο και αυτοματοποιημένο εργαλείο, το οποίο είναι και αρκετό φιλικό και σε χρήστες που δεν έχουν εμπειρία με γλώσσες προγραμματισμού. Επίσης, μπορεί να χρησιμοποιηθεί από χρήστες διαφορετικού επιστημονικού φάσματος για την ανάλυση των δεδομένων τους, από Πολιτικούς επιστήμονες μέχρι Βιολόγους κλπ.

Στο κεφάλαιο 7 παρουσιάζονται κάποια χρήσιμα συμπεράσματα από την υλοποίηση που πραγματοποιήσαμε και επίσης θέτουμε σε γενικές γραμμές κάποιες σκέψεις για μελλοντικές προεκτάσεις στη διαδικασία εξόρυξης πληροφορίας από tweets και δημιουργίας μοντέλων ταξινόμησης με τη χρήση ευφυών αλγόριθμων.

7. Συμπεράσματα και μελλοντικές προκλήσεις

7.1 Συμπεράσματα

Η εξέλιξη των κοινωνικών μέσων δικτύωσης στη γενικότερη κοινωνικοπολιτική ζωή παρατηρούμε πως έχει επιδράσει καταλυτικά και στον τρόπο με τον οποίο ασκείται και παράγεται πολιτική. Η ευρεία χρήση του twitter από πολιτικούς σε καθημερινό επίπεδο, αλλά και η αναφορά απλών ανθρώπων σε αυτούς κάνοντας θετικά ή αρνητικά σχόλια έχει δημιουργήσει ένα ζωντανό χώρο μέσα από τον οποίο μεταφέρεται πληροφορία και διαμορφώνονται πεποιθήσεις και συνειδήσεις. Η ανάλυση συναισθήματος αποτελεί ένα κρίσιμο εργαλείο προκειμένου να μπορέσουμε να επεξεργαστούμε ένα τεράστιο όγκο δεδομένων, αλλά και να εξάγουμε συμπεράσματα και εκ νέου πληροφορία.

Στην παρούσα εργασία μελετήσαμε μοντέλα μηχανικής μάθησης που χρησιμοποιούνται στην εξόρυξη κειμένου και στην κατηγοριοποίηση συναισθήματος σε tweets τα οποία αναφέρονται είτε στο Donald Trump είτε στη Hillary Clinton. Συγκεκριμένα, αναφερθήκαμε και αναλύσαμε τους αλγόριθμους Multinomial Naïve Bayes και Support Vector Machines. Στη συνέχεια με βάση ένα σύνολο δεδομένων δημιουργήσαμε δύο μοντέλα πρόβλεψης. Ένα για την επιλογή του υποψήφιου και άλλο ένα για το συναίσθημα που εκφράζεται από το χρήστη. Και για τα δύο μοντέλα πρόβλεψης χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι.

Η γλώσσα που χρησιμοποιείται είναι η Python. Πρόκειται για μια γλώσσα γενικού σκοπού, η οποία όμως είναι πολύ χρήσιμη σε όλα τα μέρη της διαδικασίας. Στην εισαγωγή/εξαγωγή δεδομένων, στην προεπεξεργασία και στο μοντέλο εκπαίδευσης και αξιολόγησης. Παρ’ όλο που η Python δεν αποτελεί τη μοναδική επιλογή, προσφέρει ένα μοναδικό συνδυασμό ευελιξίας, εύκολης ανάπτυξης και απόδοσης.

Παρατηρούμε στα αποτελέσματά μας ότι οι δύο αλγόριθμοι παρουσιάζουν υψηλά ποσοστά ακρίβειας τόσο για την πρόβλεψη του υποψηφίου όσο και για την πρόβλεψη του συναισθήματος(πάνω από 95%). Ο αλγόριθμος Support Vector Machines έχει λίγο καλύτερη απόδοση από τον αντίστοιχο Naïve Bayes.

Ο Naïve Bayes είναι αρκετά πιο γρήγορος από τον SVM, για το λόγω του ότι διαρκεί αρκετή ώρα η εκπαίδευση των συντελεστών για τη γραμμική κατηγοριοποίηση με SVM. Επίσης η εισαγωγή συνόλου λέξεων με θετική και αρνητική πολικότητα, μας βοήθησε πολύ και μας έδωσε πολύ καλά αποτελέσματα σχετικά με τη διαμόρφωση γνώμης για τα tweets. Προκειμένου να αξιολογήσουμε γενικότερα το μοντέλο μας, προσπαθήσαμε να κατηγοριοποιήσουμε κάποια τυχαία tweets, τα οποία ήταν αρκετά δημοφιλή κατά την προεκλογική περίοδο, σε έναν από τους δύο υποψήφιους με την αντίστοιχη πολικότητα. Τα αποτελέσματα που πήραμε ήταν αρκετά ενθαρρυντικά σχετικά με την ακρίβεια και την αξιοπιστία του μοντέλου που δημιουργήσαμε.

Τέλος, δημιουργήσαμε ένα αντίστοιχο μοντέλο κατηγοριοποίησης tweets σε ένα άλλο σύνολο δεδομένων, μέσα από το εργαλείο Orange της Python. Μετά από μια πρώτη γνωριμία με το εργαλείο και με τη χρήση των widgets που μας παρέχει σχετικά με το Text Mining καταφέραμε να φτιάξουμε μια ροή εργασιών που μας έδωσαν σχετικά καλά αποτελέσματα κυρίως με τον αλγόριθμο Naïve Bayes και όχι τόσο με SVM. Ενώ υψηλής ακρίβειας αποτελέσματα πήραμε με τον αλγόριθμο Random Forest.

7.2 Μελλοντική επέκταση

Είναι γνωστό πως τα περισσότερα δεδομένα είναι αδόμητη πληροφορία, άρα πολύ σημαντικό είναι η ανάπτυξη τεχνικών κατηγοριοποίησης μη-επιβλεπόμενης μάθησης. Πολύ σημαντικό είναι να διερευνηθεί η διαδικασία της ομαδοποίησης δεδομένων χωρίς ετικέτες, αλλά και η σημασιολογική ανάλυση η οποία θα βασίζεται στην εξής λογική: Λέξεις που αναφέρονται σε παρόμοια κείμενα/προτάσεις τείνουν να έχουν και παρόμοιες έννοιες. Με βάση αυτήν την ανάλυση είναι σημαντικό να εστιάσουμε στους δύο παρακάτω άξονες:

- Τη βελτιστοποίηση του υπάρχοντος μοντέλου, προκειμένου να μπορούμε να επιτυγχάνουμε μεγαλύτερη ακρίβεια. Αυτό μπορεί να γίνει τόσο με την αξιοποίηση περισσότερων χαρακτηριστικών γνωρισμάτων

που μπορούν να εξαχθούν από tweets, όσο και από τη βελτιστοποίηση των παραμέτρων των αλγόριθμων.

- Την επεκτασιμότητα του υπάρχοντος μοντέλου, προκειμένου να μπορούμε να πάρουμε περισσότερα στοιχεία για το χρήστη(φύλο, ηλικία, μορφωτικό επίπεδο) και να αξιολογήσουμε το προφίλ του.

Σχετικά με το Orange πολύ σημαντικό από εδώ και πέρα, μιας και πρόκειται και για ένα σχετικά καινούριο εργαλείο ανάλυσης δεδομένων, είναι η ευρεία διάδοσή του και η περαιτέρω ανάπτυξη του σε όσο το δυνατόν περισσότερα επιστημονικά αντικείμενα μπορεί να αναφερθεί και να φανεί χρήσιμο.

Βιβλιογραφία

Vallikannu Ramanathan, T. Meyyappan "Survey of Text Mining", International Conference on Technology and Business and Management, March 2013, pp. 508-514.

Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.

Raghavan, V. V. and Wong, S. K. M. *A critical analysis of vector space model for information retrieval.* Journal of the American Society for Information Science, Vol.37 (5), p. 279-87, 1986.

Salton, Gerard. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

Luhn, H. P. *The Automatic Creation of Literature Abstracts.* IBM Journal of Research and Development 2 (2), p. 159-165 and 317, April 1958.

van Rijsbergen, C. J. *Information retrieval.* Butterworths, 1979.

Salton, G. and Buckley, C. (1988) “Term weighting approaches In automatic text retrieval, Information Processing and Management”,Vol. 24, No.5, Pp. 513 - 523.

Diao, Q. and Diao, H. (2000) “Three Term Weighting and Classification Algorithms in Text Automatic Classification”, The Fourth International Conference on High-Performance Computing in the Asia-Pacific Region,Vol. 2, P.629.

Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319- 7242, Vol 3 Issue 5, May 2014, pp.5621-5625

Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.

Xue, X. and Zhou, Z. (2009),“Distributional Features for Text Categorization”, IEEE Transactions on Knowledge and Data Engineering,Vol. 21, No. 3, Pp. 428-442.

http://www.britannica.com/EBchecked/topic/1116194/machine-learning

Phil Simon (March 18, 2013). *Too Big to Ignore: The Business Case for Big Data*. Wiley, σελ. 89. ISBN 978-1-118-63817-0.

Ron Kohavi; Foster Provost (1998). «Glossary of terms». *Machine Learning* **30**: 271–274.

Machine learning and pattern recognition "can be viewed as two facets of the same field."

«Machine Learning: What it is and why it matters». *www.sas.com*

Mitchell, T. (1997). *Machine Learning*, McGraw Hill, *Machine Learning*, McGraw Hill, p.2

Harnad, Stevan (2008), «The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence», στο: Epstein, Robert; Peters, Grace, επιμ., *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Kluwer

K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in *Proceedings of WWW*, pp. 519–528, 2003.

S. Das and M. Chen, “Yahoo! for Amazon: Extracting market sentiment from stock message boards,” in *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

R. M. Tong, “An operational system for detecting and tracking opinions in on-line discussion,” in *Proceedings of the Workshop on Operational Text Classification (OTC)*, 2001.

Peter D. Turney, (2002), Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 417-424

V. S. Jagtap, Karishma Pawar, (2013), Analysis of different approaches to Sentence-Level Sentiment Classification, *International Journal of Scientific Engineering and Technology*, PP : 164-170

S. ChandraKala and C. Sindhu, (2012), Opinion Mining And Sentiment Classification: A Survey, *ICTACT Journal on Soft Computing*, Vol- 03, ISSUE: 01, ISSN: 2229-6956

“New user FAQs,” Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/13920?lang=en. [Accessed: 14-Feb-2016].

“How Much Data Is Generated Every Minute On Social Media?,” WeRSM | We Are Social Media. [Online]. Available: http://wersm.com/how-much-data-is-generatedevery-minute-on-social-media/. [Accessed: 14-Feb-2016].

B. Pang and Lillian Lee, “Opinion mining and sentiment analysis,” [Online]. Available: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf. [Accessed: 21-Apr- 2016]. [Accessed: 15-Feb-2016].

“Company | About,” Twitter About. [Online]. Available: https://about.twitter.com/company. [Accessed: 17-Feb-2016].

M. A. Russell, “Mining the Social Web, Second Edition,” Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472 in 2014, ISBN: 978-1-449-36761-9.

”Twitter”, Andy Murray. [Online]. Available: https://twitter.com/andy_murray/with_replies. [Accessed: 25-Feb-2016]

“Using hashtags on Twitter,” Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/49309. [Accessed: 23-Feb-2016].

“The Twitter glossary,” Twitter Help Center. [Online]. Available: https://support.twitter.com/articles/166337. [Accessed: 20-Feb-2016].

Bing Liu. “Sentiment Analysis and Subjectivity.” A Chapter in Handbook of Natural Language Processing, Second Edition, (editors: N.Indurkhya and F.J.Damerau), 2010

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 115-120). Association for Computational Linguistics.

Chung, J. E., & Mustafaraj, E. (2011, August). Can collective sentiment expressed on twitter predict political elections?. In *AAAI* (Vol. 11, pp. 1770-1771).

Younus, A., Qureshi, M. A., Asar, F. F., Azam, M., Saeed, M., & Touheed, N. (2011, July). What do the average twitterers say: A twitter model for public opinion analysis in the face of major political events. *2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM),*(pp. 618-623). IEEE.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317-332.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social science computer review*, *29*(4), 402-418.

Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society*, *14*(6), 872-895.

Bae, Y., & Lee, H. (2012). Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the Association for Information Science and Technology*, *63*(12), 2521-2535.

Monti, C., Rozza, A., Zappella, G., Zignani, M., Arvidsson, A., & Colleoni, E. (2013, August). Modelling political disaffection from Twitter data. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 3). ACM.

http://www.cbsnews.com/news/donald-trump-and-hillary-clintons-most-popular-tweets-of-2016/

http://www.businessinsider.com/most-popular-election-tweets-from-trump-and-clinton-2016-11

http://www.complex.com/life/2016/06/presidential-candidate-tweets

Janez Demšar; Tomaž Curk; Aleš Erjavec; Črt Gorup; Tomaž Hočevar; Mitar Milutinovič; Martin Možina; Matija Polajnar; Marko Toplak; Anže Starič; Miha Stajdohar; Lan Umek; Lan Žagar; Jure Žbontar; Marinka Žitnik; Blaž Zupan (2013). "Orange: data mining toolbox in Python" (PDF). JMLR. **14** (1): 2349–2353.

R.Sagayam, S.Srinivasan, S.Roshini, “A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques”. Internaltional Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5.

Mr. Rahul Patel,Mr. Gaurav Sharma,"A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319- 7242, Vol 3 Issue 5, May 2014, pp.5621-5625

Dunham Margaret H., “Data Mining Introductory and Andanced Topics”, Pearson Education Inc, (2003).

C. C. Aggarwal and C.-X. Zhai, “Mining Text Data”, New York, NY, USA: Springer, 2012.

Porter, M. (1980) “An algorithm for suffix stripping, Program”,Vol. 14, No. 3, Pp. 130–137

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258.

http://blogs.sas.com/content/sascom/2015/08/11/an-introduction-to-machine-learning/

Liao, X., Cao, D., Tan, S., Liu, Y., Ding, G., and Cheng X.Combining Language Model with Sentiment Analysis for Opinion Retrieval of Blog-Post. Online Proceedings of Text Retrieval Conference (TREC) 2006. http://trec.nist.gov/

Janardhana, Ravikiran. "How to Build a Twitter Sentiment Analyzer." Ravikiranj.net.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In Proceedings of the ACM SIGIR 2010 Poster and Demos. ACM press

https://www.kaggle.com/benhamner/clinton-trump-tweets

Haykin, Simon. Νευρωνικά δίκτυα και μηχανική μάθηση 3η έκδ. - Αθήνα : Παπασωτηρίου, 2010.

Heckerman, D.(1999). Learning in graphical models, chapter A tutorial on learning with Bayesian networks, pages 301–354. MIT Press

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan (2002). “Thumbs up? Sentiment classification using Machine learning Techniques”. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp. 79–86

Gaber, M., Zaslavsky, A., and Krishnaswamy, S. (2007). A survey of classification methods in data streams. In Aggarwal, C., editor, Data Streams, Models and Algorithms, pages 39–59. Springer

http://scikit-learn.org/0.16/tutorial/text_analytics

https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf

Introduction-to-machine-learning (May 2015), Amit Kumar.http://www.allprogrammingtutorials.com/tutorials/introduction-to-machine-learning.php