ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. «Πληροφοριακά Συστήματα και Υπηρεσίες»

Κατεύθυνση «Μεγάλα Δεδομένα και Αναλυτική»

Διπλωματική Εργασία

«140 Χαρακτήρες με Άποψη: Ανάπτυξη συστήματος ανάλυσης συναισθήματος σε Tweet και η σημασία των Emoticon»

Άννα Νικολαρέα

Α.Μ. ΜΕ1609

Επιβλέπουσα Καθηγήτρια: Μαρία Χαλκίδη

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

1.1 ΣΚΟΠΟΣ ΚΑΙ ΣΤΟΧΟΙ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στην παρούσα εργασία γίνεται έρευνα σχετικά με τις προκλήσεις στον τομέα της ανάλυσης συναισθήματος στα μέσα κοινωνικής δικτύωσης. Συγκεκριμένα, επιχειρείται η κατηγοριοποίηση tweet που συλλέγονται από το Twitter σε αρνητικά, θετικά και ουδέτερα με βάση το συναίσθημά τους. Το κύριο πρόβλημα που εντοπίζεται σε αυτό το εγχείρημα είναι ότι τα tweet καλύπτουν ένα τεράστιο φάσμα θεμάτων. Επομένως, το λογικό συμπέρασμα είναι πως η χρήση ενός έτοιμου training set για τέτοιου είδους δεδομένα δεν μπορεί να έχει βέλτιστα αποτελέσματα.

Εξετάζεται λοιπόν η εύρεση ενός συστήματος που θα κατηγοριοποιεί αυτόματα (χωρίς ανθρώπινη παρέμβαση) τα Tweet ανά Hashtag, δηλαδή ανά θέμα, ώστε να δημιουργείται κάθε φορά ένα training set, που θα είναι στοχευμένο και πιο έγκυρο.

Θεωρήθηκε λογική η υπόθεση πως ένα tweet μπορεί να χαρακτηριστεί ως θετικό, αρνητικό ή ουδέτερο ανάλογα με το συναίσθημα που αντιπροσωπεύουν τα emoticon που περιλαμβάνει. Σε πολλές έρευνες έχει γίνει κατηγοριοποίηση κειμένων με emoticon, αλλά στις περισσότερες οι ερευνητές χαρακτηρίζουν οι ίδιοι το συναίσθημα των emoticon με βάση το τι αντιλαμβάνονται όταν τα βλέπουν.

Γι’ αυτό αναζητήθηκε μια πιο αντικειμενική λύση και χρησιμοποιήθηκε ένα emoji sentiment lexicon (Petra Kralj Novak, Jasmina Smailović, Borut Sluban, Igor Mozetič, 2015)[1]. Για να φτιάξουν το λεξικό, οι ερευνητές υπολόγισαν το συναίσθημα των emoticon με βάση το συναίσθημα των tweet στα οποία βρέθηκαν και έφτιαξαν ένα χάρτη των 751 πιο συχνών.

Οι παραπάνω παραδοχές αποτελούν τη θεωρητική βάση της εργασίας. Το κύριο τμήμα της σχετίζεται με το κατά πόσο ένα τέτοιο σύστημα μπορεί να έχει μεγάλη ακρίβεια ώστε να χρησιμοποιηθεί για την κατηγοριοποίηση tweet ανάλογα με το θέμα τους.

2 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΑΙ Ο ΡΟΛΟΣ ΤΩΝ EMOTICON

2.1 ΣΥΝΤΟΜΗ ΑΝΑΦΟΡΑ

Η ανάλυση και η εξόρυξη δεδομένων από κείμενα έχει ορισθεί με πολλούς διαφορετικούς τρόπους: ανάλυση συναισθήματος, εξόρυξη γνώμης, ανάλυση υποκειμενικότητας, κ.ά.

Σύμφωνα με τους Pang και Lee [2] ωστόσο, «η «ανάλυση συναισθήματος» και η «εξόρυξη γνώμης» υποδηλώνουν το ίδιο πεδίο μελέτης (το οποίο μπορεί να θεωρηθεί ως υποπεδίο της ανάλυσης υποκειμενικότητας).»

Σε αυτή τη διπλωματική, η ανάλυση συναισθήματος αναφέρεται στην κατηγοριοποίηση ενός κειμένου ως προς την πολικότητά του, δηλαδή το κατά πόσο είναι θετικό, αρνητικό ή ουδέτερο.

2.2 ΤΟ ΚΕΙΜΕΝΟ ΩΣ ΔΕΔΟΜΕΝΟ ΑΝΑΛΥΣΗΣ

Κάθε δεδομένο είναι εκ φύσεως φορέας πληροφορίας. Σε ό, τι αφορά τα δεδομένα κειμένου η πληροφορία δεν είναι πάντα εύκολο να αποκρυπτογραφηθεί, κι αυτό γιατί ο γραπτός λόγος περιλαμβάνει φανερά (explicit) και κρυφά (implicit) δεδομένα [1].

Η αποκωδικοποίηση του νοήματος του γραπτού ή προφορικού λόγου συμπεριλαμβάνει την ανάλυση των λέξεων, αλλά χρειάζεται επιπλέον διερμηνεία για να προσεγγιστούν οι προθέσεις του ατόμου. Ο λόγος είναι πως οι λέξεις από μόνες τους φέρουν μια έννοια, που όμως μπορεί να μεταλλαχθεί πολύ εύκολα αναλόγως με τα συμφραζόμενα. Με λίγα λόγια, πολλές φορές παρατηρείται χάσμα μεταξύ της πληροφορίας που θέλει να μεταδώσει ένας πομπός και της τελικής έκφρασης που χρησιμοποιεί. Μία γραπτή φράση μπορεί να μη μεταφέρει στο μέγιστο τις σκέψεις ή να υπονοεί πληροφορίες.

Ο τομέας της ανάλυσης δεδομένων έχει ασχοληθεί ιδιαιτέρως με τις μεθόδους εξόρυξης πληροφορίας από κείμενα. Πόσο εύκολο είναι όμως να ανακαλύψεις γεγονότα ή σημασίες από ένα κείμενο χωρίς ανθρώπινη επίβλεψη; Η χρήση του συντακτικού μπορεί να προσφέρει δομή στην έκφραση του γραπτού και προφορικού λόγου, αλλά οι κανόνες που εφαρμόζει δεν είναι

αυστηροί κι αυτό γιατί ο λόγος έχει μια αυτονομία που τον καθιστά περίπλοκο στην ανάλυση. Συγκεκριμένα στα μέσα κοινωνικής δικτύωσης, τα κείμενα έχουν «θόρυβο». Αυτό σημαίνει ότι μπορεί να περιέχουν εκφράσεις argo, να είναι δυσανάγνωστα ή και ανορθόγραφα.

Τη λύση στη μελέτη και ανάλυση τέτοιων κειμένων επιχειρεί να δώσει η ανάλυση συναισθήματος ή αλλιώς η εξόρυξη γνώμης (opinion mining). Με λίγα λόγια με αυτή την ανάλυση προσπαθούμε να ανακαλύψουμε τα συναισθήματα και τις προθέσεις που είχε κάποιος όταν έγραφε ένα κείμενο.

Ο όρος opinion mining εμφανίστηκε πρώτη φορά το 2003 σε μία έρευνα των Dave κ.ά., που δημοσιεύθηκε στα πλαίσια του συνεδρίου WWW. [30] Νωρίτερα είχε εμφανιστεί ο όρος «Ανάλυση Συναισθήματος» με πρώτη αναφορά το 2001 σε άρθρα των Das, Chen και Tong [28,29], οι οποίοι επιχείρησαν να αναλύσουν τις προθέσεις της αγοράς (Bo Pang, Lillian Lee, 2008).

2.3 ΔΙΑΦΟΡΕΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Σε όλη τη βιβλιογραφία παρατηρείται συχνά η σύγχυση της εξόρυξης γνώμης με την ανάλυση συναισθήματος, αν και η λέξη γνώμη απέχει από τη λέξη συναίσθημα. Σύμφωνα με το λεξικό του Μπαμπινιώτη [4] η λέξη γνώμη ορίζεται ως αυτό που πιστεύει, που νομίζει κανείς για συγκεκριμένο ζήτημα. Αντιθέτως, η λέξη συναίσθημα ορίζεται ως η ιδιαίτερη ψυχική διάθεση στην οποία περιέρχεται κανείς από ποικίλες καταστάσεις, οι οποίες οφείλονται είτε σε εξωτερικά ερεθίσματα είτε σε λειτουργίες του οργανισμού. Το συναίσθημα δηλαδή είναι κάτι που αισθανόμαστε, ενώ μια άποψη/γνώμη είναι η λογική σκέψη μας περί ενός θέματος. Οι δύο λέξεις έχουν μεν κάποια κοινά γνωρίσματα, αλλά και διακριτές διαφορές.

Στις περισσότερες των περιπτώσεων μία άποψη εμπεριέχει συναίσθημα που μπορεί να διακριθεί σε θετικό, αρνητικό και ουδέτερο. Ίσως γι’ αυτό το λόγο έχει επικρατήσει ο όρος Ανάλυση Συναισθήματος και χρησιμοποιείται πιο συχνά από τον όρο «Εξόρυξη Γνώμης».

Στη βιβλιογραφία ο όρος Sentiment Analysis ίσως χρησιμοποιείται πιο συχνά όταν αναφέρεται η συστηματική ανάλυση του κατά πόσο ένα κείμενο αντιπροσωπεύει έναν θυμωμένο, χαρούμενο, λυπημένο, κ.ά., πομπό. Ενώ η εξόρυξη γνώμης παρατηρείται πιο συχνά

σε δημοσιεύσεις που αναφέρονται στην ανάλυση του κατά πόσο ένα κείμενο φέρει μια θετική, αρνητική ή ουδέτερη άποψη.

Η Ανάλυση Συναισθήματος μπορεί να γίνει σε επίπεδο κειμένου, πρότασης ή μεμονωμένου χαρακτηριστικού.

1. Κείμενο: Σε αυτό το επίπεδο η κύρια πρόκληση είναι να συγκεντρωθούν οι προτάσεις αυτές που θα καθορίσουν το συναίσθημα ολόκληρου του κειμένου. [5]
2. Πρόταση: Με αυτόν τον τρόπο η πρόταση αποτελεί μία ξεχωριστή οντότητα που φέρει θετικό, αρνητικό ή ουδέτερο συναίσθημα. [5]
3. Χαρακτηριστικό: Οι λέξεις μίας πρότασης γίνονται χαρακτηριστικά της. [6]

2.4 ΛΕΞΙΚΟΓΡΑΦΙΚΕΣ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Οι λεξικογραφικές μέθοδοι Ανάλυσης Συναισθήματος βασίζονται σε λεξικά που κατατάσσουν τις λέξεις με βάση το συναίσθημα που φέρουν. Ένα από τα πιο χαρακτηριστικά παραδείγματα λεξικών είναι το Sentiwordnet (Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, 2010; 7). Ένα άλλο λεξικό είναι το Sentiwords, το οποίο περιέχει περίπου 155.000 αγγλικές λέξεις με ένα σκορ μεταξύ του -1 και του 1. [8]

Πολλά περισσότερα τέτοια λεξικά είναι διαθέσιμα και μάλιστα πολλοί ερευνητές έχουν ασχοληθεί με τις μεθόδους δημιουργίας τους. Για παράδειγμα, οι Hatzivassiloglou και McKeown (1997) πρότειναν μια μέθοδο αυτόματης αναγνώρισης της πολικότητας των επιθέτων. [9]

2.5 ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Οι μέθοδοι μηχανική μάθησης έχουν τραβήξει το ενδιαφέρον των ερευνητών παγκοσμίως και στον τομέα της Ανάλυσης Συναισθήματος έχουν εξετασθεί ιδιαιτέρως. Η ακρίβεια των αποτελεσμάτων είναι ένας από τους βασικούς λόγους που προτιμώνται και εξετάζοντας τη βιβλιογραφία παρατηρούμε συνήθως μεθόδους επιβλεπόμενης μάθησης.

Η διαδικασία μάθησης περιλαμβάνει τη συλλογή των δεδομένων, την προ-επεξεργασία τους, την εκπαίδευση του μοντέλου, την κατηγοριοποίηση και την παρουσίαση των αποτελεσμάτων.

Μία από τις πρώτες και πιο αντιπροσωπευτικές έρευνες για τα αποτελέσματα αυτών των μεθόδων στην Ανάλυση Συναισθήματος έγινε από τον Pang κ.ά. [10], οι οποίοι επιχείρησαν να κατηγοριοποιήσουν κριτικές ταινιών σε θετικές και αρνητικές χρησιμοποιώντας Naïve Bayes και Support Vector Machines. Οι συγκεκριμένοι ερευνητές χρησιμοποίησαν unigrams για να φτιάξουν το feature vector τους, δηλαδή μεμονωμένες λέξεις.

Σε ό,τι αφορά τις μεθόδους μηχανικής μάθησης, ένας από τους πιο σημαντικούς παράγοντες είναι, όπως είναι φυσικό, τα δεδομένα που παρέχονται στον classifier και η αναπαράστασή τους.

Οι Pang και Lee [2] δημοσίευσαν το 2008 ένα άρθρο που αναλύει ακόμα περισσότερο τον τομέα της ανάλυσης και σχολίασαν τους διάφορους τρόπους αναπαράστασης ενός κειμένου για τους σκοπούς της μηχανικής μάθησης.

Πιο συγκεκριμένα αναφέρθηκαν στα εξής:

1. Παρουσία λέξεων (Boolean Method)
2. Συχνότητα λέξεων (TF-IDF)
3. Η Θέση των λέξεων (n-grams)
4. Part of Speech (POS)
5. Συντακτικό
6. Άρνηση
7. Features με βάση το θέμα

2.6 ΤΑ EMOTICON ΣΤΟΝ ΚΟΣΜΟ ΤΩΝ SOCIAL MEDIA

Με το πέρασμα του χρόνου ο τρόπος που επικοινωνούμε μεταξύ μας έχει αλλάξει σε πολύ μεγάλο βαθμό. Περάσαμε από την εποχή που ένα γράμμα καθόριζε το αν θα μπορείς να μάθεις τα νέα κάποιου κοντινού σου προσώπου, στο τηλέφωνο και σήμερα στο διαδίκτυο, το οποίο συνδυάζει το ακουστικό, οπτικό και γραπτό γνώρισμα της επικοινωνίας.

Το διαδίκτυο επανεφηύρε την επικοινωνία με πολλούς τρόπους. Ένας από αυτούς είναι τα emoticon που εξελίχθηκαν σε Emoji, που είναι ο πιο σύντομος τρόπος να εκφραστεί ένα μήνυμα χωρίς τη χρήση λέξεων. Αναπόφευκτα έγινε δημοφιλές σε όλες τις γενιές που χρησιμοποιούν τα σύγχρονα μέσα ανταλλαγής μηνυμάτων και εξαπλώθηκε σε όλους τους τομείς. Δεν είναι μόνο ένας «παιχνιδιάρικος» τρόπος να τονίσεις τα συναισθήματά σου για ένα συγκεκριμένο θέμα, αλλά μερικές φορές είναι απαραίτητα για να μη παρερμηνευθούν προθέσεις και πληροφορίες. Ο γραπτός λόγος πολλές φορές δεν είναι ακριβής, τουλάχιστον σε σύγκριση με τον προφορικό.

Η επικοινωνία πρόσωπο με πρόσωπο είναι πιο ακριβής και πιο άμεση, γιατί μεταφέρει διακριτά στοιχεία, όπως οι εκφράσεις του προσώπου, ο τόνος της φωνής και οι κινήσεις του σώματος. Το μήνυμα είναι πιο ξεκάθαρο όταν όλα αυτά τα στοιχεία είναι διαθέσιμα. Από την άλλη στο γραπτό λόγο οι λέξεις δεν αρκούν πάντα για να κατανοηθεί το πλήρες φάσμα της πληροφορίας. Πολύ σημαντικό παράδειγμα που επιβεβαιώνει τα παραπάνω είναι ο σαρκασμός. Είναι πολύ δύσκολο να διαχωρίσεις ένα σαρκαστικό μήνυμα από ένα κυριολεκτικό, αν δεν έχεις μπροστά σου τον μεταδότη του μηνύματος. Τα emoticon βοηθούν στο να κάνουν τα μηνύματα καλύτερα αντιληπτά και να ξεκαθαρίζουν την πληροφορία.

Η χρήση τους επομένως, δεν έχει μείνει στα στάδια ενός εφηβικού trend αλλά έχει μετατραπεί σε νόρμα, ειδικά στα μέσα κοινωνικής δικτύωσης. Συγκεκριμένα στο Twitter, που το μήνυμα δεν μπορεί να έχει περισσότερους από 140 χαρακτήρες, τα emoticon χρησιμοποιούνται συνεχώς για να τονίσουν μια πρόταση, να αναδείξουν το συναίσθημα του χρήστη, να δείξουν ότι το μήνυμα είναι ειρωνικό ή και για να συμπληρώσουν κάτι που οι λέξεις δεν έφταναν να μεταφέρουν.

2.7 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΜΕ EMOTICON ΚΑΙ Ο ΡΟΛΟΣ ΤΟΥ «EMOJI SENTIMENT LEXICON»

Πολλές έρευνες σχετικά με την ανάλυση συναισθήματος με emoji έχουν γίνει, κάποιες με καλά και κάποιες με μέτρια αποτελέσματα.

Η συγκεκριμένη προσέγγιση εισήχθη πρώτα από τον Read [11]. Στην έρευνά του χρησιμοποίησε μία λίστα από emoticon, για την κατηγοριοποίηση των tweet. Συγκέντρωσε 2000 άρθρα που περιείχαν emoticon με χαμόγελο και 2000 με θλιμμένο emoticon, ώστε να βελτιστοποιήσει το test set. Η μέση ακρίβεια που πέτυχε ήταν 61,5% με τον Naïve Bayes και 70,1% με τον SVM.

Σε μία έρευνα των Hao Wang και Jorge A. Castanon [12], αποδείχθηκε ότι η χρήση emoticon για την κατηγοριοποίηση έχει μεγάλη ακρίβεια της τάξεως του 78%. Για την κατηγοριοποίηση χρησιμοποίησαν emoticon τα οποία τα χαρακτήρισαν άνθρωποι.

Οι Soroush Vosoughi και Helen Zhou Deb Roy [13], είχαν μία πρωτότυπη ιδέα για ανάλυση συναισθήματος σε Tweets. Για το training set χρησιμοποίησαν εκτός από το κείμενο και την περιοχή, την ώρα και τον χρήστη. Τα κατηγοριοποίησαν χρησιμοποιώντας συνολικά έξι emoticons, τρία θετικά και τρία αρνητικά. Βρήκαν περισσότερα από 120 θετικά και αρνητικά emoticons σε ASCII και emojis, αλλά αποφάσισαν να χρησιμοποιήσουν μόνο τα έξι πιο κοινά. Υποστήριξαν ότι δεν υπάρχουν "ουδέτερα" emoticons, και περιόρισαν τις κατηγορίες σε θετικά ή αρνητικά tweets. Το accuracy που πέτυχαν ήταν 86,2% και επιπλέον έκαναν κάποιες πολύ ενδιαφέρουσες παρατηρήσεις. Εντόπισαν ότι τα tweet που δημοσιεύτηκαν κοντά στην Παρασκευή και το Σάββατο, ήταν πιο θετικά και παρατήσαν πτώση του συναισθήματος την Κυριακή.

Οι Jichang Zhao, Li Dong, Junjie Wu, και Ke Xu [14], υλοποίησαν το MoodLens, ένα σύστημα ανάλυσης συναισθήματος για tweets στην Κινεζική γλώσσα στο Weibo. Για την κατηγοριοποίηση των tweet χρησιμοποίησαν 95 emoticon, τα οποία τα χώρισαν σε κατηγορίες συναισθημάτων, για παράδειγμα θλίψη, χαρά, θυμός, κ.ά. Για την κατηγοριοποίηση χρησιμοποίησαν Naïve Bayes.

Οι Dayalani και Patil [15], χρησιμοποίησαν επίσης το συναίσθημα που εκφράζουν τα emoticons για να κατηγοριοποιήσουν τα δεδομένα.

Οι Alec Go, Richa Bhayani, Lei Huang [16], απέδειξαν ότι η χρήση των emoticon ως θορυβώδεις κατηγορίες για τα δεδομένα εκπαίδευσης είναι ένας αποτελεσματικός τρόπος κατηγοριοποίησης. Σύμφωνα με τους ίδιους, «οι αλγόριθμοι μηχανικής μάθησης (Naive Bayes, Maximum entropy, και SVM) μπορούν να επιτύχουν υψηλή ακρίβεια για την ταξινόμηση του συναισθήματος κατά τη χρήση αυτής της μεθόδου. Παρόλο που τα tweet έχουν μοναδικά χαρακτηριστικά σε σύγκριση με άλλα δεδομένα, οι αλγόριθμοι μηχανικής μάθησης δείχνουν ότι κατηγοριοποιούν το συναίσθημα με παρόμοιες επιδόσεις.

Το ζήτημα που δεν έχει προσεγγιστεί πλήρως από παρόμοιες έρευνες είναι ότι αυτό που αντιλαμβανόμαστε ή νιώθουμε όταν βλέπουμε ένα emoticon, είναι πολλές φορές υποκειμενικό. Πραγματικά, ένα emoticon με μια καρδιά που είναι σπασμένη στα δύο για παράδειγμα, μπορεί να χρησιμοποιηθεί και σε θετικά και σε αρνητικά μηνύματα.

«Το τέλος της ταινίας με στενοχώρησε.💔»
«Πέρασα τέλεια στις διακοπές. Γυρνάω Αθήνα για δουλειά! Αλλά δεν πειράζει γιατί και του χρόνου εκεί θα είμαι! 💔»

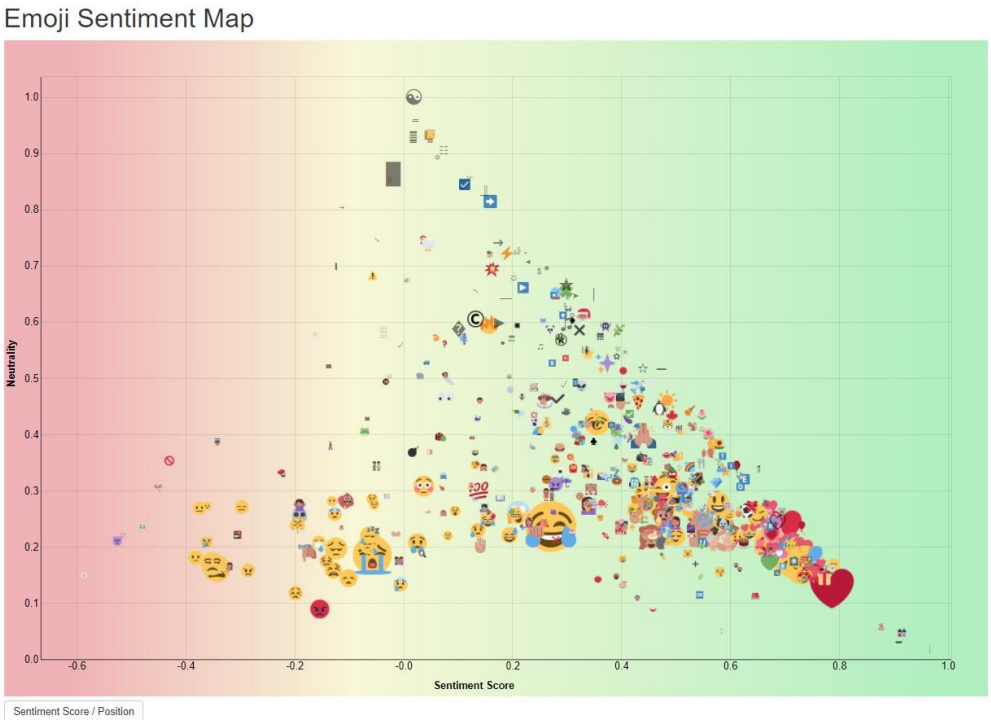
Για τους σκοπούς της εργασίας, θέλαμε να κατηγοριοποιήσουμε αυτόματα τα tweet σε θετικά και αρνητικά με βάση τα emoticon που περιέχουν. Αυτό μπορεί να ακούγεται σαν ένα εύκολο εγχείρημα, αλλά δεν είναι.

Για παράδειγμα το emoji που κλαίει, ☹, μπορεί να πει κανείς ότι είναι αντικειμενικά αρνητικό, αλλά όπως και στο παράδειγμα με την ραγισμένη καρδιά παραπάνω, βρίσκεται και σε μηνύματα με μεικτά ή και θετικά συναισθήματα.

Τέθηκε επομένως το ερώτημα του πόσο αρνητικό είναι το emoji που κλαίει ☹ από το -1 έως το 1. Για να αποφευχθεί μια υποκειμενική απάντηση σε αυτή την ερώτηση, αναζητήθηκε ένα λεξικό συναισθημάτων για τα emoji και πράγματι βρέθηκε μια έρευνα με την ονομασία “Sentiment of Emojis”. Στη συγκεκριμένη έρευνα, οι αρθρογράφοι έφτιαξαν ένα χάρτη για να καταγράψουν το συναίσθημα που έχουν τα 751 πιο δημοφιλή emoji στο twitter.

Πως το έκαναν: «Το συναίσθημα των emoji υπολογίζεται από το συναίσθημα των tweets. Ένα μεγάλο σύνολο tweets, σε 13 ευρωπαϊκές γλώσσες, χαρακτηρίστηκε ως προς συναίσθημα

από 83 άτομα στη μητρική τους γλώσσα. Τα labels για το συναίσθημα μπορούσαν να λάβουν μία από τις τρεις τιμές: αρνητική ≺ ουδέτερη ≺ θετική. Ένα label συναισθήματος, c, είναι τυπικά μια διακριτή μεταβλητή c ∈ {-1, 0, +1}. Ένα emoji παίρνει την τιμή του συναισθήματός του από όλα τα tweets στα οποία εμφανίζεται. Αρχικά, για κάθε emoji, διαμορφώθηκε μια διακριτή κατανομή πιθανότητας (p-, p0, p +). Το sentiment score των emoji υπολογίζεται έπειτα ως ο μέσος όρος της κατανομής. Τα υπόλοιπα στοιχεία της κατανομής, δηλ. P-, p0, και p + υποδηλώνουν την αρνητικότητα, την ουδετερότητα και τη θετικότητα του emoji, αντίστοιχα. Η πιθανότητα Pc υπολογίζεται από τον αριθμό των εμφανίσεων, N, των emoji σε tweets με το label c.» [1]



Εικόνα 1η[1]: Ο χάρτης των 751 Emoji.

[1] http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html

Emoji Sentiment Ranking v1.0

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
😂	😂	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
❤	❤	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
♥	♥	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
😍	😍	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
😭	😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
😘	😘	0x1f618	3848	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons
😊	😊	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES	Emoticons
👌	👌	0x1f44c	2925	0.805	0.094	0.249	0.657	0.563		OK HAND SIGN	Miscellaneous Symbols and Pictographs
💕	💕	0x1f495	2400	0.766	0.042	0.285	0.674	0.632		TWO HEARTS	Miscellaneous Symbols and Pictographs
👏	👏	0x1f44f	2336	0.787	0.104	0.271	0.624	0.520		CLAPPING HANDS SIGN	Miscellaneous Symbols and Pictographs
😁	😁	0x1f601	2189	0.796	0.127	0.296	0.577	0.449		GRINNING FACE WITH SMILING EYES	Emoticons
☺	☺	0x263a	2062	0.799	0.062	0.218	0.720	0.657		WHITE SMILING FACE	Miscellaneous Symbols
♡	♡	0x2661	1975	0.764	0.052	0.227	0.721	0.669		WHITE HEART SUIT	Miscellaneous Symbols
👍	👍	0x1f44d	1854	0.812	0.115	0.248	0.637	0.521		THUMBS UP SIGN	Miscellaneous Symbols and Pictographs
😩	😩	0x1f629	1808	0.826	0.591	0.186	0.223	-0.368		WEARY FACE	Emoticons
🙏	🙏	0x1f64f	1539	0.794	0.081	0.421	0.498	0.417		PERSON WITH FOLDED HANDS	Emoticons
✌	✌	0x270c	1534	0.790	0.113	0.310	0.576	0.463		VICTORY HAND	Dingbats
😏	😏	0x1f60f	1522	0.765	0.112	0.444	0.444	0.332		SMIRKING FACE	Emoticons
😉	😉	0x1f609	1521	0.845	0.100	0.337	0.563	0.463		WINKING FACE	Emoticons
🙌	🙌	0x1f64c	1506	0.791	0.101	0.238	0.661	0.559		PERSON RAISING BOTH HANDS IN CELEBRATION	Emoticons

Εικόνα 2[η2]: Τα πρώτα είκοσι Emoji του λεξικού.

[2] http://kt.ijs.si/data/Emoji_sentiment_ranking/

2.8 ΘΕΤΙΚΑ ΚΑΙ ΑΡΝΗΤΙΚΑ ΤΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΜΕ EMOTICON

Ανάμεσα στα πλεονεκτήματα αυτής της τεχνικής είναι η ταχύτητα αφού δεν χρειάζεται ανθρώπινη επίβλεψη, η δημιουργία training set για διαφορετικά θέματα χωρίς περαιτέρω επεξεργασία και η δημιουργία κλάσης που δεν διακρίνεται απαραίτητα στις τιμές αρνητικό, θετικό και ουδέτερο. Με τη χρήση του Emoji Sentiment Lexicon η κλάση μπορεί να πάρει τιμές από το -1 έως το 1, κάτι που αφήνει στον ερευνητή την ευχέρεια να επιλέξει το εύρος που χαρακτηρίζει ένα tweet ως αρνητικό, θετικό ή ουδέτερο. Αυτό επιτρέπει και τη χρήση Linear Regression, πέρα από αλγορίθμους μηχανικής μάθησης, για την πρόβλεψη της τιμής ενός νέου tweet.

Ένα από τα βασικά μειονεκτήματα που διακρίνονται είναι το να προσδιοριστεί το εύρος της ουδέτερης ζώνης. Αναλύοντας το λεξικό, παρατηρείται πως τα πιο χρησιμοποιημένα emoticon που δηλώνονται ως ουδέτερα παίρνουν τιμές από το -0.093 έως το 0.139. Ωστόσο, υπάρχουν και άλλα emoticon με μεγαλύτερη πιθανότητα ουδετερότητας. Για να υπολογίσουμε το εύρος τιμών της ουδέτερης ζώνης, θεωρήσαμε πως πρέπει να βρεθεί η μέση τιμή sentiment score των emoticon που έχουν πάνω από 75% πιθανότητα να είναι ουδέτερα. Η μέση τιμή αυτή είναι 0,052, οπότε το εύρος των τιμών τέθηκε ως -0,52 έως +0,052.

😭	😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
☯	☯	0x262f	992	0.384	0.006	0.987	0.007	0.001		YIN YANG	Miscellaneous Symbols
😳	😳	0x1f633	846	0.797	0.327	0.327	0.345	0.018		FLUSHED FACE	Emoticons
😢	😢	0x1f622	749	0.814	0.384	0.225	0.391	0.007		CRYING FACE	Emoticons
😴	😴	0x1f634	718	0.850	0.422	0.237	0.341	-0.080		SLEEPING FACE	Emoticons
🔥	🔥	0x1f525	651	0.616	0.124	0.613	0.263	0.139		FIRE	Miscellaneous Symbols and Pictographs
💯	💯	0x1f4af	637	0.872	0.281	0.317	0.402	0.120		HUNDRED POINTS SYMBOL	Miscellaneous Symbols and Pictographs

©		0xa9	416	0.740	0.131	0.621	0.248	0.117		COPYRIGHT SIGN	Latin-1 Supplement
		0x1f4a5	329	0.587	0.072	0.708	0.220	0.148		COLLISION SYMBOL	Miscellaneous Symbols and Pictographs

Εικόνα 3η[3]: Μερικά από τα πιο συχνά Emoji που χαρακτηρίζονται ως ουδέτερα.

[3] http://kt.ijs.si/data/Emoji_sentiment_ranking/

3 ΥΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

3.1 ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΕΞΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER

3.1.1 ΔΕΔΟΜΕΝΑ ΣΕ 140 ΧΑΡΑΚΤΗΡΕΣ

Το Twitter αποτελεί μία από τις πιο διαδεδομένες πλατφόρμες επικοινωνίας και διακρίνεται από την αμεσότητα και τη συντομία της πληροφορίας που διακινεί. Οι χρήστες γράφουν για τη ζωή, τις απόψεις, τις ανησυχίες τους, αλλά κυρίως διαδίδουν το συναίσθημά τους για τα δημοφιλή θέματα της επικαιρότητας.

Παρατηρείται ότι οι περισσότεροι χρήστες του διαδικτύου έχουν στραφεί από τα blogs και τα mail σε πλατφόρμες άμεσης και μαζικής επικοινωνίας γιατί είναι ο πιο εύκολος, πιο γρήγορος και πιο αποτελεσματικός τρόπος να μάθεις κάτι ή να μεταδώσεις μια γνώση, άποψη ή ιδέα. Γι’ αυτούς τους λόγους το Twitter έχει μετατραπεί σε μία ανοιχτή και πολύτιμη πηγή συλλογής της κοινής γνώμης.

Τα δεδομένα που εξάγονται από το Twitter μπορούν να αναλυθούν και να χρησιμοποιηθούν για marketing, για την πολιτική και για τις ειδήσεις. Πολλές φορές δεν χρειάζεται να διαβάσει κανείς ειδησεογραφικά site, αρκεί να συνδεθεί με το λογαριασμό του στο Twitter: τα φλέγοντα θέματα φτάνουν πιο γρήγορα εκεί λόγω του αριθμού των χρηστών που αναφέρονται σε αυτά. Μάλιστα, το Twitter παρουσιάζει τα δημοφιλή Hashtag της ημέρας, δηλαδή τα Hashtag που έχουν χρησιμοποιηθεί περισσότερο. Έτσι, φαίνεται με τι ασχολείται ο περισσότερος κόσμος για ένα συγκεκριμένο χρονικό διάστημα. Τα trending Hashtag συνήθως περιορίζονται ανά περιοχή και χώρα, οπότε ένας χρήστης από την Ελλάδα θα βλέπει τα δημοφιλή Hashtag κοντά του.

3.1.2 Το API του Twitter

Ο λόγος που επιλέχθηκε το Twitter για τούς σκοπούς αυτής της εργασίας είναι βασικά το γεγονός ότι έχει ανοιχτό API. Το API του Twitter αποτελείται από ένα σύνολο URL που λαμβάνουν παραμέτρους, είναι με λίγα λόγια ένας τρόπος να κάνεις query και να ζητήσεις δεδομένα με βάση συγκεκριμένες δεσμεύσεις.

Για να αποκτήσει κανείς πρόσβαση στο API πρέπει να κάνει register μια εφαρμογή κι αυτό γιατί οι εφαρμογές έχουν πρόσβαση σε δημόσιες πληροφορίες του Twitter. Αφού δημιουργηθεί η εφαρμογή το Twitter δίνει authentication και access tokens που μπορούν να χρησιμοποιηθούν για τη σύνδεση και αλληλεπίδραση με το API[4].

Τα endpoint που προσφέρει το Twitter API είναι πέντε και διακρίνονται σε:

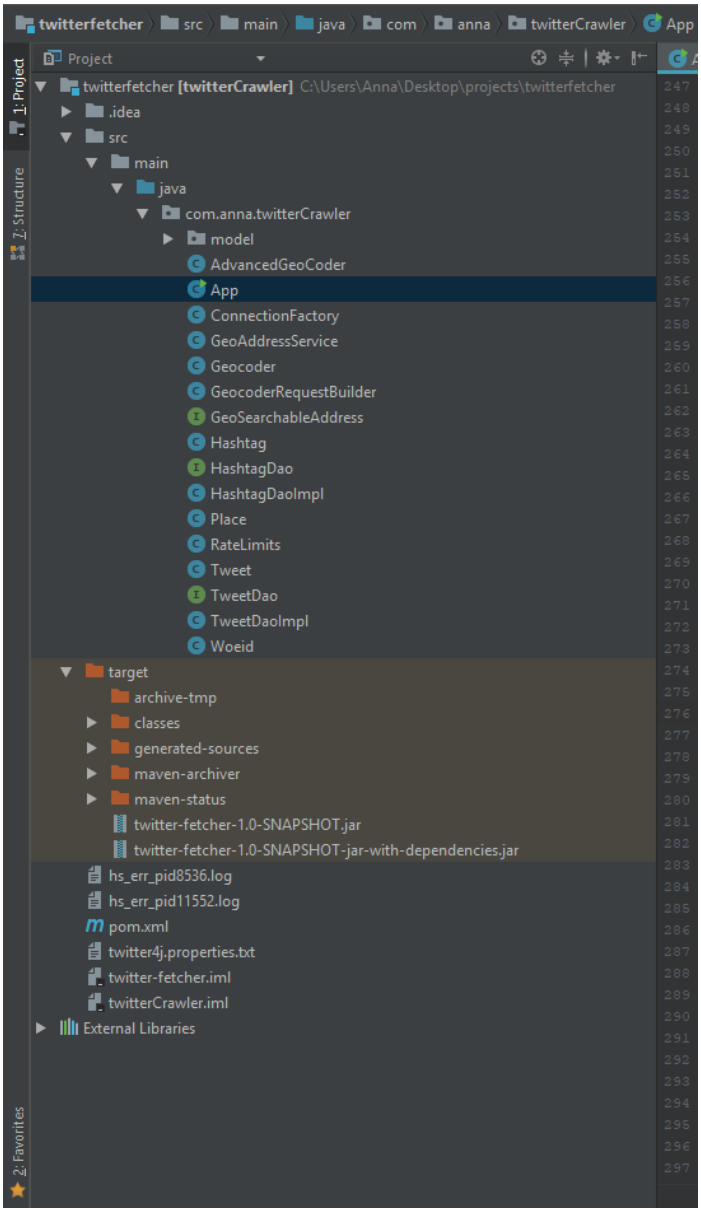
1. Λογαριασμούς και Χρήστες
2. Tweets και Απαντήσεις
3. Προσωπικά Μηνύματα
4. Διαφημίσεις
5. Εργαλεία και SDK's

Για την παρούσα εργασία χρησιμοποιήσαμε μόνο την πρόσβαση σε tweets και απαντήσεις. Η πρόσβαση σε αυτό το κομμάτι του API είναι συνεχής και βοήθησε στη συλλογή tweet για την ανάλυσή τους. Το data stream μπορεί να φιλτραριστεί με βάση hashtag, ονόματα χρηστών, περιοχές, κ.ά.

Πολύ σημαντικό είναι το γεγονός ότι μόνο το 1% των συνολικών δεδομένων μπορούν να ζητηθούν μέσω του API και επιπλέον το Twitter έχει Rate Limits σε πολλούς διαφορετικούς παράγοντες της χρήσης του API.

[4] https://developer.twitter.com/en/docs/ads/general/guides/getting-started.html

3.1.3 Περιγραφη της Εφαρμογης



Εικόνα 4η: Η δομή του κώδικα για την εφαρμογή που συλλέγει Tweet.

Για την ανάλυση συναισθήματος έπρεπε να συλλεχθούν πολλά tweet και γι' αυτό δημιουργήθηκε μια εφαρμογή σε Java.

Σκοπός της εφαρμογής είναι να συλλέγει αδιάκοπα tending hashtag και tweets αυτών των hashtag με βάση το WOEID[5] μια περιοχής, δηλαδή τον μοναδικό 32-bit κωδικό της.

Μέσω της εφαρμογής μπορεί να επιλεχθεί το WOEID ανάλογα με τη χώρα/περιοχή από την οποία θέλουμε να συλλέξουμε trending hashtags και στη συνέχεια αποθηκεύονται σε μία σχεσιακή βάση δεδομένων MySQL.

Για να γίνεται συνεχής συλλογή και αποθήκευση έπρεπε να γίνει αποφυγή των rate limit του Twitter API ώστε να τρέχει το πρόγραμμα συνεχόμενα.

Στην παρούσα εργασία επιλέχθηκαν trending hashtag από την Αγγλία. Οπότε μέσω της εφαρμογής γίνεται query που επιστρέφει τα trending hashtag της συγκεκριμένης χώρας. Στη συνέχεια με βάση το hashtag γίνεται διαφορετικό query που επιστρέφει τα tweet που έγιναν με αυτό. Τα retweet δεν χρειάζονται, οπότε δεν τα συμπεριλαμβάνουμε.

Πέρα από το κείμενο του Tweet, ζητάμε την περιοχή (αν είναι διαθέσιμη) και την ώρα δημοσίευσης.

3.1.4 Εργαλεια Αναπτυξης και Βιβλιοθηκες

Ο κώδικας Java είναι βασισμένος σε JDK 1.8 [6] και επιπλέον, για την εισαγωγή dependencies χρησιμοποιήθηκε το Apache Maven 3.5.4[7]

Τα dependencies που χρησιμοποιήθηκαν είναι τα εξής:

1. Το Twitter4j, μια βιβλιοθήκη κλήσεων στο API του Twitter.
2. Η βιβλιοθήκη geocoder-java [8], που κάνει κλήσεις στο Geocoder API της Google. Χρησιμοποιήθηκε για τον εντοπισμό των γεωγραφικών συντεταγμένων των Tweet που είχαν μόνο την πληροφορίας της χώρας και της πόλης.

[5] https://en.wikipedia.org/wiki/WOEID
[6] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[7] https://maven.apache.org/docs/3.5.4/release-notes.html
[8] https://github.com/seatgeek/geocoder-java

Σημαντικό είναι να αναφερθεί ότι για να τρέχει το πρόγραμμα χωρίς να χρειάζεται επανεκκίνηση έπρεπε να τηρηθούν τα Rate Limits που έχει το Twitter API.

Τα rate limits που έπρεπε να αποφευχθούν είναι τα ακόλουθα:

1. Tweet Limit
2. Place Limit (geocoder)
3. Hashtag Limit
4. Self Limit

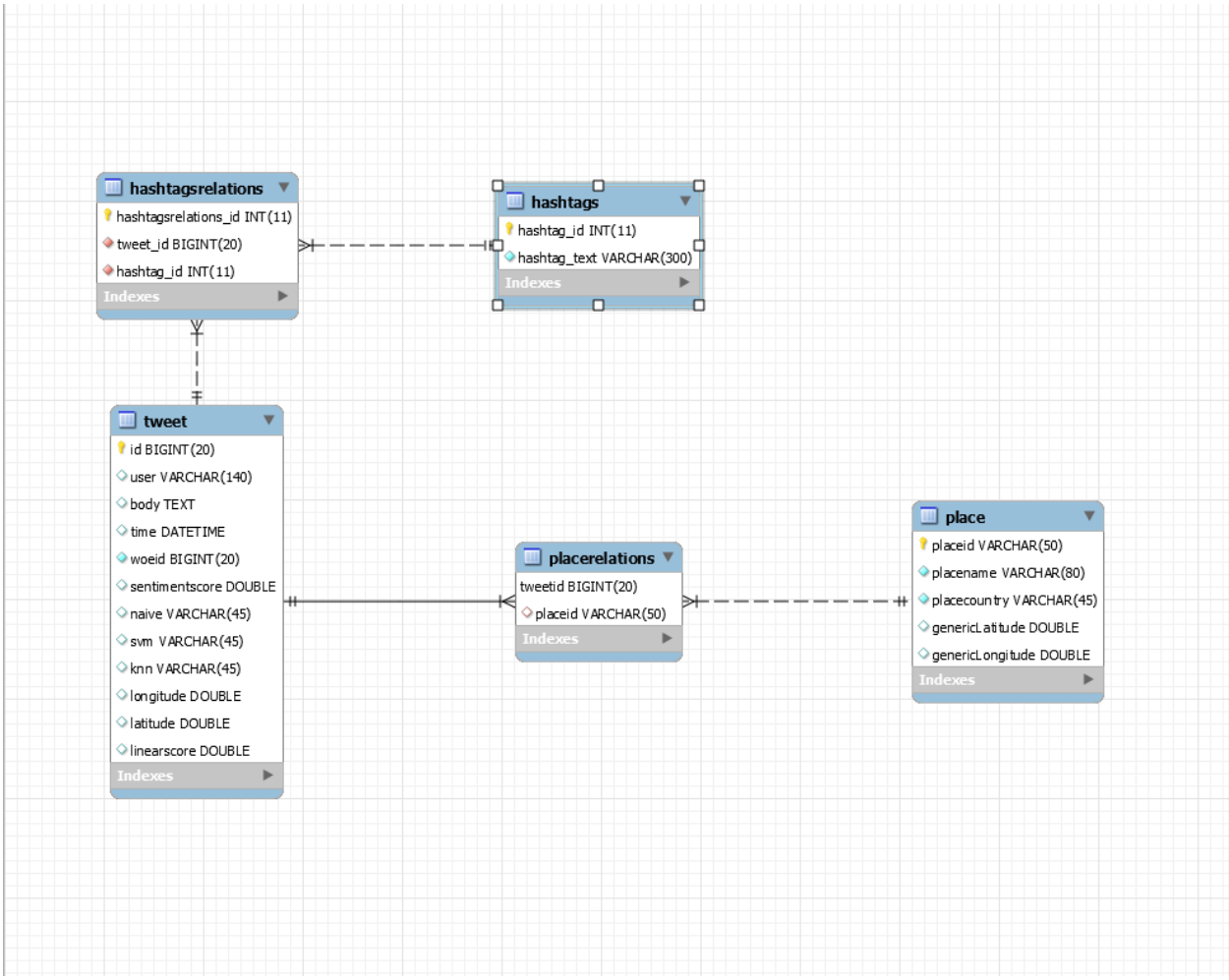
3.1.5 Η Βαση και ο Τροπος Αποθηκευσης

Η αποθήκευση των δεδομένων έγινε σε σχεσιακή βάση MySQL. Για την αποθήκευση των emoticon σε αυτό το είδος βάσης έπρεπε να γίνουν οι εξής ρυθμίσεις στο configuration αρχείο της MySQL.

```
[client]
default-character-set = utf8mb4
[mysql]
default-character-set = utf8mb4
[mysqld]
character-set-client-handshake = FALSE
character-set-server = utf8mb4
collation-server = utf8mb4_unicode_ci
init_connect='SET NAMES utf8mb4'
```

Ο λόγος που χρειάζονται αυτές οι ρυθμίσεις είναι ότι η default κωδικοποίηση UTF-8 της MySQL υποστηρίζει μόνο 3 bytes ανά χαρακτήρα, ενώ στην πραγματικότητα χρειάζεται μέχρι 4 bytes ανά χαρακτήρα.

Πρόκειται για ένα είδος bug της MySQL για το οποίο οι developers υλοποίησαν μια λύση, ένα καινούργιο character set που λέγεται utf8mb4. Έτσι, γίνεται δυνατή η αποθήκευση emoji με τη μορφή byte. Για παράδειγμα το emoticon 😃 αποθηκεύεται με αυτή τη μορφή: \xF0\x9F\x98\x83 (Unicode: U+1F604).



Εικόνα 5η: Το σχήμα της βάσης δεδομένων.

3.1.6 Δεδομενα με Geolocation

Μία από τις δυσκολίες που εμφανίστηκαν ήταν ότι το Twitter API επιστρέφει περιορισμένα Tweets με Geolocation. Πολλά από αυτά περιέχουν μόνο την πληροφορία της περιοχής και όχι συντεταγμένες γεωγραφικού μήκους και πλάτους.

Έτσι, κρίθηκε αναγκαίο να αναθέσουμε μεταγενέστερα τις συντεταγμένες στα tweet που είχαν μόνο την περιοχή, για παράδειγμα London, UK. Για να επιτευχθεί αυτό, χρησιμοποιήθηκε ένας Geocoder, ο οποίος βρίσκει τις συντεταγμένες για τα tweet που δεν είχαν αυτή την πληροφορία.

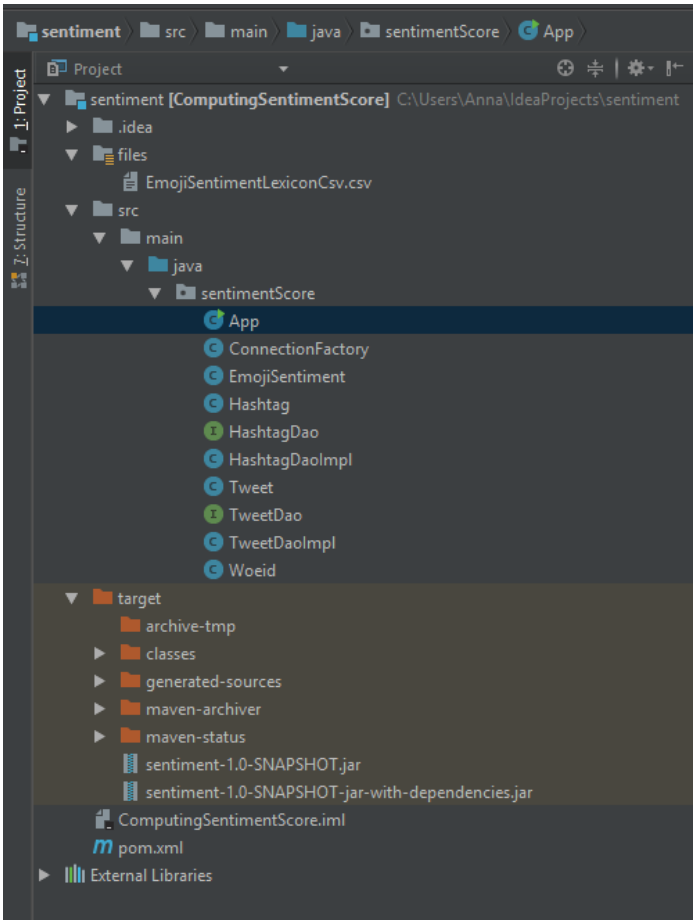
3.1.7 Το Τελικο Dataset

Τα συνολικά δεδομένα που συλλέχθηκαν αποτελούνταν από περίπου 2.500.000 tweets από τις 2 Μαρτίου έως τις 12 Αυγούστου του 2018. Όλα ανεξαιρέτως αναφέρονται σε trending θέματα που απασχόλησαν την κοινή γνώμη της Αγγλίας.

Ένα από τα προβλήματα που προέκυψαν στην πορεία της εργασίας και αφού έγιναν πολλά πειράματα, ήταν το ότι πολλά hashtag είχαν πολύ λίγα Tweet με emoticon, πράγμα που περιόριζε πολύ το training set που θα μπορούσε να δημιουργηθεί.

Λόγω του παραπάνω, πάρθηκε η απόφαση να διαγραφούν τα Hashtag, που είχαν λιγότερο από 500 Tweet με emoticon. Το dataset που έμεινε μετά από αυτή τη διεργασία αποτελείται από 89 Hashtag και 833.144 Tweet.

3.2 ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΑΠΟΔΟΣΗ SENTIMENT SCORE ΜΕ ΒΑΣΗ ΤΟ EMOTICON



3.2.1 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ

Εικόνα 6η: Η δομή της εφαρμογής που υπολογίζει το sentiment score των Tweet με βάση το emoticon.

Ο κώδικας Java είναι βασισμένος σε JDK 1.8 [9] και για την εισαγωγή dependencies χρησιμοποιήθηκε το Apache Maven 3.5.4 [10] . Για την υλοποίηση χρησιμοποιήθηκε ως

[9] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[10] https://maven.apache.org/docs/3.5.4/release-notes.html

dependency η βιβλιοθήκη emoji4j[11], για τη μετατροπή συντομεύσεων και HTML σε emoji και αντίστροφα.

Για την απόδοση του συναισθήματος ο κώδικας ακολουθεί τις εξής διαδικασίες:

1. Διαβάζει τα Tweet και βρίσκει τα emoticon που περιέχουν.
2. Για κάθε emoticon βάσει του Unicode κωδικού του βρίσκει το sentiment score του σύμφωνα με το emoji sentiment lexicon.
3. Προσθέτει το sentiment score όλων των emoticon, ακόμα κι αν υπάρχουν πάνω από μία φορά.
4. Στο τέλος αποθηκεύει το άθροισμα του sentiment score στη βάση δεδομένων.

[11] https://github.com/kcthota/emoji4j

3.3 ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

3.3.1 INFORMATION RETRIEVAL ΚΑΙ ΔΕΔΟΜΕΝΑ ΚΕΙΜΕΝΟΥ

Ο τομέας της ανάκτησης πληροφορίας εξελίχθηκε παράλληλα με τα συστήματα βάσεων δεδομένων. Το βασικό πρόβλημα σε ό, τι αφορά την ανάκτηση πληροφορίας σε δεδομένα κειμένου είναι ότι τα ίδια τα δεδομένα είναι τεράστια, και ημιδομημένα. Η έρευνα σε αυτόν τον τομέα έχει εξελιχθεί πολύ τις τελευταίες δεκαετίες και οι κύριες αρχές που έχει ιδρύσει είναι οι εξής:

1. Ένα κείμενο μπορεί να περιγραφεί από ένα σύνολο αντιπροσωπευτικών λέξεων.
2. Κάθε λέξη έχει διαφορετική βαρύτητα όταν χρησιμοποιείται για να περιγράψει το περιεχόμενο του κειμένου.
3. Η βαρύτητα μπορεί να αποδοθεί με τη χρήση αριθμητικών βαρών σε κάθε λέξη (συχνότητα, TF-IDF).

3.3.2 ΑΝΑΠΑΡΑΣΤΑΣΗ ΔΕΔΟΜΕΝΩΝ ΚΕΙΜΕΝΟΥ

BOOLEAN

Το μοντέλο Bag-Of-Words (BOW) είναι το πιο παλιό στον τομέα της ανάκτησης πληροφοριών από κείμενα και έχει χρησιμοποιηθεί ευρέως στον τομέα της επεξεργασίας της φυσικής γλώσσας. Με βάση αυτό το μοντέλο, ένα κείμενο αντιπροσωπεύεται ως το υπερσύνολο των λέξεών του. Οι τιμές που παίρνει κάθε λέξη μπορεί να είναι 0 ή 1 ανάλογα με το αν υπάρχει ή όχι στο κείμενο[12].

[12] https://en.wikipedia.org/wiki/Bag-of-words_model

VECTOR SPACE MODEL

Το μοντέλο διανυσματικού χώρου, προτάθηκε για πρώτη φορά το 1975 από τον Salton ως καλύτερη εναλλακτική του Boolean μοντέλου. Η έννοιά του προέρχεται από τη γραμμική άλγεβρα και ουσιαστικά με βάση αυτό το μοντέλο τα κείμενα μετασχηματίζονται σε διανύσματα. Με αυτόν τον τρόπο οι συντεταγμένες του μπορούν να πάρουν πραγματικούς αριθμούς και να μετρηθεί η συσχέτισή τους, δηλαδή έτσι επιτρέπεται η σύγκριση πολλών κειμένων μεταξύ τους. [17]

Πιο συγκεκριμένα τα κείμενα αντιπροσωπεύονται ως διανύσματα m διαστάσεων, όπου m είναι ο συνολικός αριθμός των λέξεων στη συλλογή των κειμένων. Το ποσοστό συσχέτισης του κειμένου d ως προς το επερώτημα q υπολογίζεται ως τη συσχέτιση των διανυσμάτων που τα αντιπροσωπεύουν, χρησιμοποιώντας μετρικές όπως η Ευκλείδεια απόσταση.

TF-IDF

Υπάρχουν πολλοί τρόποι ανάθεσης βαρών αλλά ο πιο συνηθισμένος, ονομάζεται πρότυπο συχνότητας όρου-αντίστροφης συχνότητας εγγράφου ή TF-IDF. Το μοντέλο TF-IDF υποστηρίζει ότι οι λέξεις που εμφανίζονται συχνά σε ένα έγγραφο (TF) σε σχέση με το πόσες φορές εμφανίζονται στο σύνολο των κειμένων (IDF) είναι πιο σημαντικές από αυτές που υπάρχουν σε μεγάλο αριθμό κειμένων. [17] Είναι σημαντικό να αναφερθεί ότι αποτελεί ένα από τα πιο δημοφιλή μοντέλα, ωστόσο οι Pang κ.ά. για παράδειγμα είχαν μεγαλύτερη ακρίβεια στην κατηγοριοποίηση χρησιμοποιώντας το μοντέλο BOW [24].

$$\mathbf{tf}(t,d) = \frac{f_d(t)}{\max\limits_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

$$\mathbf{tfidf}'(t,d,D) = \frac{\mathbf{idf}(t,D)}{|D|} + \mathbf{tfidf}(t,d,D)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

Εικόνα 7η[13] : Ο υπολογισμός του TF-IDF

N-GRAMS

Τα n-grams είναι μία τεχνική μοντελοποίησης ακολουθιών λέξεων. Ένα n-gram περιλαμβάνει n συνεχόμενες λέξεις. Για παράδειγμα, στη φράση «Φέτος θα πάω διακοπές στο Βόλο», τα bigrams (2-grams) είναι τα εξής: «Φέτος θα», «θα πάω», «πάω διακοπές», «διακοπές στο», «στο Βόλο» [17]. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη όταν ένα κείμενο έχει θόρυβο, όταν το σύνολο των κειμένων είναι μικρό ή όταν το κείμενο έχει μεταφορικό λόγο.

TOKENIZATION

Το βασικό βήμα για την αναπαράσταση ενός κειμένου είναι ο διαχωρισμός του σε διακριτές λέξεις, που μπορούν να ονομαστούν tokens. Σε ό,τι αφορά τα κείμενα ο διαχωρισμός γίνεται ως επί το πλείστων στα κενά και στα σημεία στίξης [17]. Το ποια από αυτά τα tokens θα

[13] https://www.joyofdata.de/blog/tf-idf-statistic-keyword-extraction/

χρησιμοποιηθούν αργότερα για την αναπαράσταση του κειμένου είναι στην ευχέρεια του ερευνητή και υπάρχουν πολλές μέθοδοι που βοηθούν σε αυτό το κομμάτι.

PART OF SPEECH TAGGING

Η σύνταξη είναι ίσως το πιο σημαντικό στοιχείο ενός κειμένου, πέρα από το λεξιλόγιο και τη φρασεολογία, και μπορεί να δώσει πολύ ενδιαφέρουσες πληροφορίες για το περιεχόμενό του. Γι' αυτούς τους λόγους στον τομέα της ανάλυσης κειμένων αποτελεί κοινή πρακτική η αναγνώριση μερών του λόγου για κάθε λέξη. Αυτή η μέθοδος μπορεί να βοηθήσει στον εντοπισμό των πιο σημαντικών λέξεων, αλλά και σε άλλες μεθόδους όπως το stemming και η λημματοποίηση. [17]

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Εικόνα 8η [14] : Part of Speech Tags για την αγγλική γλώσσα (Penn Treebank project).

[14] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

STEMMING

Το Stemming είναι η διαδικασία κατά την οποία εντοπίζεται η ρίζα της λέξης ή η λέξη μετασχηματίζεται σε μια πιο απλή μορφή. [17] Αυτή η μέθοδος είναι πολύ χρήσιμη στον τομέα της ανάλυσης συναισθήματος ώστε να μειώνονται οι διαστάσεις του training set, δηλαδή να μειώνονται οι λέξεις. Ο λόγος είναι ότι οι διαφορετικές γραμματικές μορφές μιας λέξης πολλές φορές δεν αλλάζουν το νόημά της. Για παράδειγμα η λέξη τρέχω και η λέξη τρέχοντας έχουν ουσιαστικά την ίδια βαρύτητα, αλλά αν δεν πέρναγαν από διαδικασία stemming θα υπήρχαν δύο φορές μέσα στο training set.

LEMMATIZATION

Η λημματοποίηση είναι πιο αποτελεσματική μέθοδος από το Stemming για την εύρεση της ρίζας των λέξεων. Χρησιμοποιεί τα συμφραζόμενα και τα μέρη του λόγου για να βρει το λήμμα και εφαρμόζει διαφορετικές τεχνικές αναλόγως. [18]

STOP WORD REMOVAL

Η συγκεκριμένη τεχνική είναι από τις πιο συχνές σε ό,τι αφορά την επεξεργασία ενός κειμένου και την αναπαράστασή του. Υποστηρίζεται ότι ορισμένες λέξεις, όπως τα άρθρα, οι σύνδεσμοι, τα νούμερα, κ.ά, δεν έχουν πολλά να προσφέρουν στην ανάλυση ενός κειμένου. Γι' αυτό είναι κοινή πρακτική να αφαιρούνται από το σύνολο των χαρακτηριστικών.

3.3.3 Αλγοριθμοι Κατηγοριοποιησης

ΝΑΪVE BAYES

Οι ταξινομητές Naïve Bayes αποτελούν ένα σύνολο αλγορίθμων που βασίζονται στο θεώρημα του Bayes για την κατηγοριοποίηση κάθε κλάσης. Βασίζονται στην παραδοχή ότι οι πληροφορίες για τις κλάσεις είναι γνωστές υπό μορφή προγενέστερων πιθανοτήτων και κατανομών. Δηλαδή για να κατηγοριοποιήσουν νέα δεδομένα εισόδου χρησιμοποιούν εκ των υστέρων πιθανότητες. Αυτό το επιτυγχάνουν μετατρέποντας την εκ των προτέρων πιθανότητα στην εκ των υστέρων με βάση τα δεδομένα προς κατηγοριοποίηση, χρησιμοποιώντας τις τιμές πιθανότητας [19].

«Έστω ότι X είναι τα δεδομένα που δεν ξέρουμε την κλάση τους και H_i μία υπόθεση που δείχνει σε ποια κλάση ανήκει το X. Για παράδειγμα, H_i είναι η υπόθεση ότι τα δεδομένα ανήκουν στην κλάση C_i. Ας υποθέσουμε ότι η εκ των προτέρων πιθανότητα του H_i, που δίνεται από το $P(H_i)$, είναι γνωστή. Για να κατηγοριοποιηθεί το X, πρέπει να υπολογίσουμε το $P(H_i \mid X)$, που είναι η πιθανότητα να ισχύει η υπόθεση H_i, δεδομένου του X. Το $P(H_i \mid X)$ είναι η εκ των υστέρων πιθανότητα να ισχύει το H_i δεδομένου του X. Σε αντίθεση, το $P(H_i)$ είναι η εκ των προτέρων πιθανότητα του H_i. Είναι η πιθανότητα της υπόθεσης ανεξάρτητα από την τιμή του X. Η εκ των υστέρων πιθανότητα, $P(H_i \mid X)$ βασίζεται στο X και άλλες πληροφορίες, ενώ η εκ των προτέρων πιθανότητα, $P(H_i)$ ορίζεται πριν παρατηρηθεί το X. Παρομοίως, το $P(X \mid H_i)$ είναι η πιθανότητα να ισχύει το X δεδομένου του H_i. Το θεώρημα του Bayes βοηθάει στην πρόβλεψη της εκ των υστέρων πιθανότητας.» [19]

$$P(H_i \mid X) = \frac{P(X \mid H_i)P(H_i)}{P(X)}$$

Εικόνα 14η [15]: Υπολογισμός του $P(H_i \mid X)$.

[15] https://link.springer.com/chapter/10.1007/978-0-85729-495-1_4

Για την κατηγοριοποίηση των Tweet σε αυτή την εργασία, χρησιμοποιήθηκε ο Multinomial Naïve Bayes. Με βάση αυτόν τον ταξινομητή η πιθανότητα ένα κείμενο d να ανήκει στην κατηγορία c υπολογίζεται ως εξής:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

Εικόνα 15η [16]: Υπολογισμός της πιθανότητας της κλάσης.

SUPPORT VECTOR MACHINES

Εκτός από τους ταξινομητές που χρησιμοποιούν πιθανολογικά μοντέλα, υπάρχουν και αυτοί που προσεγγίζουν τα δεδομένα γραμμικά. Ένας από αυτούς είναι ο Support Vector Machine, που εισήχθη από τους Vapnik και Chermovenkis το 1963. Αργότερα οι Boser, Guyon και Vapnik, οι οποίοι πρότειναν τη μέθοδο του πυρήνα.

Πρόκειται για έναν αλγόριθμο μηχανικής μάθησης που βασίζεται σε διανύσματα και επιχειρεί να βρει τη διαχωριστική γραμμή μεταξύ κλάσεων, με τέτοιο τρόπο ώστε να κατηγοριοποιούνται καλύτερα τα νέα δεδομένα. Ο SVM είναι γνωστός για την καλή απόδοση γενίκευσης που οφείλεται στην ιδιότητα του μέγιστου περιθωρίου. Γι' αυτό πολλές φορές τον αποκαλούν και ταξινομητή μέγιστου περιθωρίου. [20]

Πιο συγκεκριμένα, «είναι βασισμένος στην αρχή Structural Risk Minimization από τη θεωρία του computational learning. Η ιδέα της SRM είναι να βρει μια υπόθεση h για την οποία μπορεί να εγγυηθεί το χαμηλότερο πραγματικό σφάλμα. Το πραγματικό σφάλμα της h είναι η πιθανότητα ότι η h θα κάνει λάθος σε ένα μη νέο και τυχαία επιλεγμένο παράδειγμα του test set. Ένα ανώτατο όριο μπορεί να χρησιμοποιηθεί για να συνδέσει το πραγματικό σφάλμα της υπόθεσης h με το σφάλμα της h στο training set και την πολυπλοκότητα του H, το χώρο υποθέσεων που περιλαμβάνουν την h. Ο Support Vector Machine βρίσκει την υπόθεση h που ελαχιστοποιεί αυτό το όριο του πραγματικού σφάλματος ελέγχοντας τις διαστάσεις του H» [21].

[16] https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

Για να χρησιμοποιήσουμε τον SVM για την ανάλυση συναισθήματος χρησιμοποιήσαμε το Radial Basis Function Kernel:

$$K\left(x^{(i)}, x^{(j)}\right) = \phi(x^{(i)})^T \phi(x^{(j)}) = \exp\left(-\gamma \left\|x^{(i)} - x^{(j)}\right\|^2\right), \qquad \gamma > 0$$

Εικόνα 16η [17]: Radial Basis Function Kernel

K-NEAREST NEIGHBORS

Ο K-nn είναι ένας από τους πιο γνωστούς ταξινομητές στον τομέα του Machine Learning. Ουσιαστικά αυτό που κάνει είναι να διαχωρίζει τα δεδομένα σε k κλάσεις, με σημείο αναφοράς και σύγκρισης τους k κοντινότερους γείτονες.

«Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα training set με κατηγοριοποιημένα δεδομένα. Κάθε νέο δεδομένο θα πρέπει να κατηγοριοποιείται σύμφωνα με τα κοντινά του δεδομένα. Αν επομένως η κατηγοριοποίηση ενός παραδείγματος δεν είναι γνωστή, τότε μπορεί να γίνει ελέγχοντας την κλάση των κοντινών του γειτόνων. Ο K-nn υπολογίζει την απόσταση μεταξύ του testing set και του training set. Τελικά, η απόσταση με την μικρότερη τιμή, δηλαδή η κοντινότερη, χρησιμοποιείται για την κατηγοριοποίηση νέων δεδομένων.» [22]

Η μετρική της απόστασης παίζει πολύ σημαντικό ρόλο στην απόδοση του αλγορίθμου. Εξίσου σημαντικός παράγοντας είναι η τιμή του k, που αποτελεί τη βασική παράμετρο του αλγορίθμου. Αν το k είναι πολύ μεγάλο, οι κλάσεις με πολλά κατηγοριοποιημένα δεδομένα μπορεί να υπερισχύσουν και τα αποτελέσματα να είναι μεροληπτικά. Το αντίθετο συμβαίνει αν το k είναι πολύ μεγάλο. [22]

[17] http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex8/ex8.html

Δύο αποστάσεις που χρησιμοποιήσαμε για την παρούσα εργασία είναι η Ευκλείδεια και η Chebyshev.

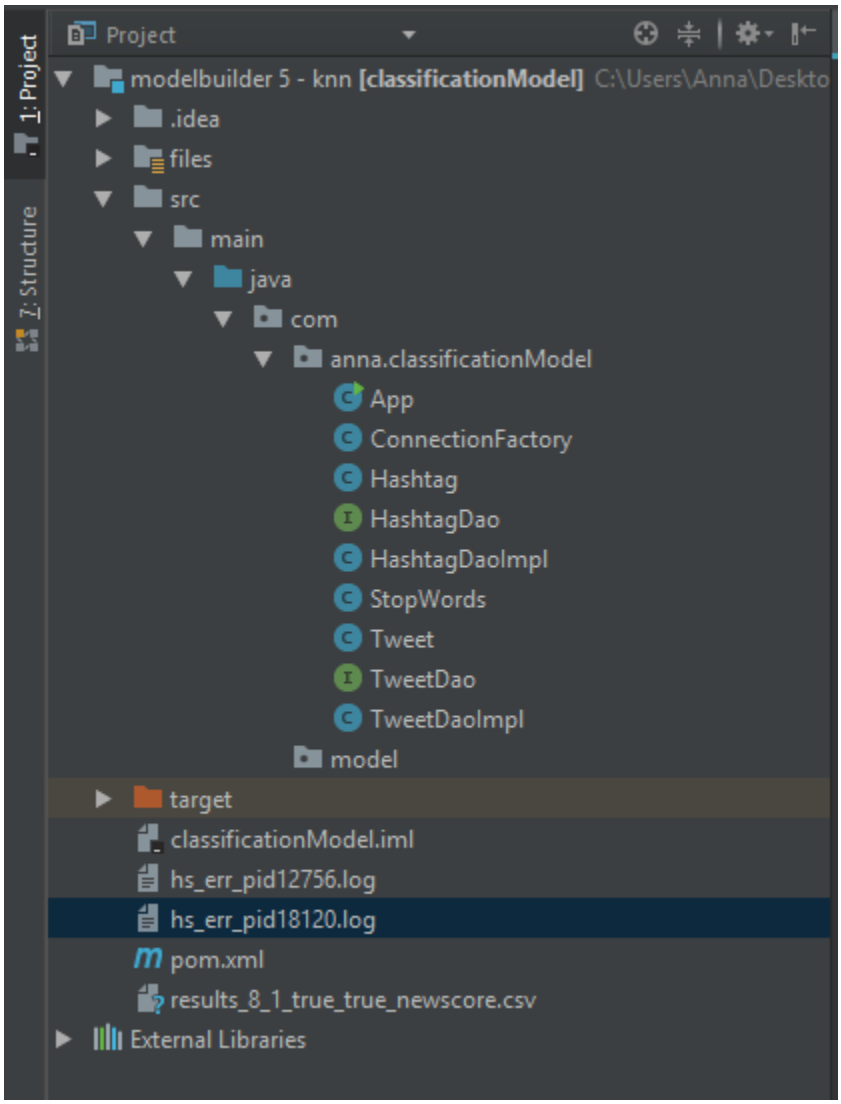
ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σύμφωνα με τον Χαράλαμπο Γναρδέλλη, «Η απλή γραμμική παλινδρόμηση ποσοτικοποιεί τη σχέση δύο συνεχών τυχαίων μεταβλητών Χ και Υ, υπό τη μορφή ενός γραμμικού υποδείγματος, στο οποίο οι τιμές της μίας μεταβλητής εκτιμώνται (ή προβλέπονται) από τις τιμές της άλλης. Αν οι τιμές της μεταβλητής Υ εκτιμώνται από τις τιμές της Χ, τότε η Υ ονομάζεται εξαρτημένη μεταβλητή και η μεταβλητή Χ ονομάζεται ανεξάρτητη». [23]

Η γραμμική παλινδρόμηση ή αλλιώς Linear Regression μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών της κλάσης ενός dataset με βάση ένα «κατηγοριοποιημένο» με τιμές training set. Μάλιστα, έχει χρησιμοποιηθεί για την πρόβλεψη κριτικών ταινιών (δηλαδή αν κρίθηκαν με 1-5 αστέρια). [26]

Εφόσον το sentiment score που λαμβάνουμε από το Emoji Sentiment Lexicon είναι σε μορφή αριθμού, θεωρήσαμε ότι για τους σκοπούς τις συγκεκριμένης εργασίας μπορούμε να χρησιμοποιήσουμε Linear Regression για να προβλέψουμε το score των νέων Tweet.

3.3.4 Περιγραφη της Εφαρμογης



Εικόνα 9η: Η δομή της εφαρμογής για την προεπεξεργασία και κατηγοριοποίηση των δεδομένων.

Για την προεπεξεργασία των Tweet, τη δημιουργία training και test set, καθώς και για την κατηγοριοποίηση, δημιουργήθηκε μια εφαρμογή σε Java.

Τα βήματα για την προεπεξεργασία των Tweet είναι τα εξής:

1. Συλλέγουμε τα Tweet που έχουν sentiment score από τη βάση δεδομένων.
2. Μετατρέπουμε τα κεφαλαία γράμματα σε μικρά.

3. Διαγράφουμε τους συνδέσμους.
4. Αφαιρούμε Hashtags και Users.
5. Αφαιρούμε τον χαρακτήρα «’».
6. Αντικαθιστούμε τον χαρακτήρα «_» με κενό.
7. Αφαιρούμε τα emoticon.
8. Από το κείμενο που έχει μείνει χωρίζουμε τις λέξεις σε όποιο χαρακτήρα δεν αποτελεί μέρος της αλφαβήτου.
9. Αφαιρούμε τις λέξεις ενός χαρακτήρα .
10. Αφαιρούμε τους αριθμούς ή τις λέξεις που περιέχουν αριθμό.
11. Αν κάποιος χαρακτήρας εμφανίζεται σε μία λέξη 3 ή παραπάνω συνεχόμενες φορές, τον αφήνουμε να υπάρχει μία φορά. (Για παράδειγμα το «Yeeees» θα γίνει «Yes»).
12. Αναθέτουμε Part of Speech tags στις λέξεις[18].
13. Κάνουμε λημματοποίηση χρησιμοποιώντας λεξικό[19].
14. Αν η λέξη δεν μπορεί να λημματοποιηθεί γιατί δεν υπάρχει στο λεξικό και βρεθεί πάνω από μία φορά στο σύνολο των Tweet του training set την κρατάμε ως attribute.
15. Αφαιρούμε τα Stop Words.

Στη συνέχεια, αφού έχουμε ετοιμάσει το training set, επιλέγουμε τον αντίστοιχο classifier (Multinomial Naïve Bayes, SVM, K-nn) ή Linear Regression. Το training γίνεται με το 90% των Tweet και το testing με το 10% μέσω 10 fold Cross Validation. Ο κώδικας μετά το Cross Validation κρατάει το training set με τη μεγαλύτερη απόδοση.

Δίνοντας νέα tweet το μοντέλο μπορεί να κάνει κατηγοριοποίηση και τα αποτελέσματα εμφανίζονται στην κονσόλα, ενώ παράλληλα αποθηκεύονται στη βάση δεδομένων.

Επειδή χρησιμοποιήθηκαν 3 διαφορετικοί αλγόριθμοι κατηγοριοποίησης, καθώς και γραμμική παλινδρόμηση, υλοποιήθηκαν 4 παραλλαγές της ίδιας εφαρμογής.

Σε ό,τι αφορά τους αλγορίθμους κατηγοριοποίησης οι εφαρμογές διατηρούν τον ίδιο τρόπο προεπεξεργασίας του κειμένου και το μόνο που αλλάζει είναι ο αλγόριθμος.

[18] http://opennlp.sourceforge.net/models-1.5/
[19] http://opennlp.sourceforge.net/models-1.5/

Για τη γραμμική παλινδρόμηση, χρειάστηκε να αλλαχθεί ο τρόπος δημιουργίας του training set.

3.3.5 Εργαλεια Αναπτυξης και Βιβλιοθηκες

Ο κώδικας Java είναι βασισμένος σε JDK 1.8 [20] και για την εισαγωγή dependencies χρησιμοποιήθηκε το Apache Maven 3.5.4[21]

Τα dependencies που χρησιμοποιήθηκαν είναι τα εξής:

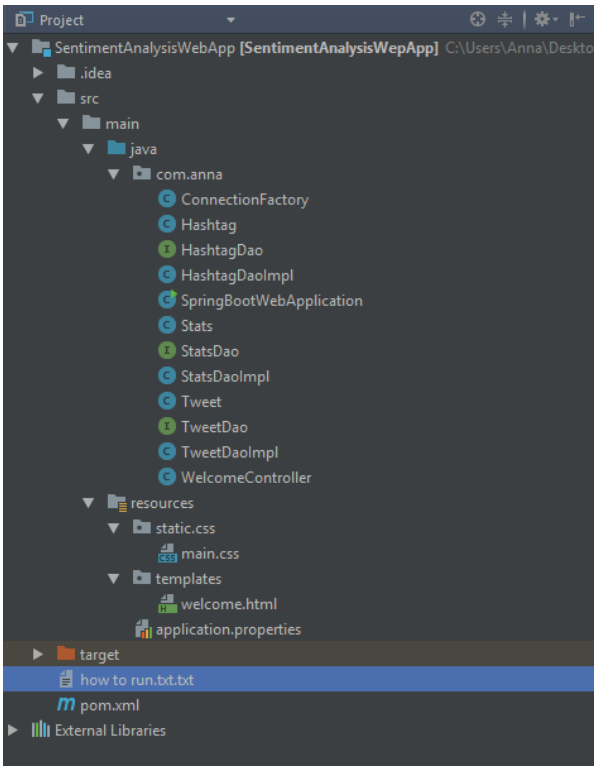
1. Η βιβλιοθήκη emoji4j, για την αναγνώριση και αφαίρεση των emoticon.[22]
2. Η βιβλιοθήκη LibSVM[23], για τη χρήση του Support Vector Machine. [25]
3. Το OpenNLP tools[24].
4. Η βιβλιοθήκη WEKA[25] για την JAVA

[20] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[21] https://maven.apache.org/docs/3.5.4/release-notes.html
[22] https://github.com/kcthota/emoji4j
[23] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[24] http://opennlp.sourceforge.net/models-1.5/
[25] https://www.cs.waikato.ac.nz/ml/weka/

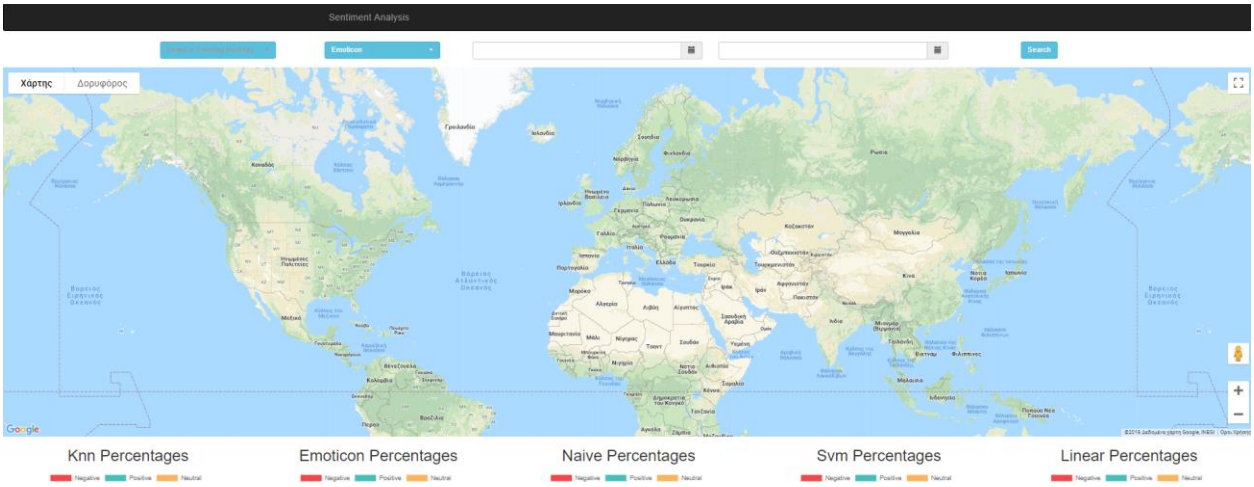
3.4 ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟ WEB

3.4.1 ΠΕΡΙΓΡΑΦΗ ΚΑΙ ΣΚΟΠΟΣ ΤΗΣ ΕΦΑΡΜΟΓΗΣ

Η οπτικοποίηση των αποτελεσμάτων βοηθάει πολύ στην αξιολόγησή και τη σύγκρισή τους, γι' αυτό δημιουργήσαμε μια δικτυακή εφαρμογή που δείχνει τα Tweet στο χάρτη και παρουσιάζει τα στατιστικά στοιχεία των αποτελεσμάτων από διαφορετικούς classifiers. Επίσης, επιλέγοντας ένα tweet εμφανίζεται το κείμενό του, η ημερομηνία δημοσίευσης και η περιοχή. Ο χρήστης έχει τη δυνατότητα να φιλτράρει τα αποτελέσματα με βάση το Hashtag, την ημερομηνία και τον classifier.



Εικόνα 10η: Η δομή της εφαρμογής οπτικοποίησης των αποτελεσμάτων.



Εικόνα 11η: Η εφαρμογή παρουσίασης των αποτελεσμάτων.

3.4.2 Εργαλεια Αναπτυξης και Βιβλιοθηκες

Εκτός από τον MySQL Server 5.7[26], που χρησιμοποιούμε για την αποθήκευση των tweet, για την εφαρμογή χρησιμοποιήθηκε το IntelliJ Community Edition[27] και το MySQL Workbench 8.0[28].

Ο κώδικας Java είναι βασισμένος σε JDK 1.8 [29] και επιπλέον, για την εισαγωγή dependencies χρησιμοποιήθηκε το Apache Maven 3.5.4[30]

Τα βασικά dependencies είναι τα εξής:

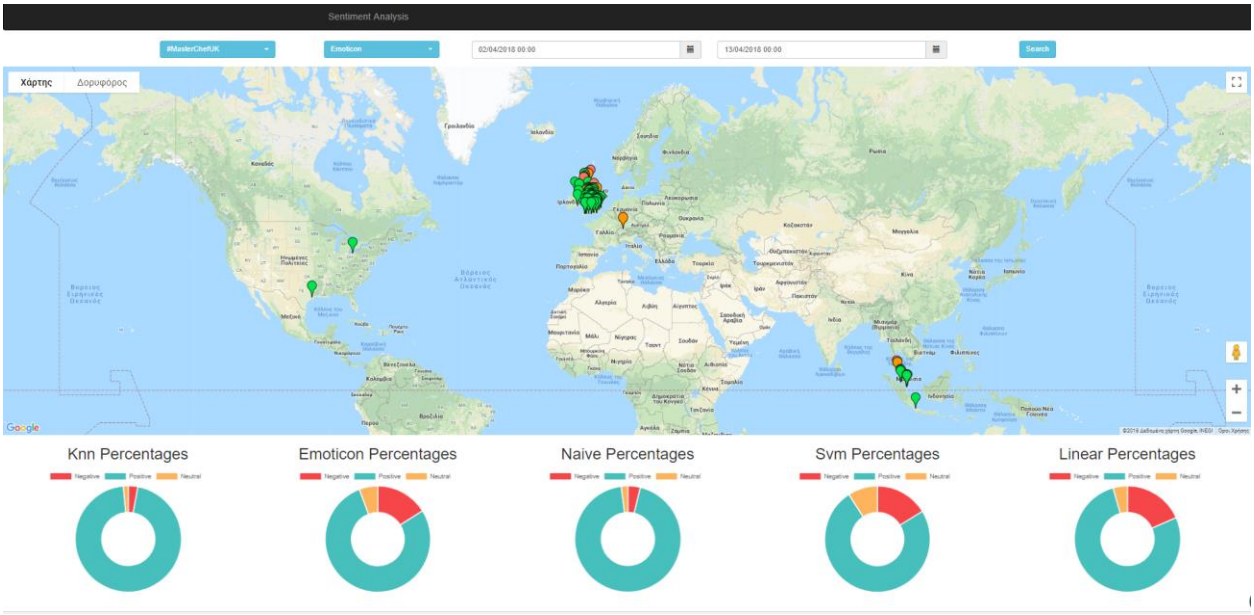
1. Για τη δημιουργία του HTTP Server χρησιμοποιήθηκε το Framework Spring Boot.[31]
2. Ως βιβλιοθήκη CSS χρησιμοποιήθηκε το Bootstrap CSS[32]

[26] https://dev.mysql.com/downloads/mysql/
[27] https://www.jetbrains.com/idea/download/
[28] https://www.mysql.com/products/workbench/
[29] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[30] https://maven.apache.org/docs/3.5.4/release-notes.html
[31] https://spring.io/projects/spring-boot
[32] https://getbootstrap.com/docs/3.3/css/

3. Ως HTML template engine χρησιμοποιήθηκε το Thymeleaf[33]
4. Για την παρουσίαση των Tweet σε χάρτη χρησιμοποιήθηκε το Google Maps.
5. Για τα γραφήματα ποσοστών χρησιμοποιήθηκε το javascript και css framework Material Design για το Bootstrap [34]
6. Για τη σύνδεση της εφαρμογής με τη βάση χρησιμοποιήθηκε το JDBC driver της MySQL.

3.4.3 Παραδειγμα Χρησης

Επιλέγοντας ένα Hashtag η σελίδα ανανεώνεται και παρουσιάζονται όλα τα Tweet με geolocation που είναι αποθηκευμένα στη βάση.



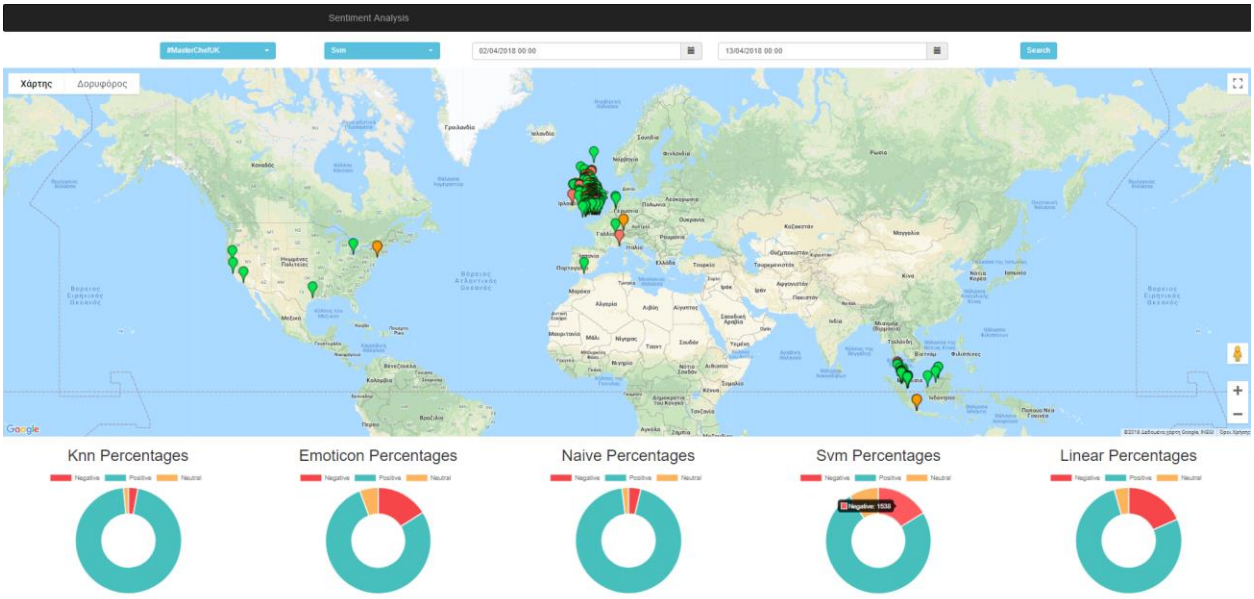
Εικόνα 12η: Τι εμφανίζεται στη σελίδα όταν επιλέγεις Hashtag.

Οι ημερομηνίες ανανεώνονται και αναγράφεται το διάστημα του χρόνου κατά τα οποίο δημοσιεύτηκαν τα συγκεκριμένα Tweet. Ο χρήστης έχει τη δυνατότητα να περιορίσει τα αποτελέσματα ανά ημέρα και ώρα.

[33] https://www.thymeleaf.org/
[34] https://mdbootstrap.com/

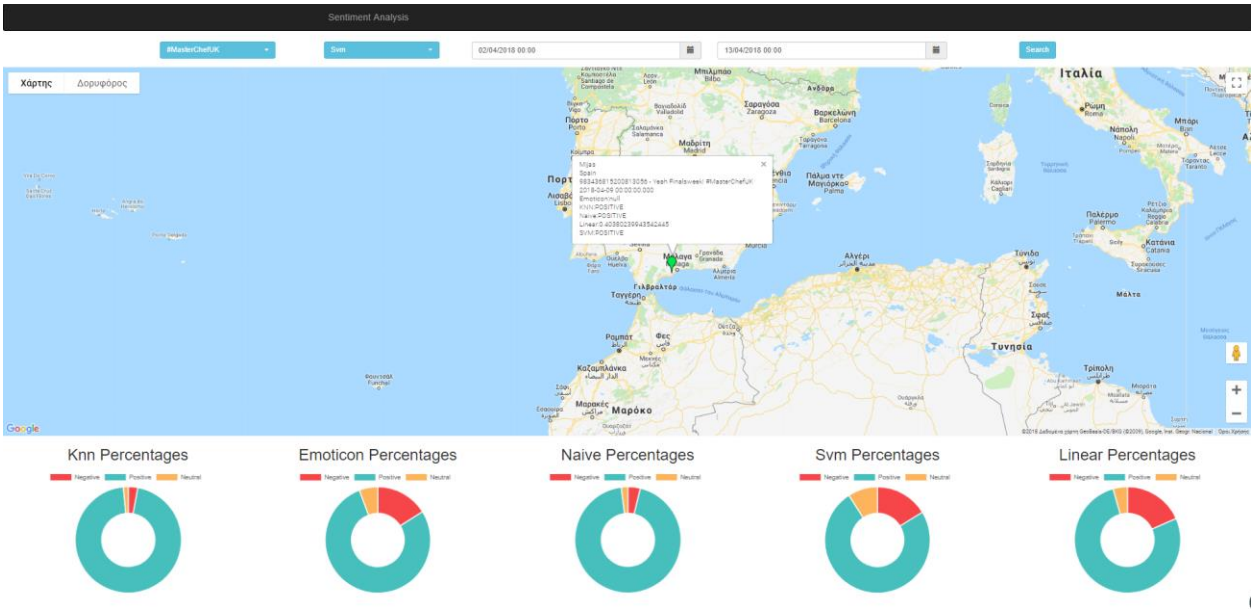
Επιλέγοντας ένα Hashtag η εφαρμογή παρουσιάζει τα tweet στο χάρτη με δείκτες κόκκινου, πράσινου ή πορτοκαλί χρώματος, αναλόγως με το συναίσθημα που έχει υπολογιστεί γι' αυτά με βάση τα emoticon που περιέχουν.

Ο χρήστης έχει τη δυνατότητα στη συνέχεια να επιλέξει classifier. Με την επιλογή του classifier η σελίδα ανανεώνεται και εμφανίζονται τα κατηγοριοποιημένα Tweet με βάση το συναίσθημα που τους έχει αποδώσει.



Εικόνα 13η: Ανανέωση της σελίδας όταν επιλέγεις διαφορετικό classifier.

Επιλέγοντας ένα tweet ο χρήστης μπορεί να δει όλες τις πληροφορίες που είναι αποθηκευμένες γι' αυτό στη βάση δεδομένων.



Εικόνα 14η: Τι εμφανίζεται όταν επιλέγεις ένα Tweet.

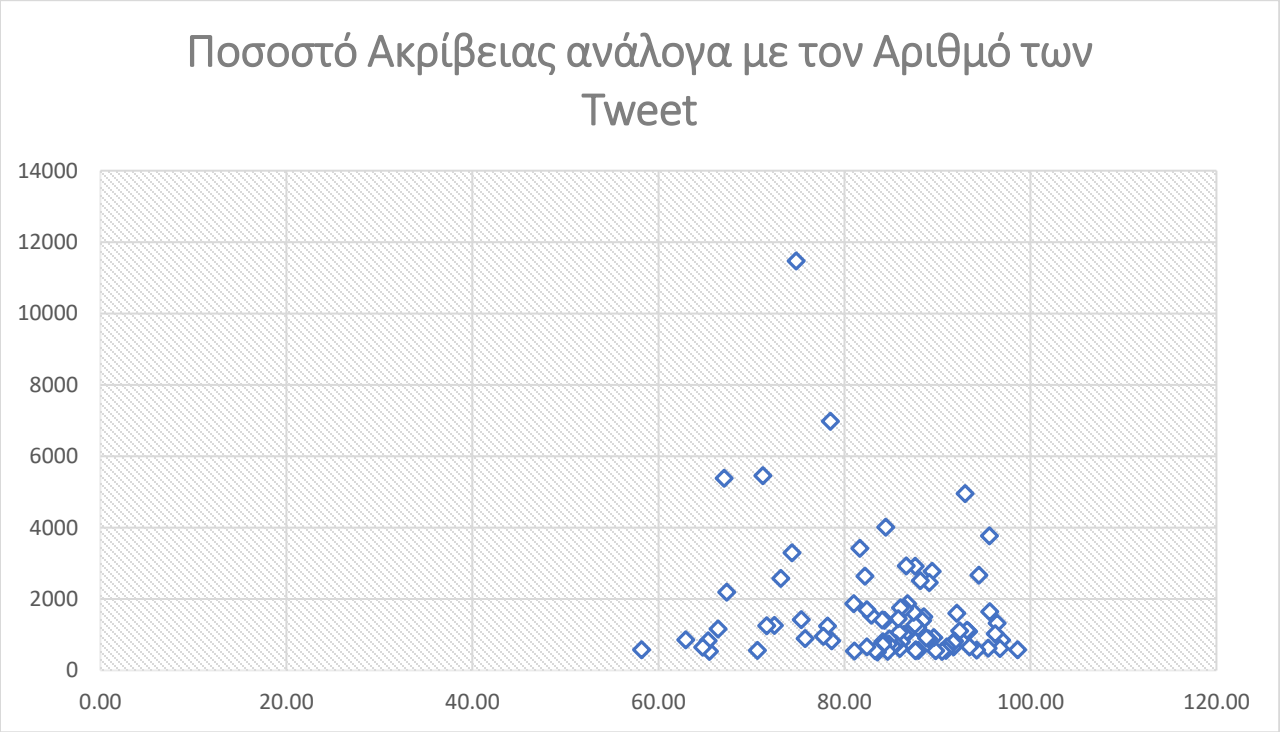
3.4.4 QR ΚΩΔΙΚΟΣ ΓΙΑ ΤΗΝ ΠΡΟΣΒΑΣΗ ΣΤΗΝ ΕΦΑΡΜΟΓΗ

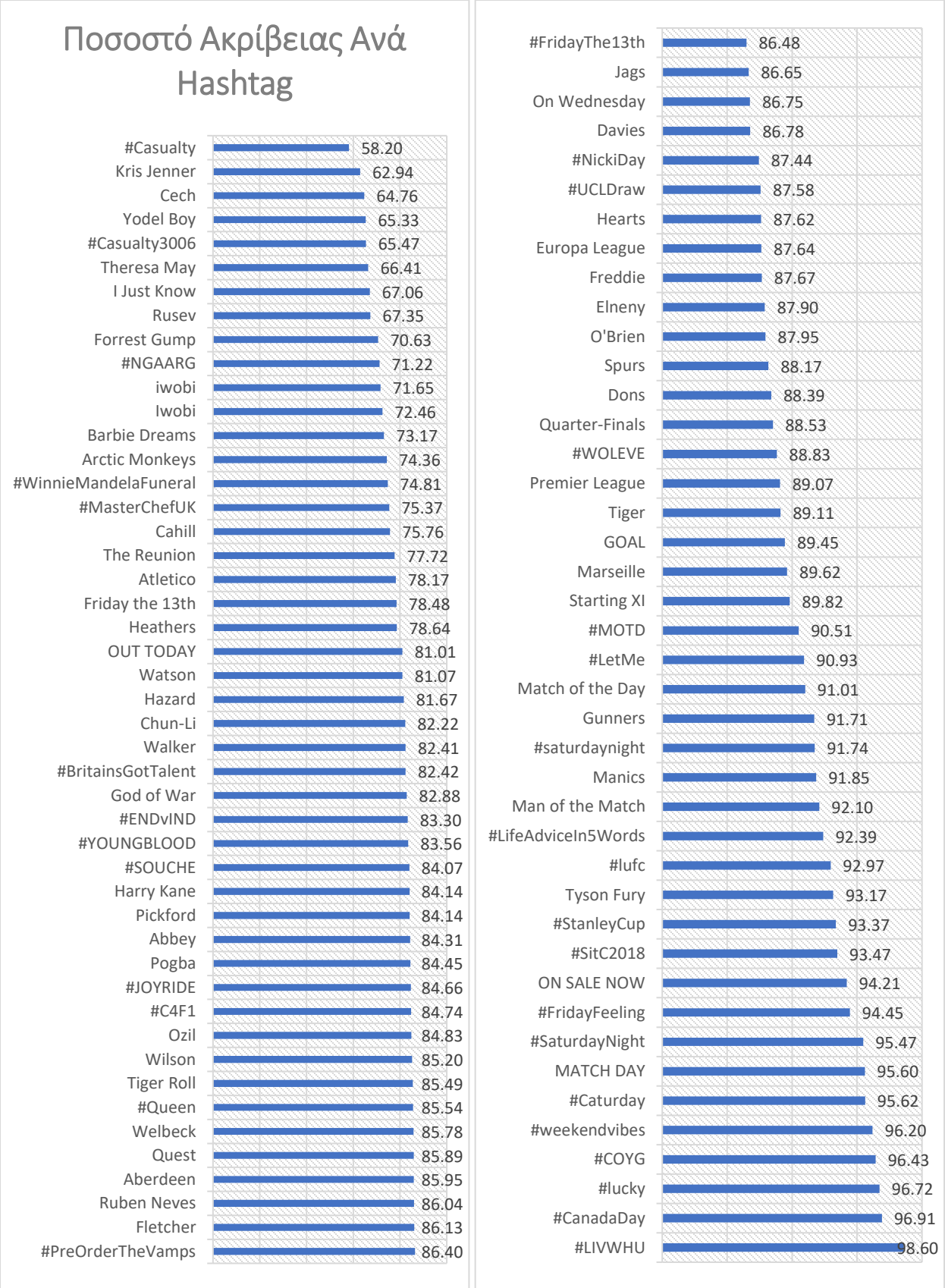
4 ΑΠΟΤΕΛΕΣΜΑΤΑ ΑΝΑ ΑΛΓΟΡΙΘΜΟ

4.1 ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΟΥ MULTINOMIAL NAÏVE BAYES

4.1.1 ΑΝΑΠΑΡΑΣΤΑΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ VECTOR SPACE MODEL

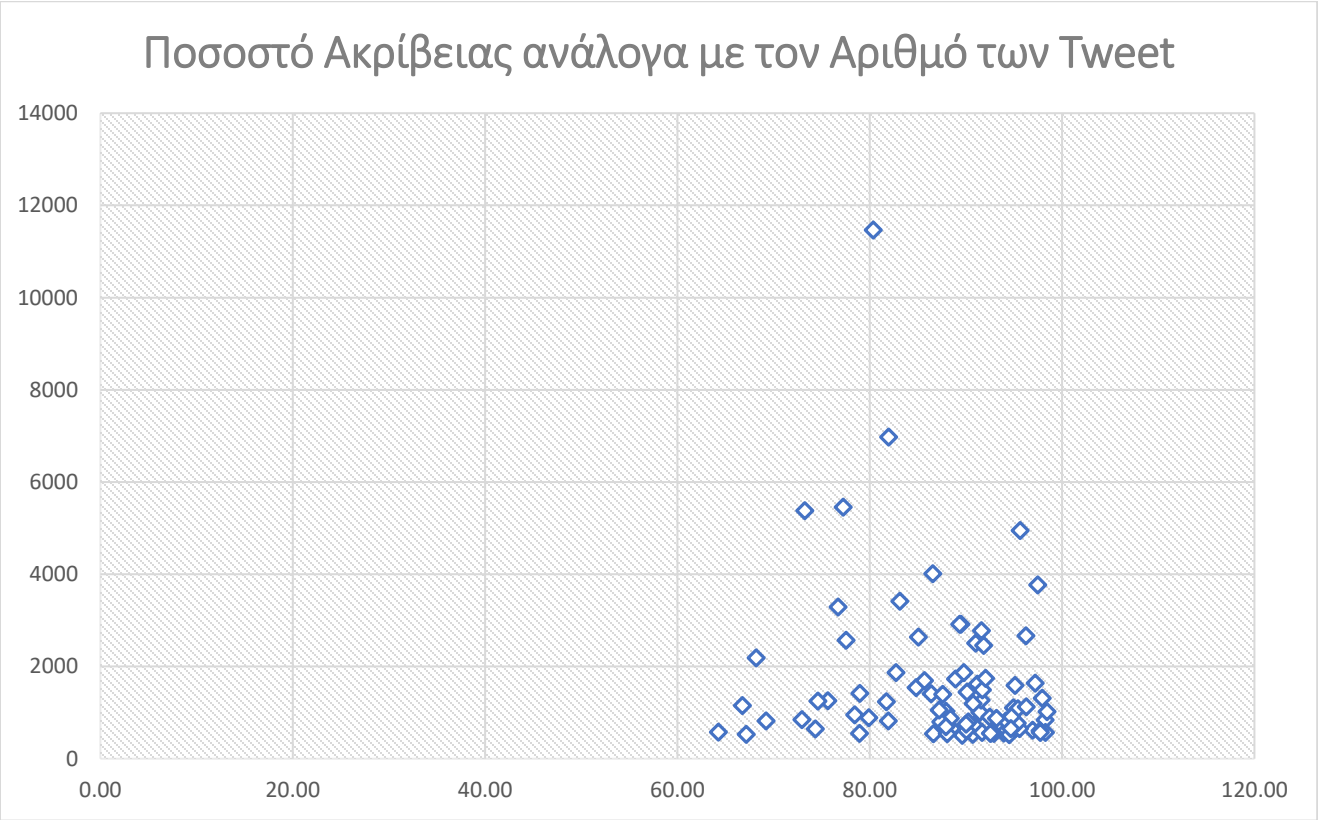
Χρησιμοποιήθηκε ο ταξινομητής Multinomial Naïve Bayes της βιβλιοθήκης του WEKA για την JAVA. Η αναπαράσταση των δεδομένων έγινε με vector space model και τα βάρη για την κάθε λέξη τέθηκαν με βάση το TF-IDF. Επίσης, χρησιμοποιήθηκαν n-grams και έγιναν πειράματα για να βρεθούν τα n-grams που δίνουν το μεγαλύτερο accuracy. Την καλύτερη απόδοση είχαν τα n-grams από 1 έως 2. Το μοντέλο έδωσε ακρίβεια της τάξεως του 84.32% (μέση ακρίβεια για όλα τα Hashtag).

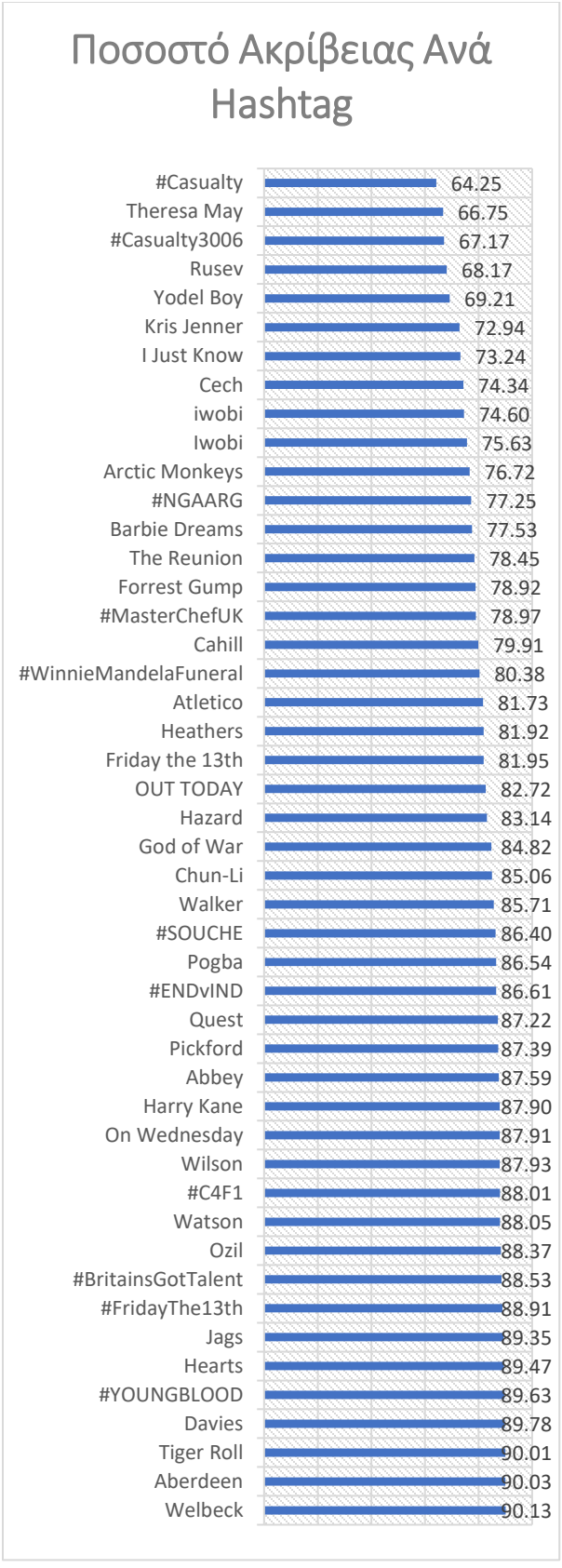



Ποσοστό Ακρίβειας Ανά Hashtag
#Casualty 58.20
Kris Jenner 62.94
Cech 64.76
Yodel Boy 65.33
#Casualty3006 65.47
Theresa May 66.41
I Just Know 67.06
Rusev 67.35
Forrest Gump 70.63
#NGAARG 71.22
iwobi 71.65
Iwobi 72.46
Barbie Dreams 73.17
Arctic Monkeys 74.36
#WinnieMandelaFuneral 74.81
#MasterChefUK 75.37
Cahill 75.76
The Reunion 77.72
Atletico 78.17
Friday the 13th 78.48
Heathers 78.64
OUT TODAY 81.01
Watson 81.07
Hazard 81.67
Chun-Li 82.22
Walker 82.41
#BritainsGotTalent 82.42
God of War 82.88
#ENDvIND 83.30
#YOUNGBLOOD 83.56
#SOUCHE 84.07
Harry Kane 84.14
Pickford 84.14
Abbey 84.31
Pogba 84.45
#JOYRIDE 84.66
#C4F1 84.74
Ozil 84.83
Wilson 85.20
Tiger Roll 85.49
#Queen 85.54
Welbeck 85.78
Quest 85.89
Aberdeen 85.95
Ruben Neves 86.04
Fletcher 86.13
#PreOrderTheVamps 86.40
#FridayThe13th 86.48
Jags 86.65
On Wednesday 86.75
Davies 86.78
#NickiDay 87.44
#UCLDraw 87.58
Hearts 87.62
Europa League 87.64
Freddie 87.67
Elneny 87.90
O'Brien 87.95
Spurs 88.17
Dons 88.39
Quarter-Finals 88.53
#WOLEVE 88.83
Premier League 89.07
Tiger 89.11
GOAL 89.45
Marseille 89.62
Starting XI 89.82
#MOTD 90.51
#LetMe 90.93
Match of the Day 91.01
Gunners 91.71
#saturdaynight 91.74
Manics 91.85
Man of the Match 92.10
#LifeAdviceIn5Words 92.39
#lufc 92.97
Tyson Fury 93.17
#StanleyCup 93.37
#SitC2018 93.47
ON SALE NOW 94.21
#FridayFeeling 94.45
#SaturdayNight 95.47
MATCH DAY 95.60
#Caturday 95.62
#weekendvibes 96.20
#COYG 96.43
#lucky 96.72
#CanadaDay 96.91
#LIVWHU 98.60

4.1.2 Αναπαραστάση Δεδομένων με BOW

Χρησιμοποιήθηκε ξανά ο Multinomial Naïve Bayes, αλλά αυτή τη φορά με Boolean μέθοδο αναπαράστασης των δεδομένων. Τα αποτελέσματα ήταν καλύτερα, με μέση ακρίβεια της τάξεως του 87,75%.

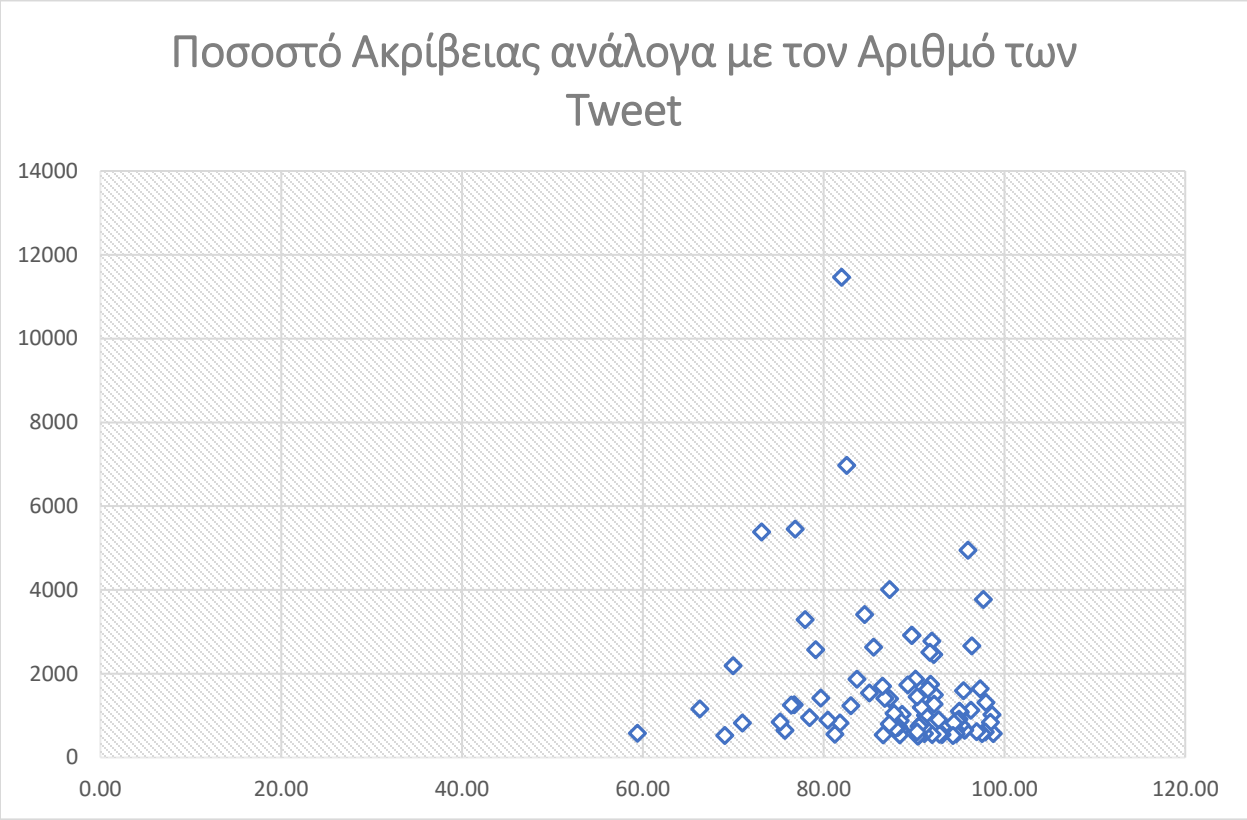


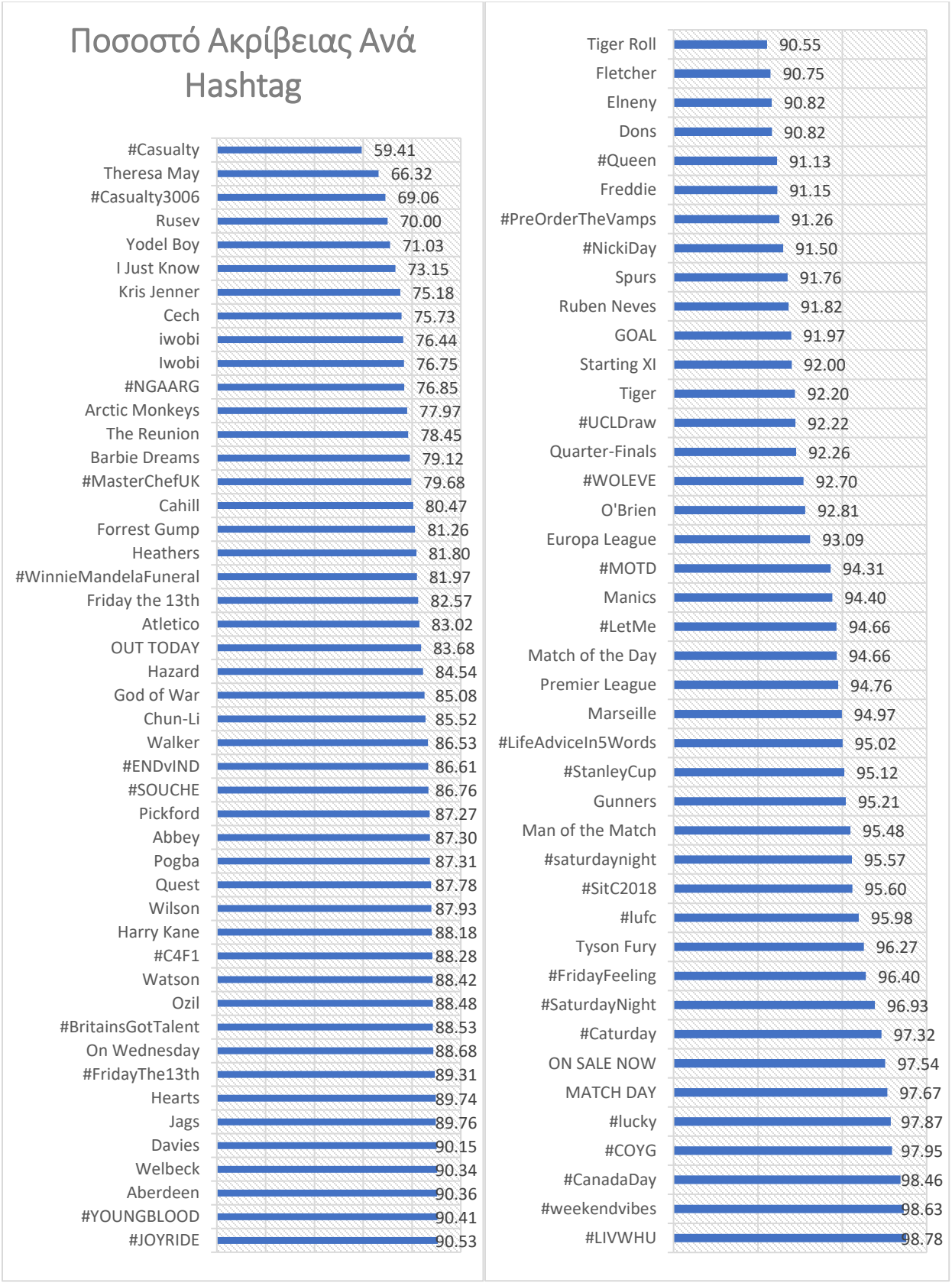

Ποσοστό Ακρίβειας Ανά Hashtag
#Casualty 64.25
Theresa May 66.75
#Casualty3006 67.17
Rusev 68.17
Yodel Boy 69.21
Kris Jenner 72.94
I Just Know 73.24
Cech 74.34
iwobi 74.60
Iwobi 75.63
Arctic Monkeys 76.72
#NGAARG 77.25
Barbie Dreams 77.53
The Reunion 78.45
Forrest Gump 78.92
#MasterChefUK 78.97
Cahill 79.91
#WinnieMandelaFuneral 80.38
Atletico 81.73
Heathers 81.92
Friday the 13th 81.95
OUT TODAY 82.72
Hazard 83.14
God of War 84.82
Chun-Li 85.06
Walker 85.71
#SOUCHE 86.40
Pogba 86.54
#ENDvIND 86.61
Quest 87.22
Pickford 87.39
Abbey 87.59
Harry Kane 87.90
On Wednesday 87.91
Wilson 87.93
#C4F1 88.01
Watson 88.05
Ozil 88.37
#BritainsGotTalent 88.53
#FridayThe13th 88.91
Jags 89.35
Hearts 89.47
#YOUNGBLOOD 89.63
Davies 89.78
Tiger Roll 90.01
Aberdeen 90.03
Welbeck 90.13
Fletcher 90.25
Dons 90.68
#JOYRIDE 90.72
Elneny 90.73
#Queen 90.86
Spurs 91.00
#NickiDay 91.19
#PreOrderTheVamps 91.46
#UCLDraw 91.59
GOAL 91.61
Freddie 91.67
Quarter-Finals 91.73
Tiger 91.83
Ruben Neves 92.05
#WOLEVE 92.48
Starting XI 92.55
O'Brien 92.63
Europa League 92.91
Premier League 93.17
#LetMe 93.95
Manics 94.16
#MOTD 94.50
Match of the Day 94.66
Marseille 94.75
#LifeAdviceIn5Words 94.93
Man of the Match 95.11
#SitC2018 95.14
Gunners 95.34
#StanleyCup 95.40
#saturdaynight 95.57
#lufc 95.64
#FridayFeeling 96.25
Tyson Fury 96.27
#SaturdayNight 96.93
#Caturday 97.20
MATCH DAY 97.48
#lucky 97.70
ON SALE NOW 97.72
#COYG 97.95
#CanadaDay 98.22
#LIVWHU 98.25
#weekendvibes 98.44

4.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΟΥ SVM

4.2.1 ΑΝΑΠΑΡΑΣΤΑΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ VECTOR SPACE MODEL

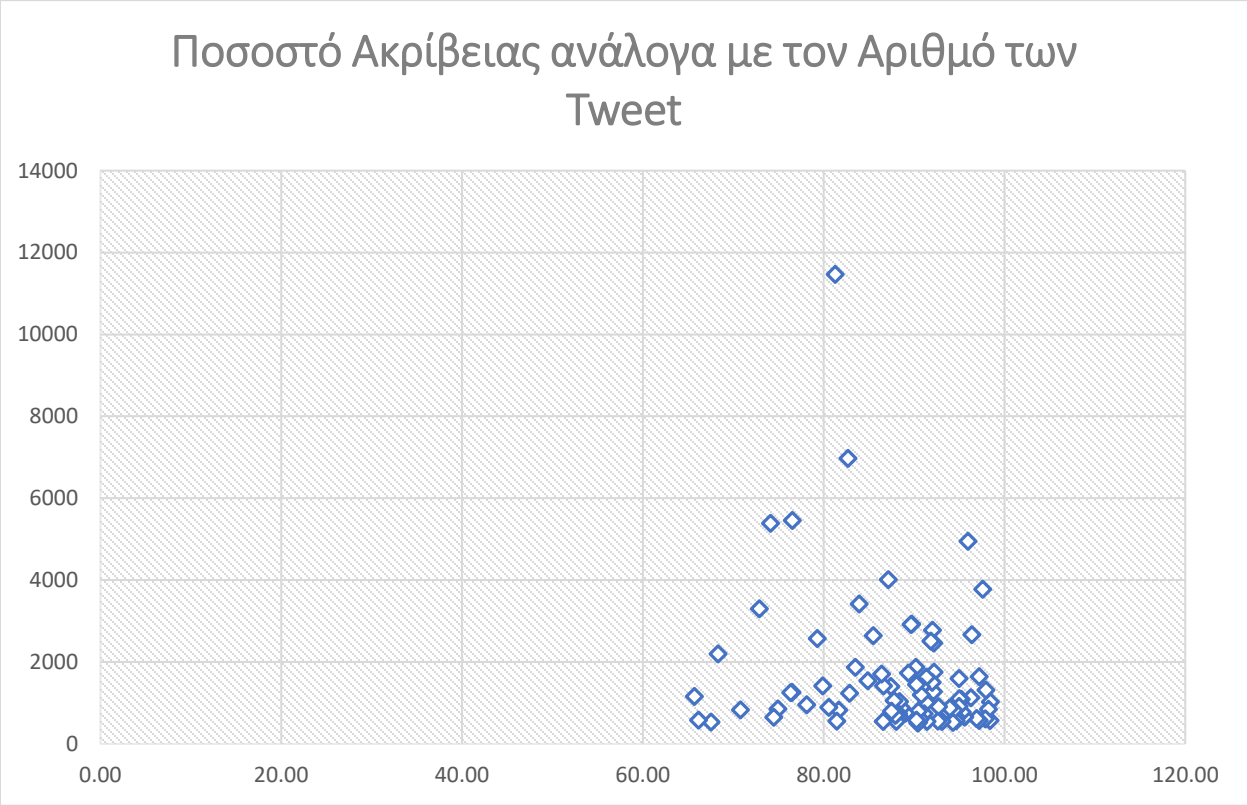
Ως ταξινομητής χρησιμοποιήθηκε ο SVM με Radial Basis Function Kernel της βιβλιοθήκης LibSVM για την JAVA. Έγιναν πολλά πειράματα για να βρεθεί το καλύτερο cost, το οποίο αποδείχθηκε πως είναι το 16. Με n-grams από 1-2 και TF-IDF ο αλγόριθμος δίνει τη μεγαλύτερη μέση ακρίβεια της τάξεως του 88,13%.

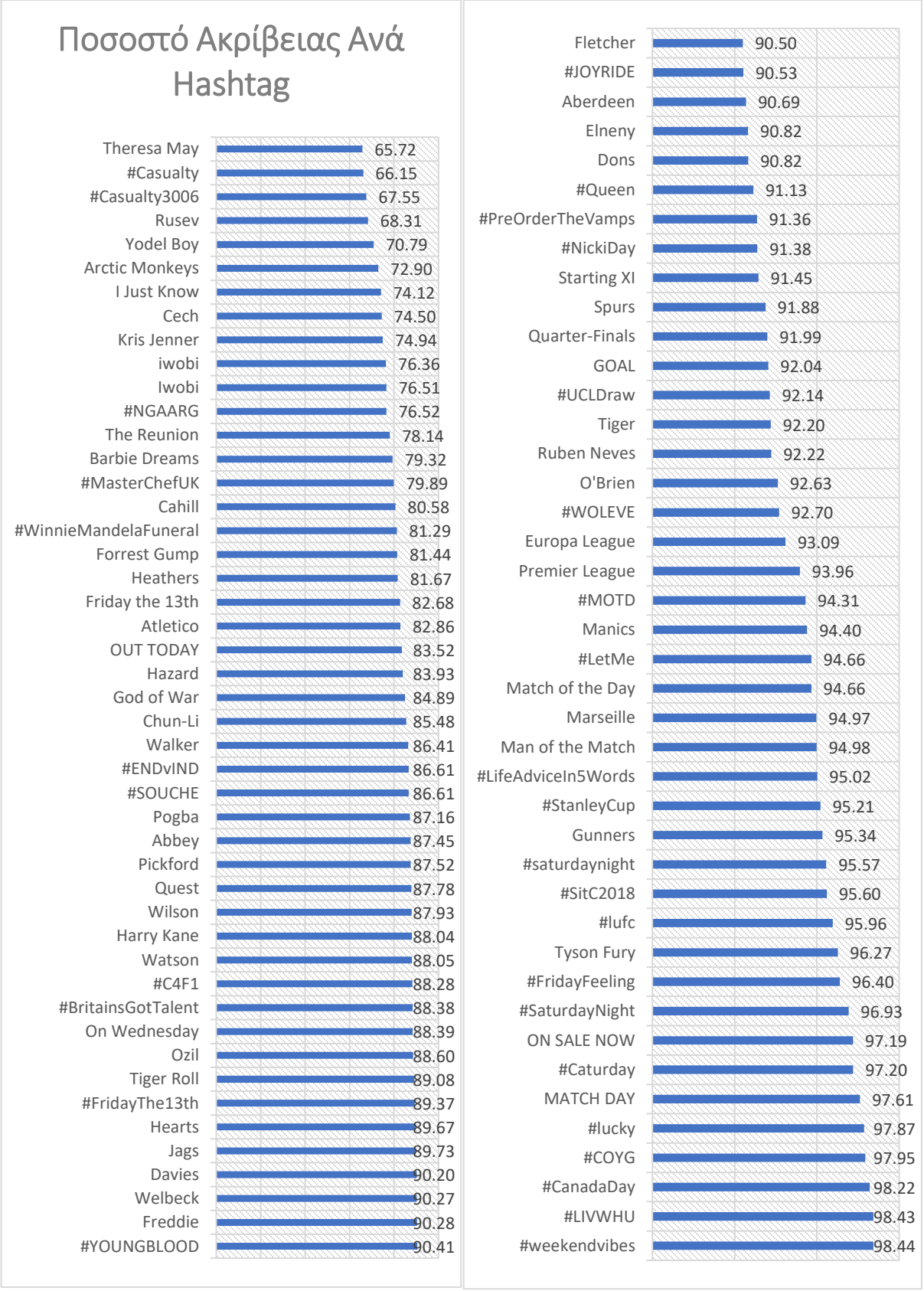



Ποσοστό Ακρίβειας Ανά Hashtag
#Casualty 59.41
Theresa May 66.32
#Casualty3006 69.06
Rusev 70.00
Yodel Boy 71.03
I Just Know 73.15
Kris Jenner 75.18
Cech 75.73
iwobi 76.44
Iwobi 76.75
#NGAARG 76.85
Arctic Monkeys 77.97
The Reunion 78.45
Barbie Dreams 79.12
#MasterChefUK 79.68
Cahill 80.47
Forrest Gump 81.26
Heathers 81.80
#WinnieMandelaFuneral 81.97
Friday the 13th 82.57
Atletico 83.02
OUT TODAY 83.68
Hazard 84.54
God of War 85.08
Chun-Li 85.52
Walker 86.53
#ENDvIND 86.61
#SOUCHE 86.76
Pickford 87.27
Abbey 87.30
Pogba 87.31
Quest 87.78
Wilson 87.93
Harry Kane 88.18
#C4F1 88.28
Watson 88.42
Ozil 88.48
#BritainsGotTalent 88.53
On Wednesday 88.68
#FridayThe13th 89.31
Hearts 89.74
Jags 89.76
Davies 90.15
Welbeck 90.34
Aberdeen 90.36
#YOUNGBLOOD 90.41
#JOYRIDE 90.53
Tiger Roll 90.55
Fletcher 90.75
Elneny 90.82
Dons 90.82
#Queen 91.13
Freddie 91.15
#PreOrderTheVamps 91.26
#NickiDay 91.50
Spurs 91.76
Ruben Neves 91.82
GOAL 91.97
Starting XI 92.00
Tiger 92.20
#UCLDraw 92.22
Quarter-Finals 92.26
#WOLEVE 92.70
O'Brien 92.81
Europa League 93.09
#MOTD 94.31
Manics 94.40
#LetMe 94.66
Match of the Day 94.66
Premier League 94.76
Marseille 94.97
#LifeAdviceIn5Words 95.02
#StanleyCup 95.12
Gunners 95.21
Man of the Match 95.48
#saturdaynight 95.57
#SitC2018 95.60
#lufc 95.98
Tyson Fury 96.27
#FridayFeeling 96.40
#SaturdayNight 96.93
#Caturday 97.32
ON SALE NOW 97.54
MATCH DAY 97.67
#lucky 97.87
#COYG 97.95
#CanadaDay 98.46
#weekendvibes 98.63
#LIVWHU 98.78

4.2.2 Αναπαραστάση Δεδομενων με BOW

Με την Boolean μέθοδο και cost 16, ο αλγόριθμος δίνει μειωμένη ακρίβεια της τάξεως του 88,01%

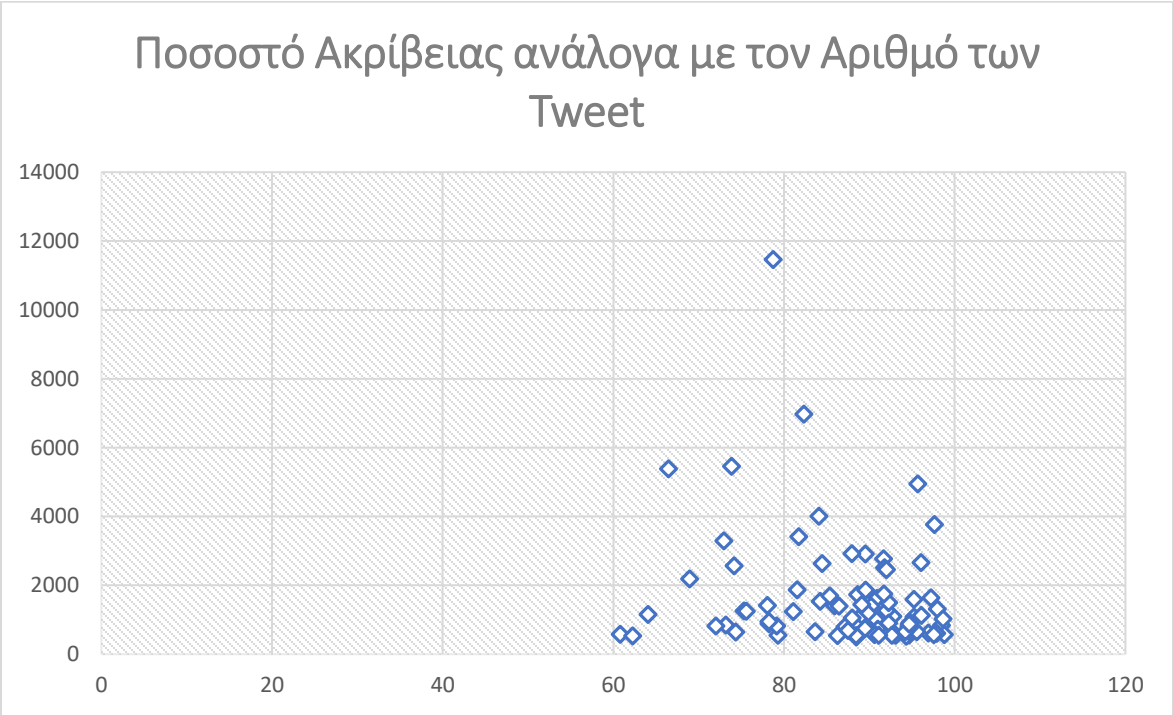


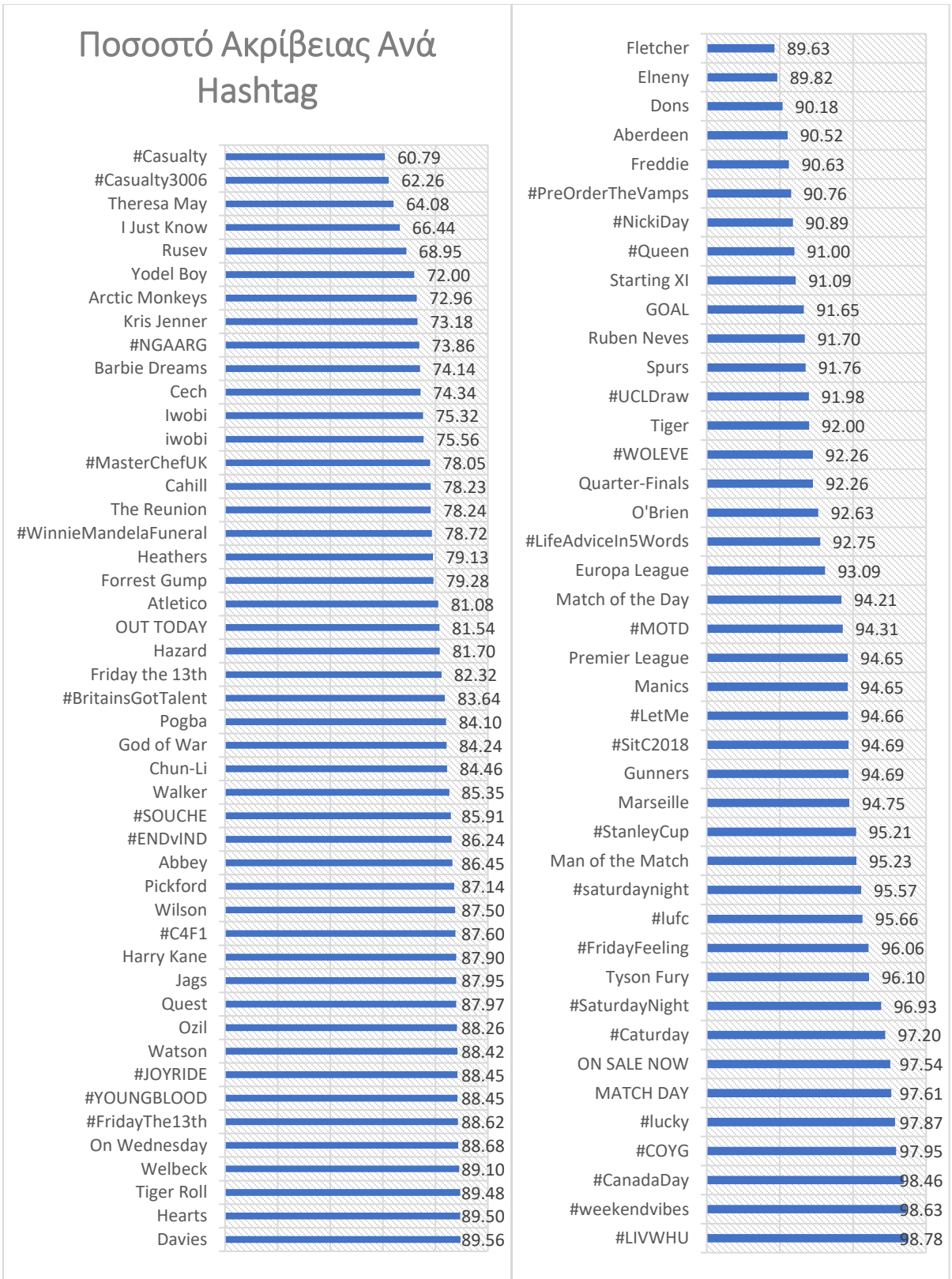

Ποσοστό Ακρίβειας Ανά Hashtag
Theresa May 65.72
#Casualty 66.15
#Casualty3006 67.55
Rusev 68.31
Yodel Boy 70.79
Arctic Monkeys 72.90
I Just Know 74.12
Cech 74.50
Kris Jenner 74.94
iwobi 76.36
Iwobi 76.51
#NGAARG 76.52
The Reunion 78.14
Barbie Dreams 79.32
#MasterChefUK 79.89
Cahill 80.58
#WinnieMandelaFuneral 81.29
Forrest Gump 81.44
Heathers 81.67
Friday the 13th 82.68
Atletico 82.86
OUT TODAY 83.52
Hazard 83.93
God of War 84.89
Chun-Li 85.48
Walker 86.41
#ENDvIND 86.61
#SOUCHE 86.61
Pogba 87.16
Abbey 87.45
Pickford 87.52
Quest 87.78
Wilson 87.93
Harry Kane 88.04
Watson 88.05
#C4F1 88.28
#BritainsGotTalent 88.38
On Wednesday 88.39
Ozil 88.60
Tiger Roll 89.08
#FridayThe13th 89.37
Hearts 89.67
Jags 89.73
Davies 90.20
Welbeck 90.27
Freddie 90.28
#YOUNGBLOOD 90.41
Fletcher 90.50
#JOYRIDE 90.53
Aberdeen 90.69
Elneny 90.82
Dons 90.82
#Queen 91.13
#PreOrderTheVamps 91.36
#NickiDay 91.38
Starting XI 91.45
Spurs 91.88
Quarter-Finals 91.99
GOAL 92.04
#UCLDraw 92.14
Tiger 92.20
Ruben Neves 92.22
O'Brien 92.63
#WOLEVE 92.70
Europa League 93.09
Premier League 93.96
#MOTD 94.31
Manics 94.40
#LetMe 94.66
Match of the Day 94.66
Marseille 94.97
Man of the Match 94.98
#LifeAdviceIn5Words 95.02
#StanleyCup 95.21
Gunners 95.34
#saturdaynight 95.57
#SitC2018 95.60
#lufc 95.96
Tyson Fury 96.27
#FridayFeeling 96.40
#SaturdayNight 96.93
ON SALE NOW 97.19
#Caturday 97.20
MATCH DAY 97.61
#lucky 97.87
#COYG 97.95
#CanadaDay 98.22
#LIVWHU 98.43
#weekendvibes 98.44

4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΤΟΝ K-NN

4.3.1 ΑΝΑΠΑΡΑΣΤΑΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ VECTOR SPACE MODEL

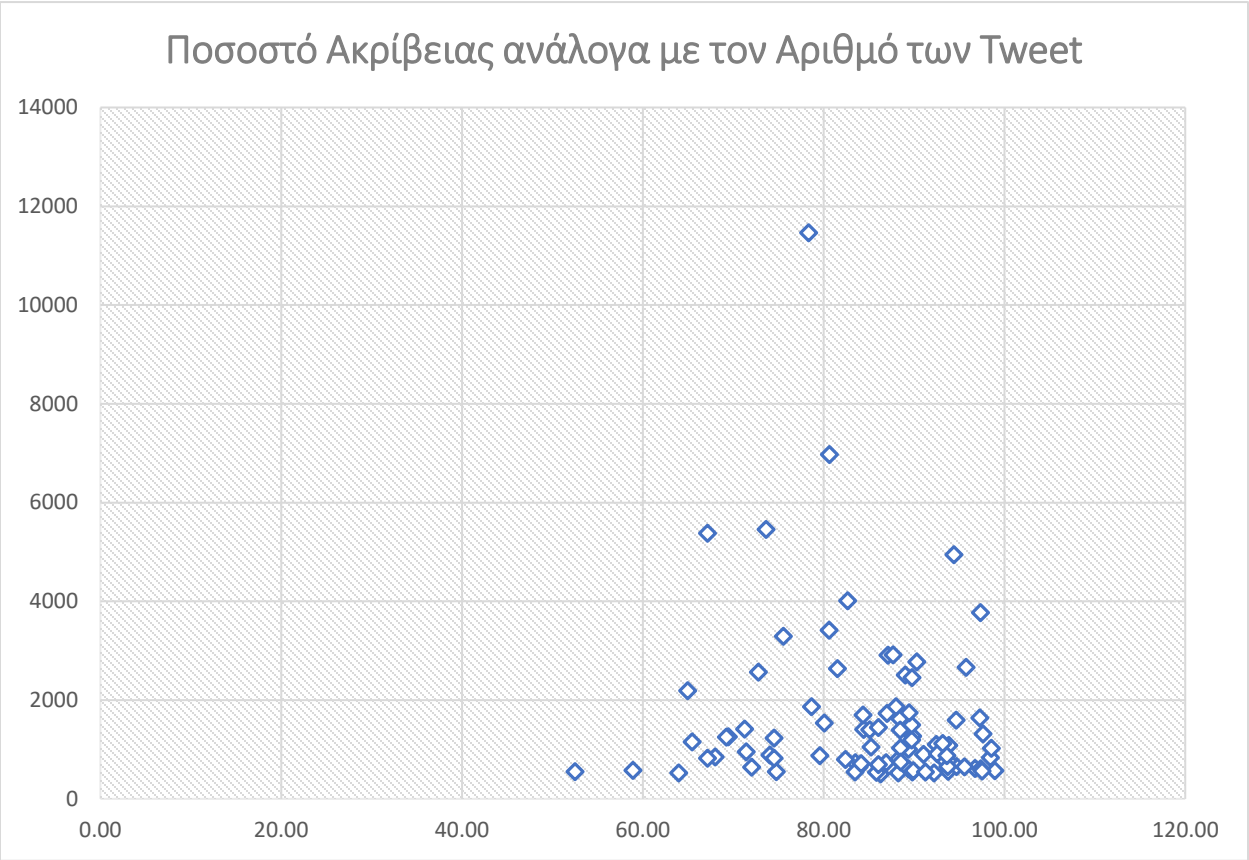
Ως ταξινομητής χρησιμοποιήθηκε ο IBK της βιβλιοθήκης WEKA για την JAVA. Ο αριθμός k επιλέγεται αυτόματα μετά από 10-fold cross validation με διαφορετικές τιμές ώστε να βρεθεί εκείνη που δίνει μεγαλύτερη ακρίβεια. Για να δοθούν βάρη στις λέξεις χρησιμοποιείται TF-IDF. Επίσης για τη μετατροπή των λέξεων σε features χρησιμοποιούνται n-grams. Με n-grams από 1 έως 8 η μέση ακρίβεια είναι 87,13%.

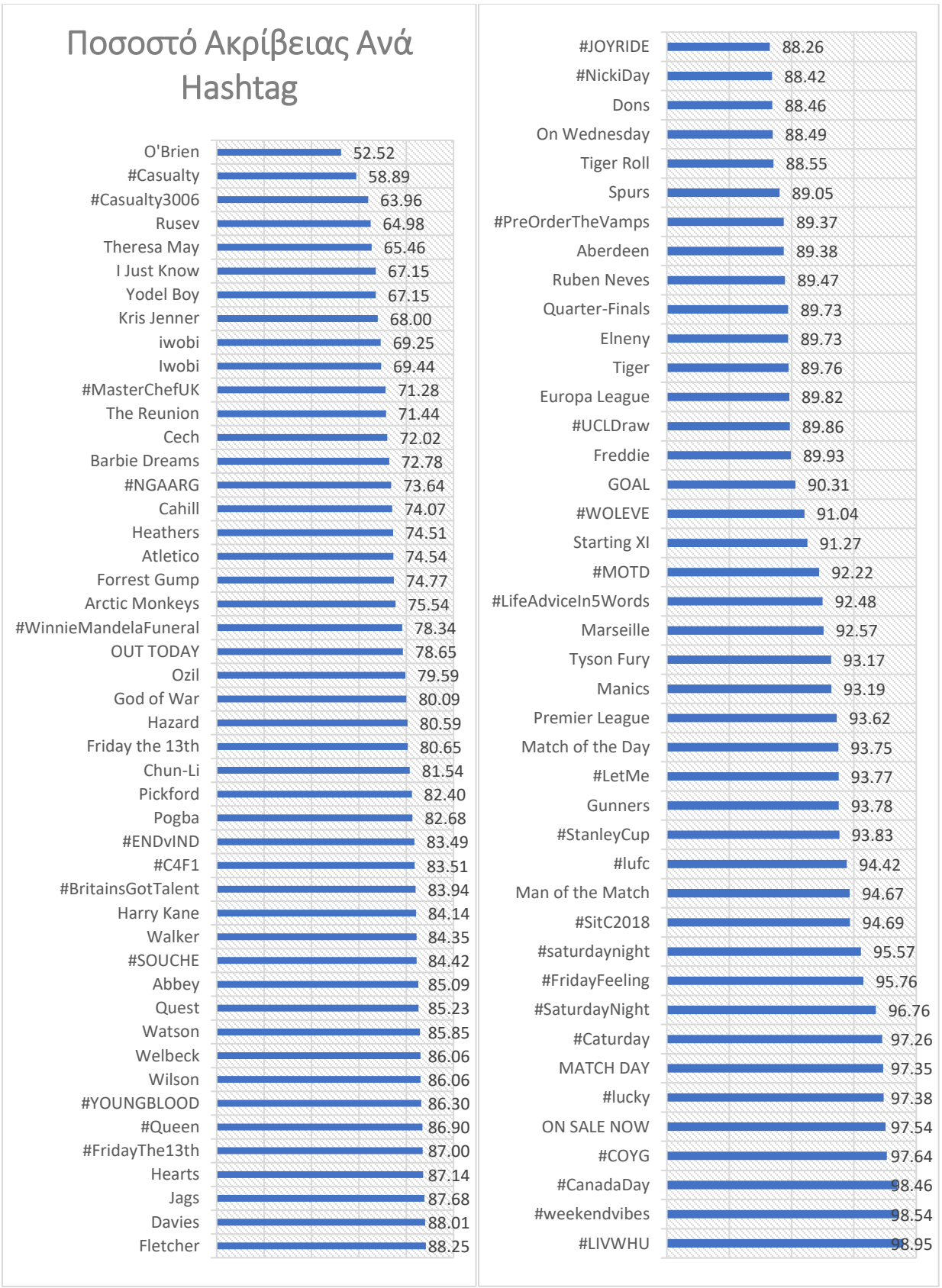



Ποσοστό Ακρίβειας Ανά Hashtag
#Casualty 60.79
#Casualty3006 62.26
Theresa May 64.08
I Just Know 66.44
Rusev 68.95
Yodel Boy 72.00
Arctic Monkeys 72.96
Kris Jenner 73.18
#NGAARG 73.86
Barbie Dreams 74.14
Cech 74.34
Iwobi 75.32
iwobi 75.56
#MasterChefUK 78.05
Cahill 78.23
The Reunion 78.24
#WinnieMandelaFuneral 78.72
Heathers 79.13
Forrest Gump 79.28
Atletico 81.08
OUT TODAY 81.54
Hazard 81.70
Friday the 13th 82.32
#BritainsGotTalent 83.64
Pogba 84.10
God of War 84.24
Chun-Li 84.46
Walker 85.35
#SOUCHE 85.91
#ENDvIND 86.24
Abbey 86.45
Pickford 87.14
Wilson 87.50
#C4F1 87.60
Harry Kane 87.90
Jags 87.95
Quest 87.97
Ozil 88.26
Watson 88.42
#JOYRIDE 88.45
#YOUNGBLOOD 88.45
#FridayThe13th 88.62
On Wednesday 88.68
Welbeck 89.10
Tiger Roll 89.48
Hearts 89.50
Davies 89.56
Fletcher 89.63
Elneny 89.82
Dons 90.18
Aberdeen 90.52
Freddie 90.63
#PreOrderTheVamps 90.76
#NickiDay 90.89
#Queen 91.00
Starting XI 91.09
GOAL 91.65
Ruben Neves 91.70
Spurs 91.76
#UCLDraw 91.98
Tiger 92.00
#WOLEVE 92.26
Quarter-Finals 92.26
O'Brien 92.63
#LifeAdviceIn5Words 92.75
Europa League 93.09
Match of the Day 94.21
#MOTD 94.31
Premier League 94.65
Manics 94.65
#LetMe 94.66
#SitC2018 94.69
Gunners 94.69
Marseille 94.75
#StanleyCup 95.21
Man of the Match 95.23
#saturdaynight 95.57
#lufc 95.66
#FridayFeeling 96.06
Tyson Fury 96.10
#SaturdayNight 96.93
#Caturday 97.20
ON SALE NOW 97.54
MATCH DAY 97.61
#lucky 97.87
#COYG 97.95
#CanadaDay 98.46
#weekendvibes 98.63
#LIVWHU 98.78

4.3.2 Αναπαραστάση Δεδομενων με BOW

Με τη Boolean μέθοδο αναπαράστασης η μέση ακρίβεια ήταν 84,80.



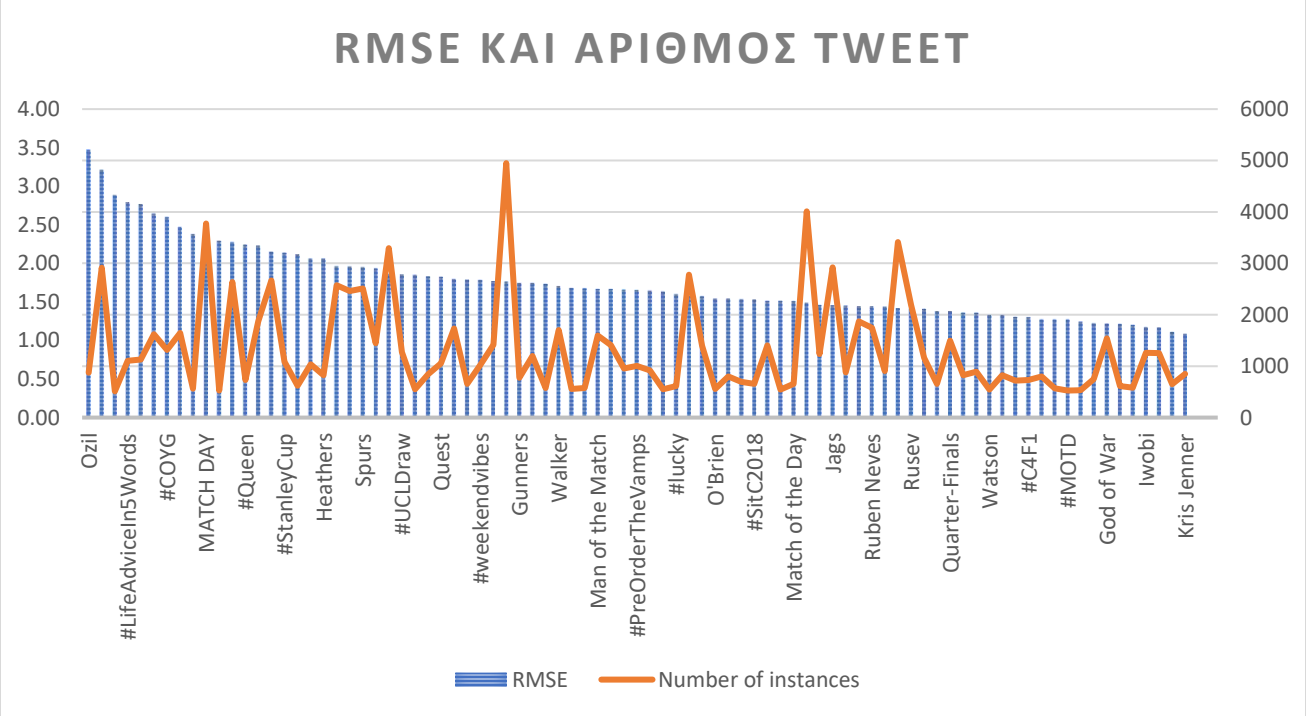

Ποσοστό Ακρίβειας Ανά Hashtag
O'Brien 52.52
#Casualty 58.89
#Casualty3006 63.96
Rusev 64.98
Theresa May 65.46
I Just Know 67.15
Yodel Boy 67.15
Kris Jenner 68.00
iwobi 69.25
Iwobi 69.44
#MasterChefUK 71.28
The Reunion 71.44
Cech 72.02
Barbie Dreams 72.78
#NGAARG 73.64
Cahill 74.07
Heathers 74.51
Atletico 74.54
Forrest Gump 74.77
Arctic Monkeys 75.54
#WinnieMandelaFuneral 78.34
OUT TODAY 78.65
Ozil 79.59
God of War 80.09
Hazard 80.59
Friday the 13th 80.65
Chun-Li 81.54
Pickford 82.40
Pogba 82.68
#ENDvIND 83.49
#C4F1 83.51
#BritainsGotTalent 83.94
Harry Kane 84.14
Walker 84.35
#SOUCHE 84.42
Abbey 85.09
Quest 85.23
Watson 85.85
Welbeck 86.06
Wilson 86.06
#YOUNGBLOOD 86.30
#Queen 86.90
#FridayThe13th 87.00
Hearts 87.14
Jags 87.68
Davies 88.01
Fletcher 88.25
#JOYRIDE 88.26
#NickiDay 88.42
Dons 88.46
On Wednesday 88.49
Tiger Roll 88.55
Spurs 89.05
#PreOrderTheVamps 89.37
Aberdeen 89.38
Ruben Neves 89.47
Quarter-Finals 89.73
Elneny 89.73
Tiger 89.76
Europa League 89.82
#UCLDraw 89.86
Freddie 89.93
GOAL 90.31
#WOLEVE 91.04
Starting XI 91.27
#MOTD 92.22
#LifeAdviceIn5Words 92.48
Marseille 92.57
Tyson Fury 93.17
Manics 93.19
Premier League 93.62
Match of the Day 93.75
#LetMe 93.77
Gunners 93.78
#StanleyCup 93.83
#lufc 94.42
Man of the Match 94.67
#SitC2018 94.69
#saturdaynight 95.57
#FridayFeeling 95.76
#SaturdayNight 96.76
#Caturday 97.26
MATCH DAY 97.35
#lucky 97.38
ON SALE NOW 97.54
#COYG 97.64
#CanadaDay 98.46
#weekendvibes 98.54
#LIVWHU 98.95

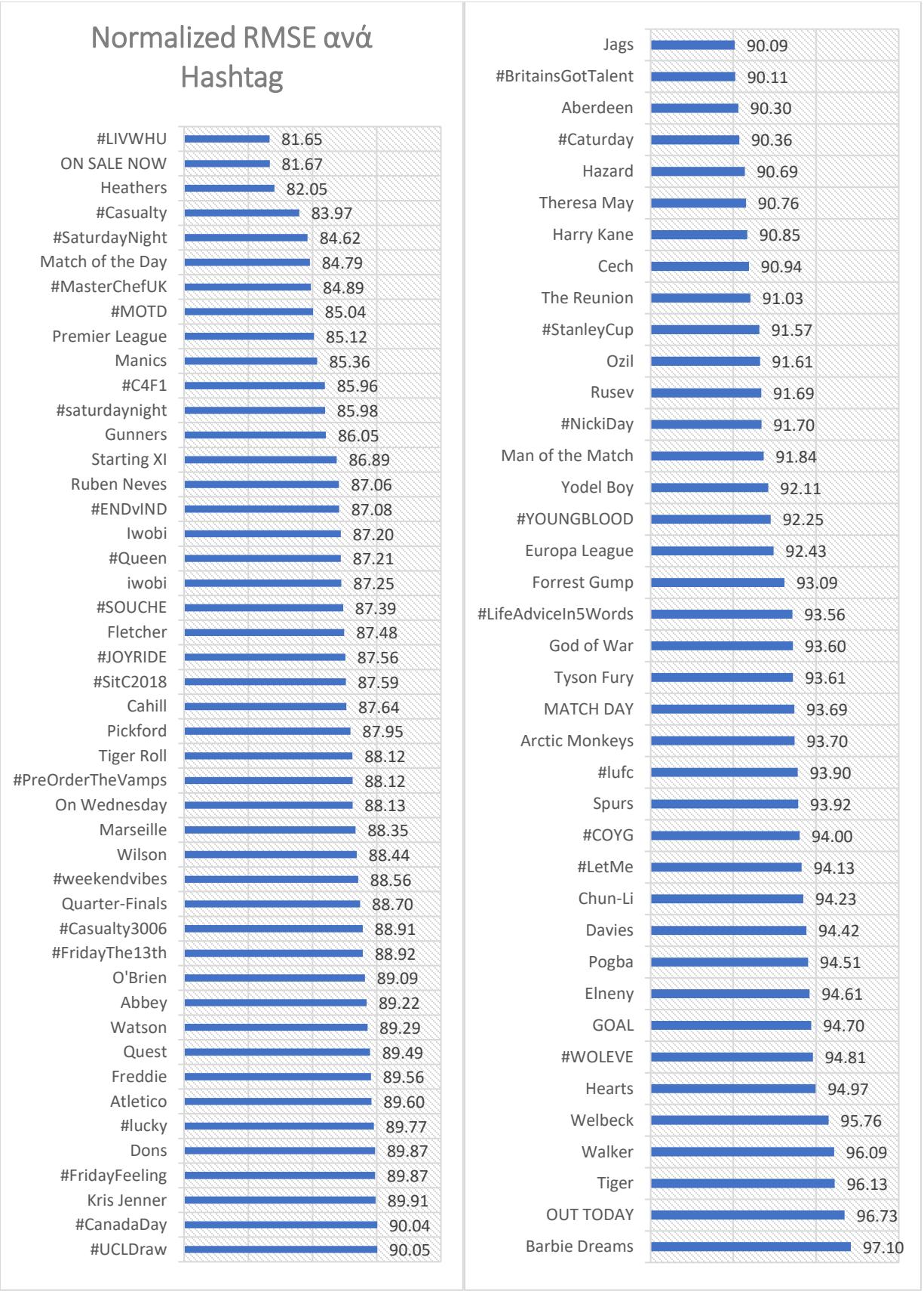
4.4 ΑΠΟΤΕΛΕΣΜΑΤΑ LINEAR REGRESSION

Για την υλοποίηση χρησιμοποιήσαμε τη βιβλιοθήκη WEKA για την JAVA. Επειδή η απλή γραμμική παλινδρόμηση δεν έδινε τόσο καλά αποτελέσματα, αναζητήσαμε μια πιο υβριδική εκδοχή. Γι' αυτό χρησιμοποιήσαμε τη μέθοδο Locally Weighted Learning [27], που πραγματοποιεί Γραμμική Παλινδρόμηση με αλγόριθμο εύρεσης κοντινότερων γειτόνων τον KD-tree και k 10.

Για την εκτίμηση των αποτελεσμάτων χρησιμοποιήσαμε το RMSE και μία normalized εκδοχή του, που υπολογίζεται ως εξής:

$$\left(1 - \left(\frac{RMSE}{\text{MaxSentimentScore} - \text{MinSentimentScore}}\right)\right) * 100$$




Normalized RMSE ανά Hashtag
#LIVWHU 81.65
ON SALE NOW 81.67
Heathers 82.05
#Casualty 83.97
#SaturdayNight 84.62
Match of the Day 84.79
#MasterChefUK 84.89
#MOTD 85.04
Premier League 85.12
Manics 85.36
#C4F1 85.96
#saturdaynight 85.98
Gunners 86.05
Starting XI 86.89
Ruben Neves 87.06
#ENDvIND 87.08
Iwobi 87.20
#Queen 87.21
iwobi 87.25
#SOUCHE 87.39
Fletcher 87.48
#JOYRIDE 87.56
#SitC2018 87.59
Cahill 87.64
Pickford 87.95
Tiger Roll 88.12
#PreOrderTheVamps 88.12
On Wednesday 88.13
Marseille 88.35
Wilson 88.44
#weekendvibes 88.56
Quarter-Finals 88.70
#Casualty3006 88.91
#FridayThe13th 88.92
O'Brien 89.09
Abbey 89.22
Watson 89.29
Quest 89.49
Freddie 89.56
Atletico 89.60
#lucky 89.77
Dons 89.87
#FridayFeeling 89.87
Kris Jenner 89.91
#CanadaDay 90.04
#UCLDraw 90.05
Jags 90.09
#BritainsGotTalent 90.11
Aberdeen 90.30
#Caturday 90.36
Hazard 90.69
Theresa May 90.76
Harry Kane 90.85
Cech 90.94
The Reunion 91.03
#StanleyCup 91.57
Ozil 91.61
Rusev 91.69
#NickiDay 91.70
Man of the Match 91.84
Yodel Boy 92.11
#YOUNGBLOOD 92.25
Europa League 92.43
Forrest Gump 93.09
#LifeAdviceIn5Words 93.56
God of War 93.60
Tyson Fury 93.61
MATCH DAY 93.69
Arctic Monkeys 93.70
#lufc 93.90
Spurs 93.92
#COYG 94.00
#LetMe 94.13
Chun-Li 94.23
Davies 94.42
Pogba 94.51
Elneny 94.61
GOAL 94.70
#WOLEVE 94.81
Hearts 94.97
Welbeck 95.76
Walker 96.09
Tiger 96.13
OUT TODAY 96.73
Barbie Dreams 97.10

5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΑΡΑΤΗΡΗΣΕΙΣ ΣΧΕΤΙΚΑ ΜΕ ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Από τις τέσσερις υλοποιήσεις με στόχο την ανάλυση συναισθήματος σε Tweet τα καλύτερα αποτελέσματα με βάση την ακρίβεια δείχνει να έχει ο SVM και ο K-nn. Ωστόσο, στην πραγματικότητα οι δύο αυτοί αλγόριθμοί δίνουν τη χειρότερη ποιοτικά κατηγοριοποίηση.

Την καλύτερη κατηγοριοποίηση με βάση την παρατήρηση του δείγματος κάνει ο Multinomial Naïve Bayes. Πιο αναλυτικά, το Boolean μοντέλο φαινομενικά έχει μεγαλύτερο accuracy και άρα καλύτερα αποτελέσματα όταν κάποιος επεξεργαστεί τα δείγματα. Ωστόσο, ελέγχοντας τα αποτελέσματα μετά από κάθε παραμετροποίηση αποδείχθηκε ότι ενώ η ακρίβεια είναι υψηλότερη η ανάλυση συναισθήματος δεν έχει επιτύχει στο βαθμό που θα θέλαμε. Τα πιο σωστά αποτελέσματα φαίνεται να δίνει το μοντέλο με TF-IDF και n-grams από 1 έως δύο. Στα περισσότερα hashtag η πρόβλεψη της κλάσης ήταν άνω των προσδοκιών μας, αφού βρίσκονταν ελάχιστα λάθη ανάμεσα στα κατηγοριοποιημένα παραδείγματα.

Τα αποτελέσματα της γραμμικής παλινδρόμησης είναι επίσης πολύ καλά, ωστόσο σε σχέση με τον Multinomial Naïve Bayes ίσως είναι ελάχιστα προς το χειρότερο. Αυτό δεν αφορά μόνο την ακρίβεια, αλλά τα συμπεράσματά μας παρατηρώντας τις προβλέψεις των τιμών του συναισθήματος των Tweet.

Αν και στη βιβλιογραφία ο Support Vector Machine φαίνεται να έχει τα καλύτερα αποτελέσματα σε ό,τι αφορά την ανάλυση συναισθήματος, στην παρούσα εργασία και δεδομένου του training set δεν άγγιξε τις προσδοκίες μας. Πιο αναλυτικά, έχει υψηλότερη ακρίβεια από τον Multinomial Naïve Bayes, αλλά φαίνεται ότι μεροληπτεί υπέρ της υπερισχύουσας κλάσης ανεξάρτητα από τις πολλές παραμετροποιήσεις που δοκιμάσαμε. Αυτό πιθανώς να οφείλεται στο γεγονός ότι τα Tweet έχουν πολύ λίγες λέξεις και το πλεονέκτημα του SVM είναι γνωστό πώς είναι τα μεγάλα κείμενα.

Ο Knn έχει ομολογουμένως τα χειρότερα αποτελέσματα σε σύγκριση με τις άλλες τρεις μεθόδους. Παρατηρώντας το αληθινό δείγμα των Tweet είναι ξεκάθαρο πως η κατηγοριοποίηση που κάνει ωχριά μπροστά σε άλλα εντυπωσιακά αποτελέσματα.

6 ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Λόγω των αποτελεσμάτων που έδωσε ο Multinomial Naïve Bayes και η Γραμμική Παλινδρόμηση, συμπεραίνουμε πως η κατηγοριοποίηση με emoticon είναι δόκιμη. Μάλιστα φαίνεται πως στις περισσότερες περιπτώσεις το μέγεθος του training set δεν χρειάζεται να είναι ιδιαίτερα μεγάλο ώστε να μπορεί να «μάθει» ο ταξινομητής τα παραδείγματα.

Περαιτέρω έρευνα μπορεί να γίνει με διαφορετικούς ταξινομητές και συνδυασμούς μεθόδων. Για παράδειγμα, συνδυασμό ταξινομητή με λεξικογραφικές μεθόδους ανάλυσης συναισθήματος.

Αξίζει να σημειωθεί πως το σύστημα που υλοποιήθηκε σε αυτή την εργασία θα μπορούσε να μετασχηματιστεί σε ανάλυση συναισθήματος των Tweet σε πραγματικό χρόνο, αφού δεν χρειάζεται ανθρώπινη επίβλεψη για τη δημιουργία του training set.

Επιπλέον, θα μπορούσαν να εξεταστούν τα αποτελέσματα αυτής της μεθόδου σε άλλες γλώσσες πέραν της Αγγλικής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, Igor Mozetič. (2015). *Sentiment of Emojis*. PLoS One.

[2] Bo Pang, Lillian Lee. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval, Vol. 2.

[3] C. Fabricius-Hansen, B. Behrens, M. F. Krave. (2006*). Explicit and Implicit Information in Text Information Structure across Languages*. Pre-Proceedings of the SPRIK Conference 2006.

[4] Μπαμπινιώτης, Γ. (1998). *Λεξικό της Νέας Ελληνικής Γλώσσας*. Κέντρο Λεξικολογίας.

[5] V. S. Jagtap, Karishma Pawar. (2013) *Analysis of different approaches to Sentence-Level Sentiment Classification*. International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3

[6] Zhongwu Zhai, Bing Liu, Hua Xu,Hua Xu. (2011). *Clustering Product Features for Opinion Minin*g. WSDM'11

[7] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Proceedings of the International Conference on Language Resources and Evaluation.

[8] Gatti, L., Guerini, M., & Turchi, M. (2016). *SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis*. IEEE Transactions on Affective Computing.

[9] Hatzivassiloglou, V., McKeown, K. (1997). *Predicting the semantic orientation of adjectives*. ACL-EACL.

[10] B. Pang, L. Lee, S. Vaithyanathan, (2002). *"Thumbs up? Sentiment classification using machine learning techniques"*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[11] Jonathon Read (2005). *Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification*. Proceedings of the ACL Student Research Workshop.

[12] Hao Wang, Jorge A. Castanon. (2015). *Sentiment Expression via Emoticons on Social Media*. Proceedings of the 2015 IEEE International Conference on Big Data.

[13] Soroush Vosoughi, Helen Zhou, Deb Roy. (2016). *Enhanced Twitter Sentiment Classification Using Contextual Information*. Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

[14] Jichang Zhao, Li Dong, Junjie Wu, Ke Xu. (2012). MoodLens: *An Emoticon-Based Sentiment Analysis System for Chinese Tweets*. KDD'12.

[15] Geeta. G. Dayalani, B. K. Patil. (2014). *Emoticon-based unsupervised sentiment classifier for polarity analysis in tweets*. International Journal Of Engineering Research and General Science.

[16] Alec Go, Richa Bhayani, Lei Huang. (2009). *Twitter sentiment classification using distant supervision.*

[17] Grant S. Ingersoll Thomas S. Morton Andrew L. Farris. (2013). *Taming Text*. Manning Publications Co

[18] Nitin Hardeniya. (2015). *NLTK Essentials*. Packt Publishing Ltd

[19] Murty M.N., Devi V.S. (2011) *Bayes Classifier In: Pattern Recognition*. Undergraduate Topics in Computer Science, vol 0. Springer, London

[20] Yihong Gong, Wei Xu. (2007) M*ax-Margin Classifications. In: Machine Learning for Multimedia Content Analysi*s. Springer.

[21] Thorsten Joachims. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence

[22] Mucherino A., Papajorgji P.J., Pardalos P.M. (2009) *k-Nearest Neighbor Classification. In: Data Mining in Agriculture*. Springer Optimization and Its Applications, vol 34.

[23] Χαράλαμπος Γναρδέλλης. (2003). *Εφαρμοσμένη Στατιστική*. Εκδόσεις Παπαζήση.

[24] Pang, B.and Lee, L. (2004). *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*. 42nd Meeting of the Association for Computational Linguistics.

[25] C.-C. Chang, C.-J. Lin. (2011). *LIBSVM : a library for support vector machines*. ACM Transactions on Intelligent Systems and Technology.

[26] B. Pang, L. Lee. (2005). *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.* Proceedings of the Association for Computational Linguistics (ACL).

[27] C. Atkeson, A. Moore, S. Schaal (1996). *Locally weighted learning*. AI Review.

[28] S. Das, and M. Chen. (2001*). Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards*. Proceedings of the Asia Pacific Finance Association Annual Conference.

[29] Tong, R.M. (2001*). An Operational System for Detecting and Tracking Opinions in On-Line Discussion*. Proceedings of SIGIR Workshop on Operational Text Classification.

[30] K. Dave, S. Lawrence, D. M. Pennock. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. Proceedings of the 12th international conference on World Wide Web.