ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΝΑΛΥΣΗ ΚΛΙΜΑΤΟΣ ΣΕ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Γεώργιος Ο. Τζουμάνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Φεβρουάριος 2017

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΝΑΛΥΣΗ ΚΛΙΜΑΤΟΣ ΣΕ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ

Γεώργιος Ο. Τζουμάνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Φεβρουάριος 2017

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

-Ν Πελέκης. (Επιβλέπων)

- Μ Κούτρας

- Ελ. Κοφίδης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

SENTIMENT ANALYSIS ON SOCIAL MEDIA

By

George O. Tzoumanis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
February 2017

Στην σύζυγό μου

Νίκη

.

Ευχαριστώ τον επιβλέποντα καθηγητή κύριο Νίκο Πελέκη για την καθοδήγηση και τις υποδείξεις του κατά την διάρκεια της συγγραφής της παρούσας εργασίας. Επίσης ευχαριστώ την κυρία Μαρία Καρανάσιου και την ομάδα DSUNIPI για την πρόσβαση στα αποτελέσματα και στον κώδικα Python που δημιούργησαν.

Περίληψη

Η εποχή που διανύουμε πιθανότατα στο μέλλον θα αναφέρεται ως η εποχή της ηλεκτρονικής επανάστασης. Παραδοσιακά η τεχνολογία προσπαθούσε να δώσει λύσεις στις ανάγκες των ανθρώπων. Στις μέρες μας η ραγδαία πρόοδος που σημειώνει η τεχνολογία και ιδιαίτερα αυτή των ηλεκτρονικών συστημάτων προηγείται αρκετές φορές της δημιουργίας αναγκών. Με την χρήση του διαδικτύου, ιδιαίτερα με την μορφή που έχει αποκτήσει αυτό, μετά τα μέσα της δεκαετίας του 2000 ο χρήστης έχει συμμετοχή και πρόσβαση σε μια τεράστια δεξαμενή πληροφοριών- δεδομένων. Ένα μεγάλο κομμάτι της δεξαμενής αυτής αποτελείται από κείμενα γραπτού λόγου.

Η παρούσα εργασία ασχολείται με την ανάλυση κλίματος ή αλλιώς ανάλυση συναισθήματος σε δεδομένα κειμένου που προέρχονται από την εφαρμογή Twitter. Σαν σκοπό έχει την δημιουργία ενός αυτοτελούς συστήματος ανάλυσης γραπτού λόγου και την κατηγοριοποίηση του σε θετικά, αρνητικά ή ουδέτερα συναισθηματικά φορτισμένα κείμενα. Το σύστημα αυτό δημιουργήθηκε με την χρήση του στατιστικού πακέτου προγραμματισμού R.

Στα πλαίσια της παρούσας εργασίας μελετούμε την αξία της αλληλεπίδρασης των διαφόρων χαρακτηριστικών που έχουν αναδειχθεί από την βιβλιογραφία ως χρήσιμα για την κατηγοριοποίηση ενός κειμένου με βάση τον συναισθηματικό προσανατολισμό του.

Περιεχόμενα

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

Ανάλυση κλίματος (Sentiment Analysis)

Ο ανθρώπινος λόγος, γραπτός και προφορικός ορίζεται ως το σύστημα λέξεων και συμβόλων που χρησιμοποιούν οι άνθρωποι για να εκφράσουν σκέψεις, συναισθήματα ανάγκες, να μεταφέρουν πληροφορία και γενικά να επικοινωνήσουν μεταξύ τους. Για τον εκπαιδευμένο ανθρώπινο εγκέφαλο είναι ένα σύστημα που αν και περίπλοκο μπορεί το να επεξεργαστεί και να το χρησιμοποιήσει εύκολα αξιοποιώντας ένα μεγάλο ποσοστό των δυνατοτήτων που του παρέχει. Το σχετικά εύκολο αυτό, στην χρήση, σύστημα επικοινωνίας για τους ανθρώπους είναι μια ιδιαίτερα δύσκολη δοκιμασία για την τεχνητή νοημοσύνη.

Η επιστήμη έχει ασχοληθεί συχνά με την ανάλυση του ανθρώπινου λόγου χρησιμοποιώντας αυτοματοποιημένες διαδικασίες προκειμένου να επιτύχει διάφορους σκοπούς. Από τις πρώτες προσπάθειες για δημιουργία αυτόματου μεταφραστικού μηχανισμού την δεκαετία του 30 και την δημιουργία του Turing test[1] από τον Alan Turing το 1950 μέχρι τις μεθόδους τεχνητής νοημοσύνης και machine learning στις ημέρες μας.

Ένα σημαντικό μέρος της επικοινωνίας μεταξύ ανθρώπων είναι η χρήση και η αναγνώριση του συναισθήματος σε ένα γραπτό ή προφορικό κείμενο. Το συναίσθημα που διοχετεύεται μέσα από τον λόγο είναι συχνά το πρώτο ερέθισμα που φθάνει στον παραλήπτη της επικοινωνίας και είναι άρρηκτα συνδεδεμένο με το κείμενο. Είναι η ουσία του κειμένου ή η κύρια συνιστώσα του, σε πολλές περιπτώσεις. Τα συναισθήματα των ανθρώπων είναι αφηρημένα και πολυδιάστατα. Ένα κείμενο μπορεί να εκφράζει χαρά, λύπη, θυμό, οργή κλπ. Μπορεί να είναι μίγμα συναισθημάτων που εμφανίζονται ταυτόχρονα ή σε εναλλαγές ή να μην εκφράζει συναίσθημα. Καθώς τα συναισθήματα είναι αφηρημένες έννοιες με συγκεχυμένα όρια μεταξύ τους είναι δύσκολο να ανιχνευθούν από ένα αυτοματοποιημένο σύστημα. Ακόμα και η προσπάθεια να τα ομαδοποιήσουμε σε 2 μεγάλες κατηγορίες θετικών και αρνητικών συναισθημάτων είναι δύσκολη για έναν αλγόριθμο.

Στην παρούσα εργασία θα ασχοληθούμε με την ανάλυση κλίματος κειμένου (sentiment analysis ή opinion mining). Η ανάλυση κλίματος σύμφωνα με τον ορισμό του Wikipedia αναφέρεται στην χρήση μεθόδων natural language processing, ανάλυσης κειμένου και computational linguistics προκειμένου να αναγνωρίσει και να εξάγει πληροφορίες υποκειμενικότητας και συναισθηματικής φόρτισης από δεδομένα γραπτού λόγου.

Ένας βασικός τομέας της SA με τον οποίο θα ασχοληθούμε είναι η κατηγοριοποίηση κειμένων σε θετικά, αρνητικά και ουδέτερα φορτισμένα. Οι δυσκολίες που παρουσιάζονται είναι πολλές και περιλαμβάνουν την ύπαρξη σαρκασμού στο κείμενο, την μεταφορική σημασία των λέξεων, τις διαφορετικές έννοιες των λέξεων, τη χρήση συμβόλων και emoticons, την χρήση αργκό, αρνήσεων κλπ.

Η ανάλυση κλίματος ή SA έχει ευρεία εφαρμογή. Μεταξύ άλλων, εφαρμόζεται σε μεγάλες βάσεις δεδομένων αξιολόγησης προϊόντων/υπηρεσιών και σε δεδομένα εφαρμογών κοινωνικής δικτύωσης. Σαν σκοπό μπορεί να έχει την εξαγωγή συμπερασμάτων για την

[1] Το Turing Test αξιολογούσε την ικανότητα ενός υπολογιστή να προσομοιάσει τον ανθρώπινο γραπτό λόγο

εικόνα μιας εταιρείας ή συγκεκριμένων προϊόντων/υπηρεσιών στην συνείδηση των καταναλωτών καθώς και για την απήχηση συγκεκριμένων πολιτικών ή πρακτικών στο κοινό.

Σαν καταναλωτές προσπαθούμε να συγκεντρώσουμε τις περισσότερες δυνατές πληροφορίες που αφορούν υποψήφια για κατανάλωση προϊόντα-υπηρεσίες. Είτε σκοπεύουμε να αγοράσουμε αυτοκίνητο, είτε σχεδιάζουμε τις διακοπές μας σε κάποιον προορισμό, είτε επιλέγουμε μεταξύ κινηματογραφικών ταινιών, η πληροφόρηση από την εμπειρία άλλων καταναλωτών είναι πολύ σημαντική και συνήθως καθοριστική για την απόφασή μας. Η πληροφόρηση αυτή με την χρήση του διαδικτύου είναι διαθέσιμη άλλα συχνά ο τεράστιος όγκος της, κάνει δύσκολη την διαχείριση και την περίληψη της. Η ανάλυση κλίματος αποτελεί ένα εργαλείο που μπορεί να συμπτύξει τεράστιο όγκο δεδομένων σε πραγματικό χρόνο προκειμένου να δώσει στον καταναλωτή μια συγκεντρωτική εικόνα σε αντίθεση με τον περιορισμένο αριθμό σχολίων-αξιολογήσεων που μπορεί ο ίδιος να διαβάσει.

Η ανάλυση κλίματος είναι χρήσιμη και από την πλευρά των εταιρειών που παρέχουν τα προϊόντα και τις υπηρεσίες στους καταναλωτές και ενδιαφέρονται για την εταιρική τους εικόνα στην αντίληψη του καταναλωτικού κοινού. Πέρα από τα παραδοσιακά μέσα δημοσκοπήσεων, έρευνας αγοράς η ανάλυση κλίματος δίνει την δυνατότητα αποτύπωσης της εταιρικής εικόνας σε μια πολύ μεγαλύτερη βάση ανθρώπων που δεν είναι απαραίτητα καταναλωτές των προϊόντων-υπηρεσιών. Η κλιματική ανάλυση μπορεί να έχει στο επίκεντρό της συγκεκριμένα προϊόντα ή τα αποτελέσματα μιας διαφημιστικής καμπάνιας κλπ.

Σε κάποιες περιπτώσεις όπως στις χρηματιστηριακές αγορές η ψυχολογία του κοινού και η γενική αντίληψη που επικρατεί είναι συχνά ο σημαντικότερος παράγοντας που καθορίζει τις εξελίξεις στον κλάδο. Συνεπώς τα εργαλεία που μπορούν να δώσουν αντίστοιχη πληροφόρηση όπως είναι η ανάλυση κλίματος είναι πολύ χρήσιμα.

Μέσω της ανάλυσης κλίματος μπορούμε να σφυγμομετρήσουμε την θετική ή αρνητική στάση των πολιτών απέναντι σε ζητήματα που απασχολούν μια κοινωνία, απέναντι σε πολιτικές ή άλλες εξελίξεις.

Όλες οι παραπάνω εφαρμογές της ανάλυσης κλίματος αλλά και πολλές άλλες, μπορούν και έχει νόημα να πραγματοποιηθούν σε δεδομένα που προέρχονται από μέσα κοινωνικής δικτύωσης. Ο όγκος δεδομένων είναι παγκόσμιος και τεράστιος, οι χρήστες αυθόρμητα εκφράζονται μέσα από τα μέσα για όλα τα ζητήματα που ανακύπτουν και ήδη έχει δημιουργηθεί ένας παράλληλος ηλεκτρονικός κόσμος που διαρκώς επεκτείνεται μέσα από τα μέσα κοινωνικής δικτύωσης.

Ένα μέσο κοινωνικής δικτύωσης που γνωρίζει ραγδαία ανάπτυξη τα τελευταία χρόνια και αποτελεί πρόσφορο έδαφος για ανάλυση κλίματος ή αλλιώς ανάλυση συναισθηματικού προσανατολισμού είναι το Twitter με το οποίο θα ασχοληθούμε στην παρούσα εργασία.

Στο πρώτο μέρος του δευτέρου κεφαλαίου που ακολουθεί παρουσιάζεται το Twitter και η πορεία του στο χρόνο. Στο δεύτερο κεφάλαιο παρουσιάζονται κάποιες από τις κυριότερες εργασίες στην προσπάθεια των σύγχρονων επιστημών να δημιουργήσουν μηχανισμούς που αναλύουν ικανοποιητικά τον συναισθηματικό προσανατολισμό κειμένου και αντιμετωπίζουν τις δυσκολίες που αναφέραμε παραπάνω. Στο τρίτο κεφάλαιο παρουσιάζεται το σύστημα που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας και μια ανάλυση πιθανοθεωρητικών μοντέλων που χρησιμοποιούνται κατά κόρον. Στο Τέταρτο και τελευταίο κεφάλαιο παρουσιάζονται τα αποτελέσματα της εφαρμογής καθώς και τα συμπεράσματα.

Κ Ε Φ Α Λ Α Ι Ο 2

2.1 Twitter

Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης που παρέχει την δυνατότητα στον χρήστη να στέλνει και να διαβάζει σύντομα μηνύματα 140 χαρακτήρων που ονομάζονται tweets ή να κοινοποιεί μηνύματα άλλων (Retweets). Τα tweets περιέχουν κείμενο, σύμβολα (emojis, emoticons) και φυσικά φωτογραφίες και βίντεο που οι χρήστες μοιράζονται με άλλους χρήστες.

Το Twitter δημιουργήθηκε τον Μάρτιο του 2006 από τους Jack Dorsey, Evan Williams, Biz Stone και Noah Glass και ξεκίνησε να λειτουργεί στο διαδίκτυο τον Ιούλιο του 2006. Τον Απρίλιο του 2007 ενσωματώθηκε σε ξεχωριστή εταιρεία την Twitter Inc. με έδρα το Σαν Φρανσίσκο της Καλιφόρνια. Πολύ γρήγορα κερδίζει δημοτικότητα και εδραιώνεται ως το sms (Short Message Service) του διαδικτύου.

Το 2009 αεροπλάνο των αμερικανικών αερογραμμών πραγματοποιεί αναγκαστική προσγείωση στον ποταμό Hudson της Νέας Υόρκης. Η φωτογραφία του δημοσιεύεται στο twitter μαζί με την είδηση προλαβαίνει τα διεθνή ειδησεογραφικά μέσα στην προβολή.

Φωτογραφία δημοσιευμένη στο twitter 15.1.2009

Το Twitter εδραιώνεται ταχύτατα και ως μέσο άμεσης πληροφόρησης. Τα ειδησεογραφικά πρακτορεία ενημερώνουν διαρκώς το κοινό με tweets. Τον Μάρτιο του 2011 η εταιρεία ανακοινώνει μεταξύ άλλων την αποστολή 1 δισεκατομμυρίου tweets την εβδομάδα και τον Σεπτέμβριο του ίδιου έτους ότι οι ενεργοί χρήστες της εφαρμογής έχουν φτάσει τα 100 εκατομμύρια μηνιαίως. Το 2012 ο Νέος Πρόεδρος των Ηνωμένων Πολιτειών ανακοινώνει την νίκη του στις εκλογές στον λογαριασμό του στο twitter. Αύγουστο του 2013 η εταιρεία ανακοινώνει ότι 500 εκατομμύρια tweets στέλνονται ημερησίως. Τα νούμερα του twitter αυξάνονταν διαρκώς μέχρι το τελευταίο 3μηνο του 2014. Από το 2014 και μετά τα νούμερα του Twitter αν και παραμένουν υψηλά φθίνουν.

Το Twitter μπορούμε να πούμε ότι προσομοιάζει στην ανθρώπινη επικοινωνία καθώς το φάσμα των περιεχομένων των tweets είναι ευρύ. Περιλαμβάνει από ειδησεογραφία και πολιτικό διάλογο υψηλόβαθμων αξιωματούχων μέχρι διαφημίσεις, αυτοπροβολή χρηστών, ανούσιες ενημερώσεις, χιούμορ, spam κλπ. Σύμφωνα με την εταιρεία έρευνας αγοράς Pear Analytics που ανέλυσε 2000 tweets για περίοδο 2 εβδομάδων τον Αύγουστο του 2009 κατέληξε ότι το 40% των tweets ήταν άνευ ουσίας (Pointless Bubble), το 38% συζητήσεις, 9% είχαν κάποια αξία ανάλυσης, 6% ήταν tweets αυτοπροβολής, 4% spam και 4% ειδήσεις.

Η εταιρεία (twitter Inc) δεν συμφώνησε με τον χαρακτηρισμό Pointless Bubble θεωρώντας πως υπάρχει ουσία στην κοινωνική δικτύωση και στην αλληλεπίδραση των χρηστών.

Το Twitter είναι σίγουρα μια τεράστια δεξαμενή πληροφοριών από την οποία μπορούμε να αντλήσουμε γνώση για ζητήματα που απασχολούν την κοινή γνώμη μιας χώρας ή του πλανήτη, για την εικόνα μιας μεγάλης πολυεθνικής εταιρείας στα μάτια των καταναλωτών, για το οικονομικό κλίμα, την απήχηση διαφημιστικής καμπάνιας κλπ. Σύμφωνα με την εργασία των Federica Giummolè, Salvatore Orlando, Gabriele Tolomei: «**Trending Topics on Twitter Improve the Prediction of Google Hot Queries**» **[1]** χρησιμοποιώντας ανάλυση χρονολογικών σειρών (autoregressive distributed lag models) ένα μεγάλο ποσοστό από τα δημοφιλή θέματα που αναδεικνύονται μέσω Twitter αναδεικνύονται στην συνέχεια και από τις αναζητήσεις της Google σε μια εφαρμογή με μεγαλύτερη βαρύτητα και απήχηση.

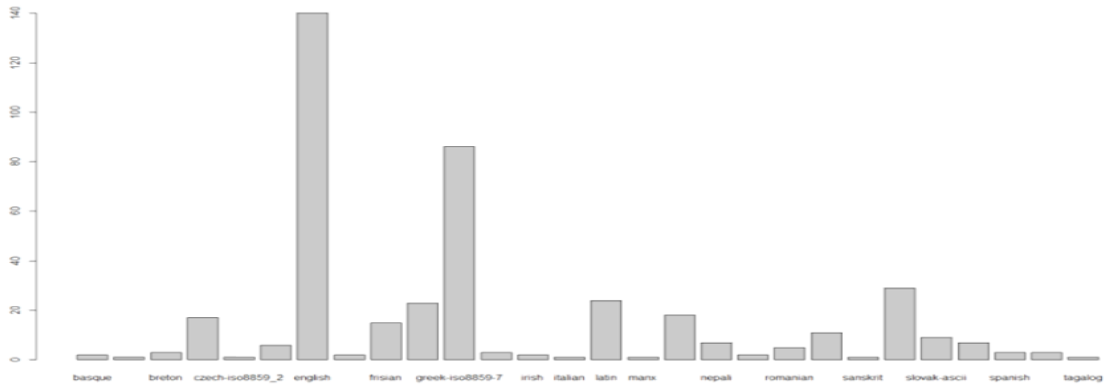
Η πολιτική της εταιρείας ενθαρρύνει τη χρήση των δεδομένων που δημοσιεύονται στην εφαρμογή της για σκοπούς ανάλυσης και επεξεργασίας με κάποιους σημαντικούς περιορισμούς. Υπάρχουν πολλές εφαρμογές που αξιοποιούν τα δεδομένα του Twitter και άλλων μέσων κοινωνικής δικτύωσης.

Twitter API

Η εφαρμογή Twitter δίνει την δυνατότητα στους χρήστες να έχουν πρόσβαση σε περιορισμένο αριθμό αναρτήσεων (tweets) μέσω της πλατφόρμας για επεξεργασία ή ανάλυση. Παρέχει επίσης τη δυνατότητα να επιλέξουν τη γλώσσα που θα είναι δημοσιευμένα, λέξεις κλειδιά ή/και άλλα περιεχόμενα που επιθυμούν να περιέχουν τα tweets καθώς επίσης να έχουν δημοσιευτεί από συγκεκριμένο χρήστη ή/και σε συγκεκριμένες συντεταγμένες ή/και χρόνο.

Η βιβλιοθήκη **twitter** χρησιμοποιείται μεταξύ άλλων, προκειμένου να συνδεθούμε με το twitter και να εισάγουμε στην R tweets με την εντολή ***searchTwitter*.**
Με τις επιλογές αυτές δίνεται η δυνατότητα να εστιάσουμε σε συγκεκριμένα χωροχρονικά πλαίσια και να συλλέξουμε δείγμα tweets που αναφέρεται σε συγκεκριμένο αντικείμενο ή από συγκριμένους χρήστες.

Για παράδειγμα ζητήσαμε από την εφαρμογή 500 tweets που δημοσιεύτηκαν σε ακτίνα 20 μιλίων από την πλατεία Ομονοίας χωρίς άλλο περιορισμό. Αφαιρώντας τα Retweets από το δείγμα έμειναν 428 tweets. Στο ιστόγραμμα που ακολουθεί παρουσιάζονται τα αποτελέσματα του αλγόριθμου της βιβλιοθήκης **textcat** για την αναγνώριση της γλώσσας (δεν υπάρχει η δυνατότητα αναγνώρισης των διαφόρων ειδών greeklish).
Η συγκεκριμένη βιβλιοθήκη εφαρμόζει και βελτιστοποιεί την κατηγοριοποίηση των tweets χρησιμοποιώντας τον αλγόριθμο των William, B. Cavnar και John M. Trenkle, ο οποίος βασίζεται στην χρήση των N grams που προκύπτουν από τις λέξεις των κειμένων. Ο συγκεκριμένος αλγόριθμος επιτυγχάνει πολύ υψηλά ποσοστά όταν πρόκειται για την αναγνώριση της γλώσσας σε κανονικό κείμενο. Στα tweets όμως λόγω της ιδιομορφίας τους δεν είναι εξίσου επιτυχημένη η κατηγοριοποίηση.



Π1.Ιστόγραμμα συχνότητας διαφόρων γλωσσών που ανιχνεύθηκαν

2.2 Ιστορική Αναδρομή Ανάλυσης Κλίματος

Σύμφωνα με τον ορισμό Web 2.0 που έγινε παγκοσμίως γνωστός από τον Tim O' Riell, το διαδίκτυο από το 2004 και μετά έχει περάσει επίσημα στην 2η φάση του. Αυτή χαρακτηρίζεται από την ενεργή συμμετοχή των χρηστών του στην δημιουργία των τεράστιων διαθέσιμων όγκων δεδομένων. Σε αντιδιαστολή με την πρώτη περίοδο κατά την οποία ο μεγάλος όγκος χρηστών ήταν κυρίως παθητικός καταναλωτής πληροφοριών, χωρίς ο ίδιος να δημιουργεί δεδομένα, η δεύτερη φάση του διαδικτύου χαρακτηρίζεται από την χρήση μέσων κοινωνικής δικτύωσης, ελεύθερη διακίνηση ηλεκτρονικών αρχείων και γενικά την έντονη και ενεργή δραστηριότητα και αλληλεπίδραση των χρηστών του Internet με την δημιουργία, δημοσιοποίηση και ανταλλαγή δεδομένων.

Η έρευνα για την δημιουργία μηχανισμών που αναλύουν την συναισθηματική φόρτιση γραπτού λόγου είχε αναπτυχθεί αρκετά χρόνια πριν την ανάπτυξη του Internet. Η εργασία όμως των **Vasileios Hatzivassiloglou και Kathleen R. McKeown το 1997** « **Predicting the Semantic Orientation of Adjectives**» μπορεί να θεωρηθεί ως ένα σημείο αναφοράς. Σε αυτή διαχωρίζουν τα επίθετα σε θετικού και αρνητικού συναισθηματικού προσανατολισμού, χρησιμοποιώντας ένα μοντέλο λογαριθμογραμμικής παλινδρόμησης. Πιο συγκεκριμένα εξάγουν 15.048 συνδέσεις-ζευγάρια επιθέτων (9.296 διαφορετικά ζευγάρια) που ενώνονται με λέξεις συνδέσμους όπως είναι το "and", "or", "but", "either-or","neither-nor" από ένα σώμα κειμένων της Wall Street Journal με 21 εκατομμύρια λέξεις. Χρησιμοποιώντας ένα μοντέλο λογαριθμογραμμικής παλινδρόμησης που περιέχει σαν ερμηνευτικές μεταβλητές, παράγοντες όπως είναι η λέξη σύνδεσμος και ή μορφολογική τροποποίηση που ασκεί στο ουσιαστικό της πρότασης η σύνδεση (προσδιοριστική, κατηγορηματική, παραθετική κλπ.) αναθέτουν σε κάθε ζευγάρι επιθέτων μια εκτίμηση y (σκορ) μεταξύ (0,1) για την πιθανότητα να ανήκουν στον ίδιο προσανατολισμό (είτε θετικό είτε αρνητικό) με ποσοστό επιτυχίας 82% . Τέλος το σκορ μετατρέπεται σε διαφορά προσανατολισμού φόρτισης (ή απόσταση) μεταξύ των επιθέτων (1-y) και με την βοήθεια μη ιεραρχικού αλγόριθμου συσταδοποίησης τα αποτελέσματα ομαδοποιούνται σε 2 ομάδες θετικά και αρνητικά φορτισμένων επιθέτων.

Το **2002** Ο **Peter D. Turney** στην εργασία του: «**Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews»** εστιάζει την ανάλυσή του σε επίπεδο φράσης δύο λέξεων και σε δεδομένα αξιολόγησης καταναλωτών διαφόρων προϊόντων και υπηρεσιών από την ιστοσελίδα EPINION όπως είναι: η αγορά

αυτοκινήτων, οι τράπεζες, οι ταινίες και οι ταξιδιωτικοί προορισμοί. Χρησιμοποιεί φράσεις 2 λέξεων (Bigrams) που περιέχουν επίθετα και επιρρήματα ώστε να συμπυκνώσει καλύτερα το νόημα του κειμένου. Όπως αναφέρει για παράδειγμα το επίθετο απρόβλεπτη (unpredictable) μπορεί να έχει θετική φόρτιση όταν αφορά αξιολόγηση κινηματογραφικής ταινίας (unpredictable plot) αλλά αρνητική όταν πρόκειται για αξιολόγηση επιδόσεων αυτοκινήτου (unpredictable steering). Ο αλγόριθμος που χρησιμοποιεί για την εκτίμηση του κλίματος κάθε φράσης είναι ο PMI-IR ή Point wise Mutual Information and Information Retrieval. Για κάθε φράση κατέγραψε το πλήθος των αποτελεσμάτων 3 αναζητήσεων στο διαδίκτυο μέσω της Near επιλογής που παρείχε η μηχανή αναζήτησης ALTAVISTA. 1. Το πλήθος των αποτελεσμάτων που περιέχουν την φράση, 2. Το πλήθος των αποτελεσμάτων που περιέχουν την φράση μαζί με την λέξη «Excellent» και 3. Το πλήθος των αποτελεσμάτων που περιέχουν την φράση μαζί με την λέξη «Poor». Επίσης κατέγραψε το πλήθος των αποτελεσμάτων των αναζητήσεων «Excellent» και «Poor» χωριστά. Χρησιμοποίησε τον παρακάτω τύπο ο οποίος αποτελεί μέτρο του βαθμού στατιστικής εξάρτισης μεταξύ δύο λέξεων:

$$PMI(word_1, word_2) = \log_2\left[\frac{p(word_1 \& word_2)}{p(word_1)\,p(word_2)}\right]$$

Για να διευκρινίσει τον συναισθηματικό προσανατολισμό της φράσης υπολόγισε την διαφορά:

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$

ή αλλιώς

$$\log_2\left[\frac{hits(phrase\,NEAR\,"excellent")(hits("poor")}{hits(phrase\,NEAR\,"poor")(hits("excellent")}\right]$$

που αποτελεί Log-Odds Ratio. Ο αλγόριθμος κατηγοριοποίησε σωστά το 74.39% των αξιολογήσεων επιτυγχάνοντας καλύτερα αποτελέσματα στις αξιολογήσεις αυτοκινήτων με 84% και στις τράπεζες με 80%, ενώ στις αξιολογήσεις των ταξιδιωτικών προορισμών πέτυχε 70.53% Το χειρότερο ποσοστό κατέγραψε στις αξιολογήσεις κινηματογραφικών ταινιών όπου πέτυχε μόλις 65.83%.

Την ίδια χρονιά σε δεδομένα αξιολόγησης κινηματογραφικών ταινιών, οι **Bo Pang, Lillian Lee και Shivakumar Vaithyanathan** με την εργασία τους «**Thumbs up? Sentiment Classification using Machine Learning Techniques»** εφαρμόζουν μεθόδους machine learning προκειμένου να επιτύχουν καλύτερα αποτελέσματα. Μάλιστα, έδειξαν ότι οι λέξεις που θεωρούμε ότι αντιπροσωπεύουν θετικές ή αρνητικές κριτικές διαφέρουν στην πράξη. Λαμβάνοντας υπόψη την συχνότητα λέξεων ή συμβόλων στίξης στα θετικά και αρνητικά σχόλια-αξιολογήσεις κατέληξαν στην παρακάτω λίστα με στοιχεία που δεν θα σκεφτόταν εύκολα ένας άνθρωπος.

	Proposed word lists	Accuracy	Ties
Human 3 + stats	**Positive**: **love**, **wonderful**, **best**, **great**, **superb**, **still**, **beautiful** **Negative**: **bad**, **worst**, **stupid**, **waste**, **boring**, **?**, **!**	69%	16%

Π2. Λέξεις-σύμβολα που συνδέονται με πολικότητα του κειμένου

Τα δεδομένα αποτελούνταν από 1.301 θετικά σχόλια-αξιολογήσεις και 752 αρνητικά. Συμβολίζοντας το κείμενο (document) με d, την κλάση (class) που επιθυμούμε να χωρίσουμε τα δεδομένα με c, με f_i τα χαρακτηριστικά επάνω στα οποία εκπαιδεύονται οι αλγόριθμοι και διαμερίζουν το κείμενο d. Οι μέθοδοι που εφάρμοσαν στα δεδομένα είναι:

- **Naive Bayes**
  Αν και οι υποθέσεις ανεξαρτησίας μεταξύ των f_i δεν ευσταθούν στην πραγματικότητα, τα αποτελέσματά του είναι συνήθως ικανοποιητικά και προκύπτουν από τον τύπο:

$$P_{NB}(c/d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i/c)^{n_i(d_i)})}{P(d)}$$

- **Maximum Entropy**
  Ένας δεύτερος αλγόριθμος που δοκίμασαν και συνήθως δίνει καλύτερα αποτελέσματα από τον Naïve Bayes ιδιαίτερα όταν η ανεξαρτησία μεταξύ των f_i είναι παραπλανητική. Ο εκθετικός τύπος που χρησιμοποίησαν είναι:

$$P_{ME}(c/d) := \frac{1}{Z(d)} \exp(\sum_i \lambda_{i,c} F_{i,c}(d,c)$$

$$F_{i,c}(d,c') := \begin{cases} 1, n_i(d) > 0 and c' = c \\ 0, otherwise \end{cases}$$

- **Support Vector Machines**
  Τέλος χρησιμοποίησαν τον SVM αλγόριθμο που δεν χρησιμοποιεί πιθανότητες αλλά κατηγοριοποιεί σε 2 ομάδες τα δεδομένα προσπαθώντας να καταλήξει στην καλύτερη δυνατή διαχώριση των δεδομένων σε διάφορα υπερεπίπεδα που αντιπροσωπεύει το διάνυσμα w με την μεγαλύτερη δυνατή απόσταση μεταξύ των 2 ομάδων. Συνίσταται σε πρόβλημα βελτιστοποίησης με περιορισμούς. Έστω ότι τα c_j Ε {-1,1} αντιστοιχούν στις 2 σωστές κατηγορίες του κειμένου d_j η λύση μπορεί να γραφεί ως:

$$\vec{w} := \sum_j a_j c_j \vec{d}_j, a_j \geq 0$$

Τα αποτελέσματα έδειξαν ότι και οι τρείς μέθοδοι κατηγοριοποιούν καλύτερα τα σχόλια-αξιολογήσεις σε σχέση με τις Unsupervised προσπάθειες. Μάλιστα ο αλγόριθμος SVM πέτυχε την μεγαλύτερη ακρίβεια σχεδόν σε όλες τις δοκιμές με διαφορετικά χαρακτηριστικά για εκμάθηση. Στον πίνακα που ακολουθεί παρουσιάζουν αναλυτικά τα αποτελέσματα και τα συγκρίνουν με διαφορετικές μεθόδους όπως είναι των **Vasileios Hatzivassiloglou** και **Kathleen R. McKeown.**

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Figure 3: Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%.

Π3. Παρουσίαση αποτελεσμάτων της εργασίας των Bo Pang, Lilian Lee

Ενδιαφέρον επίσης παρουσιάζει το ότι η χρήση Unigrams +Bigrams δεν βελτιώνει την αποτελεσματικότητα των αλγορίθμων όπως θα περιμέναμε. Επίσης αρκούν τα πιο συχνά Unigrams (2633 αντί 16165) για δώσουν ικανοποιητικά αποτελέσματα. Η συχνότητα που εμφανίζονται τα Unigrams δεν βελτιώνει τα ποσοστά που επιτυγχάνει η απλή παρουσία τους και μόνο στο σχόλιο-αξιολόγηση. Στην προετοιμασία των δεδομένων χρησιμοποιήθηκε η μέθοδος των **Das και Chen (2001)** σύμφωνα με την οποία προσθέτουν το NOT_ σε κάθε λέξη που εκφράζει άρνηση (“not”,”isn ‘t”, “didn ‘t” κλπ) για να βοηθήσουν τον αλγόριθμο.

Την ανωτερότητα των Supervised μεθόδων επιβεβαιώνουν με την εργασία τους «**Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches**» (2005) οι **Pimwadee Chaovalit** και **Lina Zhou**.

Το 2005 οι **Bo Pang** και **Lillian Lee** παρουσιάζουν τα οφέλη της τροφοδότησης των παραδοσιακών μεθόδων Machine Learning που χρησιμοποίησαν το 2002 με δεδομένα που έχουν πρώτα φιλτραριστεί σε σχέση με την υποκειμενικότητά τους. Προτάσεις που δεν μπορούν να χαρακτηριστούν υποκειμενικές συνήθως δεν είναι συναισθηματικά φορτισμένες και δεν περιγράφουν την θέση του ατόμου απέναντι στο θέμα. Για παράδειγμα, όσον αφορά τα σχόλια-αξιολογήσεις ταινιών η περίληψη της πλοκής που μπορεί να γίνεται εντός του σχολίου αποτελεί θόρυβο. Για να πετύχουν τον διαχωρισμό των δεδομένων στην εργασία τους «**A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts**» δεν χρησιμοποιούν τους αλγόριθμους που είχαν χρησιμοποιήσει για την ανάλυση κλίματος. Δημιουργούν ένα νοητό σχήμα που συνδέει όλα τα δεδομένα (φράσεις-λέξεις) μεταξύ τους αλλά και με τις 2 διαφορετικές ομάδες: των υποκειμενικών προτάσεων και των αντικειμενικών. Λαμβάνοντας υπόψη, ως αποστάσεις, τα ατομικά σκορ της κάθε εγγραφής που τα κατατάσσουν στις 2 ομάδες αλλά και την σχέση που συνδέει την μια εγγραφή με την άλλη (association) ο αλγόριθμος διαχωρίζει τα δεδομένα βρίσκοντας την βέλτιστη λύση που ελαχιστοποιεί το κόστος της κατηγοριοποίησης. Τα οφέλη για την ανάλυση κλίματος είναι: ότι τα κείμενα συμπυκνώνονται και καταφέρνουν να έχουν περίπου την ίδια απόδοση με λιγότερα δεδομένα για ανάλυση αλλά και μεγαλύτερα ποσοστά επιτυχίας για τον αλγόριθμο Naïve Bayes από την χρήση των φιλτραρισμένων δεδομένων.

Την ίδια χρονιά με την εργασία τους «**Extracting Product Features and Opinions from Reviews**» οι **Ana-Maria Popescu** και **Oren Etzioni** εστιάζουν σε μια άλλη πτυχή της ανάλυσης κλίματος και αυτή είναι η ανίχνευση των επιμέρους χαρακτηριστικών που αφορά η κριτική. Δηλαδή όχι μόνο αν είναι αρνητική ή θετική η αξιολόγηση του Laptop αλλά και αν

το σχόλιο αφορά την διάρκεια της μπαταρίας της συσκευής, το design, την τιμή κλπ. Αναλύουν το πρόβλημα ως εξής:

I. Αναγνώριση των χαρακτηριστικών του προϊόντος
II. Ανίχνευση σχολίων που αφορούν τα παραπάνω χαρακτηριστικά.
III. Κατηγοριοποίηση με βάση τον συναισθηματικό προσανατολισμό (ανάλυση κλίματος)
IV. Ταξινόμηση των σχολίων με βάση την έντασή τους.

I. Για την αναγνώριση χαρακτηριστικών χρησιμοποιούν τον αλγόριθμο Point wise Mutual Information για να αποδώσουν σκορ σε ουσιαστικά που εμφανίζονται με υψηλή συχνότητα στις κριτικές και συνδέονται με το προϊόν και λέξεις σύνδεσης στο διαδίκτυο.

II. Για την ανίχνευση σχολίων που συνδέονται με τα παραπάνω χαρακτηριστικά βασίζονται σε 10 κανόνες εξαγωγής των χαρακτηριστικών μαζί με συγκεκριμένα μέρη του λόγου.

III. Για την ανάλυση κλίματος χρησιμοποιούν unsupervised αλγόριθμο που ονομάζεται Relaxation Labelling και αποδίδει συναισθηματικό προσανατολισμό στην λέξη-φράση χρησιμοποιώντας πιθανότητες που λαμβάνουν υπόψη τις γειτονικές σε αυτή λέξεις-φράσεις και τον προσανατολισμό τους σύμφωνα με κάποιους κανόνες.

Οι παραπάνω λειτουργίες αποτελούν το σύστημα που ονομάζουν OPINE και η αποτελεσματικότητά του συγκρίνεται με την μέθοδο **του Turney** και **των Hu and B. Liu. 2004** που έχουν παρεμφερή προσέγγιση. Τα αποτελέσματα είναι ανώτερα.

Μια ειδική περίπτωση του Naïve Bayes που παρουσιάστηκε παραπάνω είναι ο αλγόριθμος **multinomial Naïve Bayes (MNB)** που αντιμετωπίζει τα χαρακτηριστικά f (μεμονωμένες λέξεις και σημεία στίξης, ή φράσεις 2 λέξεων κλπ), σαν παρατηρήσεις που προέρχονται από την Multinomial κατανομή (εξ ου και multinomial Naïve Bayes).

$$P(f_i / c)^{n_i(d_i)} \square Mn(n, p_1, p_2, ..., p_n)$$

Σε αντίθεση με την Κατανομή Bernoulli που χρησιμοποιήθηκε παραπάνω η πολυωνυμική κατανομή δίνει την δυνατότητα στον αλγόριθμο κατηγοριοποίησης να λάβει υπόψη του και την συχνότητα των χαρακτηριστικών f του κειμένου πέρα από την παρουσία τους. Η χρήση της συγκεκριμένης παραλλαγής που είχε ήδη γνωρίσει επιτυχία στην ταξινόμηση κειμένων σε θεματικές κατηγορίες και στην ανίχνευση spam emails χρησιμοποιήθηκε και στην ανάλυση κλίματος γραπτού λόγου. Η Alyssa Liang το 2006 με την εργασία της «**Rotten Tomatoes: Sentiment Classification in Movie Reviews**» σε δεδομένα αξιολόγησης κινηματογραφικών ταινιών έδειξε πώς η χρήση του συγκεκριμένου αλγόριθμου σε δεδομένα που δεν έχουν δεχθεί ιδιαίτερη τροποποίηση, παρά μόνο μεγαλύτερο βάρος στο τελευταίο 20% του κειμένου, όπου συνήθως η κριτική γίνεται πιο έντονη και συμπυκνωμένη, δίνει καλύτερα αποτελέσματα από τον SVM (2 παραλλαγές του: linear και radial basis function kernel). Ενδιαφέρον επίσης παρουσιάζει το γεγονός ότι ο SVM δίνει καλύτερα αποτελέσματα όταν τον τροφοδοτεί με επιπλέον πληροφορία για την συχνότητα-σημασία των χαρακτηριστικών σε σχέση με τα υπόλοιπα σχόλια-αξιολογήσεις. Τα αποτελέσματα αυτά είναι σε αντίθεση με τα αποτελέσματα των **Bo Pang, Lillian Lee και Shivakumar Vaithyanathan** σχετικά με την σημασία της παρουσίας και όχι της συχνότητας των χαρακτηριστικών.

Σε καλύτερη κατηγοριοποίηση κλίματος κατέληξαν και οι **Neil O'Hare1 , Michael Davy2 , Adam Bermingham1 , Paul Ferguson1 , Páraic Sheridan2 , Cathal Gurrin1 , Alan F. Smeaton** στην εργασία τους «**Topic-Dependent Sentiment Analysis of Financial Blogs**» το 2009 χρησιμοποιώντας τον αλγόριθμο MNB σε σύγκριση με τον SVM (linear Kernel). Τα αποτελέσματα τους, παρουσίασαν στον παρακάτω πίνακα σε επίπεδο λέξης, πρότασης και παραγράφου.

	Paragraph		Sentence		Words		
N	SVM	MNB	SVM	MNB	N	SVM	MNB
0	67.9462	**73.3429**	69.2377	71.8958	5	68.9565	71.2904
1	64.5829†	71.8679301	**70.7022**	**72.5925**	10	72.6119	72.1599
2	66.3369	70.99617	68.0999	72.1534	15	72.1901	73.6156
3	66.9230	70.855509	70.5656	72.5839	20	73.3301	74.6280
4	67.0724	69.83894466	67.9247	71.7143	25	**74.3683**	74.0460
5	67.7949	69.09919†	66.4883	70.5636	30	71.8807	**75.0691**
6	**68.2383**	69.38905†	66.7825	71.4331	40	72.1728	74.4787
7	64.8618†	69.8217	64.8791†	70.4143	50	71.5779	74.0482
8	63.4127†	69.96663	67.7925	69.8367	60	68.3722	74.3511
9	65.1604†	69.966631	66.1920	70.2758	70	68.3301	73.3190
10	64.8749†	70.260748	65.4586†	69.6789	80	69.0925	73.1722
Baseline	66.0601	69.5447	66.0601	69.5447		66.0601	69.5447

Table 5: Binary classification results for paragraph, sentence and word text extraction. Maximum accuracy is represented by bold text, while accuracy below that of the baseline are indicated with †

Π4. Αποτελέσματα εργασίας Neil O'Hare1 , Michael Davy2 , Adam Bermingham1

Πρέπει να σημειώσουμε, ότι τα αποτελέσματα αφορούν κατηγοριοποίηση σε αρνητικά και θετικά φορτισμένα κείμενα και δεν διαχωρίζουν σε ουδέτερα. Τα αποτελέσματα των αλγόριθμων αν λάβουμε υπόψη μας και τα ουδέτερα κείμενα είναι αρκετά χαμηλότερα.

Μια διαφορετική προσέγγιση του θέματος έγινε από τους Edoardo Airoldi, Xue Bai, και Rema Padman με την εργασία τους «**Markov Blankets and Meta-Heuristics Search: Sentiment Extraction from Unstructured Texts»**. Ασχολήθηκαν με την κατηγοριοποίηση διαδικτυακών αξιολογήσεων κινηματογραφικών ταινιών σε 2 κατηγορίες (θετικές/αρνητικές κριτικές) και οικονομικών ειδησεογραφικών δεδομένων (από 3 δεξαμενές ύλης: M&A, Finance και Mixed news) σε 3 κατηγορίες (θετικά, αρνητικά και ουδέτερα άρθρα). Η μέθοδος που χρησιμοποίησαν χωρίζεται σε 2 βασικά στάδια.

1ο Στάδιο Markov Blanket

Το πρώτο στάδιο χρησιμοποιεί κατηγοριοποίηση Markov Blanket σε μπευσιανό δίκτυο (1) για να δημιουργήσει το μικρότερο δυνατό υποσύνολο λέξεων που εκφράζουν με τον καλύτερο δυνατό τρόπο τον συναισθηματικό προσανατολισμό του σχολίου-κειμένου και (2) για να καταγράψει την εξάρτηση των λέξεων μεταξύ τους και των λέξεων με την μεταβλητή κατηγοριοποίησης. Για την καταγραφή της εξάρτησης (ή γραφικά απόστασης) χρησιμοποιεί το G^2 test ανεξαρτησίας.

Λαμβάνοντας υπόψη την ύπαρξη ή μη της κάθε λέξης {0,1} αλλά και την στατιστική σημαντικότητα του G^2 test για την ανεξαρτησία μεταξύ των λέξεων ο αλγόριθμος καταφέρνει να αμβλύνει την επίδραση των διαφορετικών θεμάτων στην κατηγοριοποίηση και να εστιάσει σε λέξεις που συνδέονται στενά με τον συναισθηματικό προσανατολισμό. Σε πρώτη φάση ο αλγόριθμος συνδέει την μεταβλητή που απεικονίζει την κατηγοριοποίηση με λέξεις και σε δεύτερη φάση τις λέξεις αυτές με άλλες λέξεις. Στο Τρίτο βήμα ο αλγόριθμος αποδίδει στην σύνδεση σχέση «αιτίας-αιτιατού» και τέλος στο τέταρτο βήμα κρατάει μόνο τις συνδέσεις και

τις λέξεις που έχουν προκριθεί. Στο γράφημα που ακολουθεί από την εργασία τους παρουσιάζονται στα αριστερά ένα δίκτυο Bayes και στα δεξιά ένα Markov Blanket.

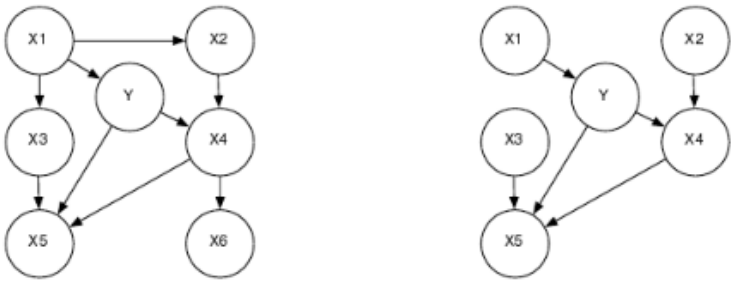


Fig. 1. (left) A sample Bayesian Network (S, P), and (right) the Markov Blanket for the variable encoding the overall sentiment of a document, Y.

Π5. Bayes δίκτυο και δίκτυο Markov Blanket

2ο Στάδιο Tabu Search Heuristic

Στο Δεύτερο Στάδιο ο αλγόριθμος βελτιστοποιεί το αποτέλεσμα του προηγούμενου σταδίου δοκιμάζοντας τροποποιήσεις που να μεγιστοποιούν την ικανότητα κατηγοριοποίησης όπως φαίνεται στο γράφημα που ακολουθεί. Για κάθε νέο κείμενο $\{x_1,\ldots,x_V\}$ υπολογίζει την ποσότητα:

$$l_i = \log\left[\frac{P(Y - y_i / \{x_1,...,x_V\})}{P(Y - y_0 / \{x_1,...,x_V\})}\right]$$

Όπου Y_i είναι η πιθανή τιμή της μεταβλητής κατηγοριοποίησης και επιλέγουμε την κατηγοριοποίηση που μεγιστοποιεί αυτή την ποσότητα.

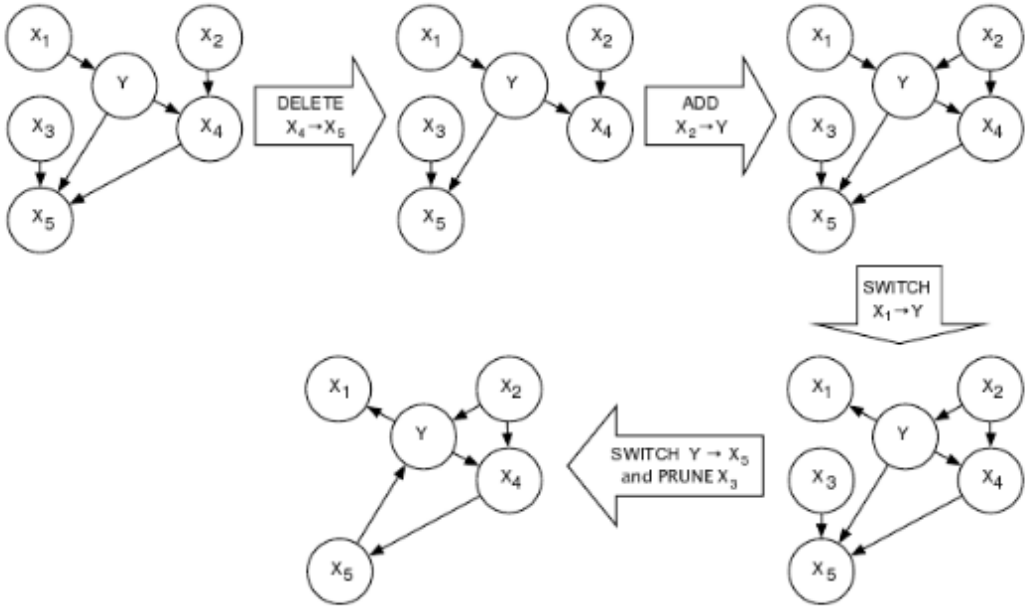


Fig. 4. An example of the moves allowed in Tabu Search

. Π6.Διαδικασία βελτιστοποίησης δίκτυο Markov Blanket

Για την τελική κατηγοριοποίηση των κειμένων-σχολίων ο αλγόριθμος χρησιμοποιεί απλή λογιστική παλινδρόμηση. Για τα δεδομένα κριτικών ταινιών η συγκεκριμένη μέθοδος συγκρίνεται μεταξύ άλλων με τις μεθόδους SVM και NB και παρουσιάζει καλύτερα αποτελέσματα σε όρους AUC: Area under the Curve (87,52% έναντι 81,32% και 82,61 αντίστοιχα) χρησιμοποιώντας μόνο 32 λέξεις από τις 7.716. Μάλιστα οι 32 λέξεις που προκρίνονται από την μέθοδο εκπαιδεύουν καλύτερα τον αλγόριθμο σε σχέση με 32 λέξεις που θα επιλέγαμε σύμφωνα με το κέρδος πληροφορίας. Αντίστοιχα είναι τα αποτελέσματα και για το δεύτερο σετ δεδομένων (χωρίς να συγκρίνονται με τον SVM αλγόριθμο).

Με την ραγδαία ανάπτυξη των μέσων κοινωνικής δικτύωσης και την εδραίωση της Δεύτερης φάσης του διαδικτύου, η ανάλυση κλίματος βρίσκεται στο επίκεντρο της έρευνας μεγάλου όγκου δεδομένων.

Το 2009 οι Alec Go, Richa Bhayani και Lei Huang στην εργασία τους «**Twitter Sentiment Classification using Distant Supervision**» εφαρμόζουν τις κλασικές μεθόδους machine learning (Multinomial Naïve Bayes, Maximum Entropy, SVM) για πρώτη φορά σε δεδομένα twitter. Για να εκπαιδεύσουν τους αλγόριθμους χρησιμοποιούν tweets τα οποία διαχωρίζονται σε θετικά και αρνητικά με μοναδικό κριτήριο το αν περιέχουν θετικά ή αρνητικά emoticons όπως αυτά που φαίνονται στον παρακάτω πίνακα.

Table 3: List of Emoticons

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Π7. Λίστα δημοφιλών Emoticons

Με την συγκεκριμένη μέθοδο που χρησιμοποιήθηκε πρώτη φορά από τον J. Read στην εργασία «**Using emoticons to reduce dependency in machine learning techniques for sentiment classification**» καταφέρνουν να δημιουργήσουν αυτόματα και άμεσα μια μεγάλη βάση θετικών και αρνητικών tweets (όχι όμως και ουδέτερων) προκειμένου να εκπαιδεύσουν τους τρείς κλασικούς αλγόριθμους. Το βασικό πλεονέκτημα της μεθόδου είναι ότι δεν χρειάζεται να αναθέσουν σε ανθρώπους την χρονοβόρα διαδικασία της αντιστοίχισης θετικής ή αρνητικής φόρτισης του κάθε σχολίου. Το βασικό μειονέκτημα της μεθόδου είναι ότι τα δεδομένα που χρησιμοποιούν κατηγοριοποιούνται με βάση μόνο τα emoticons που στην πραγματικότητα μπορεί να είναι αντίθετα φορτισμένα σε σχέση με το κείμενο, να εκφράζουν σαρκασμό κλπ. Επίσης εκπαιδεύουν τον αλγόριθμο με συγκεκριμένη υποκατηγορία μηνυμάτων (αυτών που περιέχουν emoticons).

Μετά τους απαραίτητους μετασχηματισμούς και αφού έχουν εκπαιδεύσει τους αλγόριθμους, χρησιμοποιώντας διάφορα χαρακτηριστικά των 1,600,000 tweets (50% θετικά και 50% αρνητικά) για εκμάθηση, εφαρμόζονται οι αλγόριθμοι σε 177 αρνητικά και 182 θετικά tweets. Τα αποτελέσματα είναι πολύ κοντά στα αποτελέσματα των **Bo Pang, Lillian Lee και Shivakumar Vaithyanathan** (MNB 81,3%, ME 80,5% και SVM 82,2% αντίστοιχα). Όσον αφορά τα χαρακτηριστικά, κατέληξαν ότι η χρήση Unigrams+Bigrams σε σχέση με την χρήση μόνο Unigrams βελτιώνει την απόδοση του Naïve Bayes (από 81.3% σε 82.7%) και του Maximum Entropy (από 80.5 σε 82.7) αλλά όχι του SVM (από 82.2% σε 81.6%). Επίσης η αξιοποίηση της πληροφορίας τι μέρος του λόγου είναι η κάθε λέξη, ωφελεί μόνο τον αλγόριθμο Maximum Entropy.

Οι Alexander Pak και Patrick Paroubek χρησιμοποίησαν την ίδια μέθοδο για να δημιουργήσουν την βάση των θετικών και αρνητικών tweets αλλά μέσω του twitter API σχημάτισαν και μια 3η κατηγορία tweets που προέρχονταν από λογαριασμούς ειδησεογραφικών πρακτορείων όπως οι NY Times, Washington Post κλπ. Σκοπός τους ήταν να αντιπροσωπεύουν κατά κύριο λόγο αντικειμενικά tweets που δεν εκφράζουν συναισθηματικό προσανατολισμό. Αφού κάθε λέξη αντιστοιχήθηκε με το μέρος του λόγου που ανήκει (Part of Speech Tagging), ανέλυσαν την σχέση που συνδέει τα μέρη του λόγου με την υποκειμενικότητα ή την αντικειμενικότητα ενός κειμένου. Χρησιμοποίησαν τον παρακάτω τύπο για να αποδώσουν σκορ υποκειμενικότητας-αντικειμενικότητας.

$$P_{1,2}^{T} = \frac{N_1^T - N_2^T}{N_1^T + N_2^T}$$

Όπου τα N_1^T, N_2^T αντιπροσωπεύουν την συχνότητα εμφάνισης του μέρους του λόγου στην κάθε κατηγορία. Κατέληξαν στο παρακάτω γράφημα που παρουσιάζει την κατανομή των διαφόρων μερών του λόγου (POS) σε αντικειμενικά και υποκειμενικά tweets. Για παράδειγμα οι προσωπικές αντωνυμίες έχουν υψηλό θετικό σκορ και συνδέονται με υποκειμενικά σχόλια ενώ η χρήση κύριων και κοινών ουσιαστικών συνδέεται με αντικειμενικά tweets.

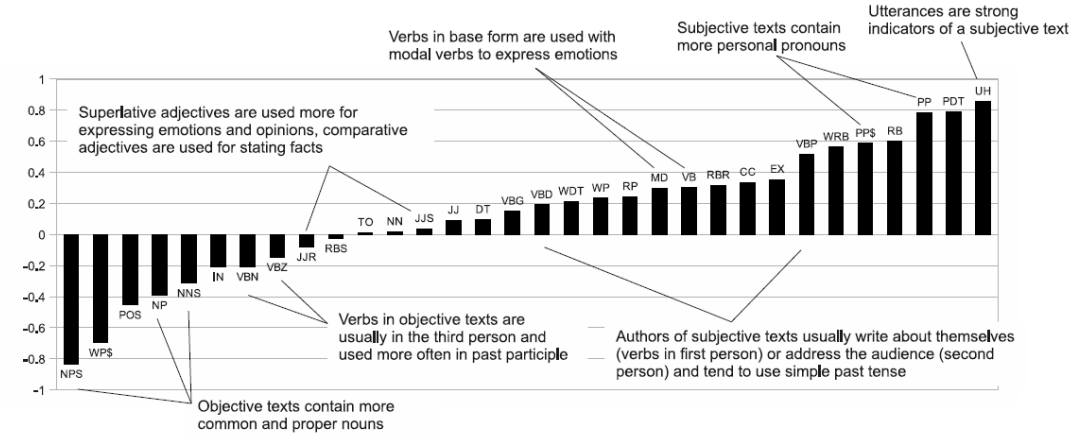


Figure 2: P^T values for objective vs. subjective

Π8. συχνότητα εμφάνισης των μερών του λόγου με βάση το συναίσθημα

Παρόμοια ανάλυση πραγματοποιούν για την σχέση που συνδέει την θετική-αρνητική φόρτιση του κειμένου με τα μέρη του λόγου που το αποτελούν.

Για την ανάλυση κλίματος χρησιμοποίησαν Multinomial Naïve Bayes, SVM και CRF πετυχαίνοντας τα καλύτερα αποτελέσματα μέσω του Multinomial Naïve Bayes που χρησιμοποιούσε για εκμάθηση την παρουσία n grams και την κατανομή των λέξεων στα διάφορα μέρη του λόγου (POS).

Σύμφωνα με τα αποτελέσματα που παρουσίασαν τα Bigrams (ανά 2 λέξεις) αποδίδουν τα καλύτερα αποτελέσματα σε συνδυασμό με τα POS tags.

Όπως και στην κλασική ανάλυση κλίματος η ανάλυση κλίματος που στοχεύει σε tweets παρουσιάζει βελτιωμένα αποτελέσματα όταν το μοντέλο λαμβάνει υπόψη του την

άρνηση εντός της πρότασης. Επίσης σύμφωνα με τον Go et al (2009) οι δημοσιεύσεις που περιέχουν URL τείνουν να είναι θετικά φορτισμένες.
Την ανωτερότητα του MNB για την ανάλυση κλίματος δεδομένων tweeter επιβεβαιώνουν και οι SidaWang και Christopher D. Manning στην εργασία τους «**Baselines and Bigrams: Simple, Good Sentiment and Topic Classification**»

Αν και ο αλγόριθμος MNB παρουσιάζει καλύτερα αποτελέσματα όταν το κείμενο είναι βραχύ όπως είναι τα tweets η αποτελεσματικότητα των διαφόρων αλγορίθμων εξαρτάται από τα χαρακτηριστικά που χρησιμοποιούμε για εκμάθηση και την προετοιμασία που έχουμε κάνει στα δεδομένα. Με την εργασία «**Sentiment analysis of twitter Data**» οι Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow και Rebecca Passonneau δοκιμάζουν διαφορετικούς τρόπους τροφοδότησης ενός μοντέλου SVM προκειμένου να πάρουν τα καλύτερα δυνατά αποτελέσματα. Μεταξύ των 3 μοντέλων που εξετάζουν είναι ένα μοντέλο Tree Kernel. Αυτό χρησιμοποιεί δενδροδιάγραμμα, αντί για διάνυσμα χαρακτηριστικών όπως παρουσιάζεται παρακάτω για να απεικονίσει το κάθε tweet και υπολογίζει την ομοιότητα μεταξύ των tweet σύμφωνα με την ομοιογένεια των δέντρων και όλων των πιθανών κλάδων. Το δεύτερο μοντέλο λαμβάνει υπόψη του χαρακτηριστικά που αντιπροσωπεύονται από σκορ συναισθηματικής φόρτισης και το τρίτο μοντέλο είναι μοντέλο unigrams ως βάση αναφοράς.

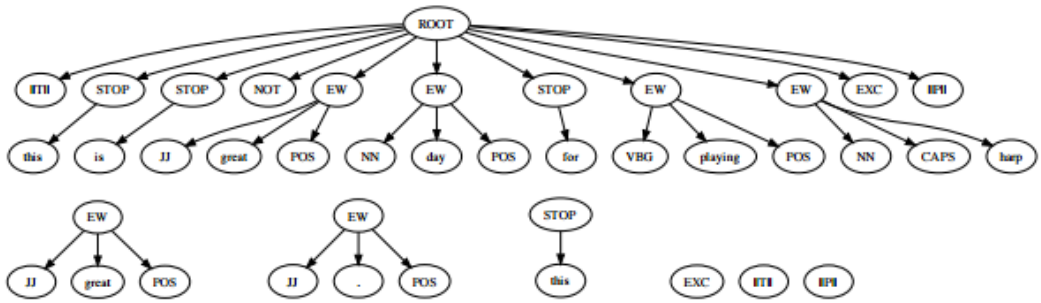


Figure 1: Tree kernel for a synthesized tweet: "@Fernando this isn't a great day for playing the HARP! :)"

Π9. Δενδροδιάγραμμα Χαρακτηριστικών

Τα αποτελέσματα του Tree Kernel είναι εξίσου ικανοποιητικά με το πιο σύνθετο μοντέλο που λαμβάνει υπόψη του POS tags και τον συναισθηματικό προσανατολισμό των λέξεων.

Ο χαρακτήρας των tweets και γενικότερα της επικοινωνίας που εμφανίζεται στο διαδίκτυο μεταξύ χρηστών διαφέρει σημαντικά από τον κλασικό γραπτό λόγο. Σίγουρα είναι λιγότερο έως καθόλου επίσημος, πολλές φορές προσπαθεί να προσομοιάσει τον προφορικό λόγο και κάποιες φορές περιλαμβάνει αργκό (slang). Οι χρήστες προκειμένου να δώσουν έμφαση χρησιμοποιούν κεφαλαία γράμματα, πολλαπλά θαυμαστικά ή/και άλλα σημεία στίξης ή/και πολλαπλά γράμματα συνεχόμενα μέσα στην λέξη (π.χ I Looooove going to the Beach). Αρκετές φορές ο χρήστης προσπαθεί να βρει τον συντομότερο δυνατό τρόπο για να αποδώσει το νόημα που θέλει να εκφράσει χρησιμοποιώντας συντμήσεις, emoticons, emojis κλπ.

Αρκετές εργασίες έχουν προσπαθήσει να μετρήσουν την αξία των ιδιαίτερων αυτών χαρακτηριστικών και να την εντάξουν στην ανάλυση κλίματος. Έχουμε ήδη αναφερθεί στην ισχυρή σχέση που συνδέει τα emoticons με την συναισθηματική φόρτιση του κειμένου καθώς και την υποκειμενικότητά του.Οι Efthymios Kouloumpis, Theresa Wilson και Johanna Moore με την εργασία τους «**Twitter Sentiment Analysis: The Good the Bad and the**

OMG!» αναδεικνύουν την αξία των hashtags προκειμένου να συλλέξουν δεδομένα για να εκπαιδεύσουν ικανοποιητικά τον αλγόριθμο. Επίσης δοκιμάζουν συνδυασμό δεδομένων hashtags με δεδομένα που περιέχουν emoticons.

Positive	#iloveitwhen, #thingsilike, #bestfeeling, #bestfeelingever, #omgthatssotrue, #imthankfulfor, #thingsilove, #success
Negative	#fail, #epicfail, #nevertrust, #worst, #worse, #worstlies, #imtiredof, #itsnotokay, #worstfeeling, #notcute, #somethingaintright, #somethingsnotright, #ihate
Neutral	#job, #tweetajob, #omgfacts, #news, #listeningto, #lastfm, #hiring, #cnn

Table 3: Top positive, negative and neutral hashtags used to create the HASH data set

Π10. Hastag με βάση την συναισθηματική φόρτιση

Επίσης καταλήγουν πως σημαντική βοήθεια για την αναγνώριση του συναισθηματικού προσανατολισμού παρέχουν τα κεφαλαία γράμματα, τα επαναλαμβανόμενα γράμματα μαζί με τα emoticons.

Με την εργασία τους «**Sentiment of Emojis**» οι Petra Kralj Novak, Jasmina Smailović, Borut Sluban και Igor Mozetič αναλύουν την σχέση των emojis με τον συναισθηματικό προσανατολισμό των tweets. Τα emojis , μερικά από τα πιο δημοφιλή παρουσιάζονται στην εργασία τους με τον παρακάτω πίνακα, δημιουργήθηκαν στην Ιαπωνία στο τέλος του 20ου αιώνα και έγιναν δημοφιλή σε όλο τον κόσμο όταν τα συμπεριέλαβε η εταιρεία Apple στις συσκευές iphone το 2010. Χωρίζονται σε πολλές κατηγορίες με σκοπό μεταξύ άλλων τον εμπλουτισμό των κειμένων, την συντόμευση του γραπτού λόγου και φυσικά την συναισθηματική έκφραση.

Emoji	N	Position	p_-	p_0	p_+	$\overline{s}$	Name
😂	14,622	0.80	0.25	0.29	0.47	0.22	face with tears of joy
❤	8,050	0.74	0.04	0.17	0.79	0.75	heavy black heart
♥	7,144	0.75	0.04	0.27	0.69	0.66	black heart suit
😍	6,359	0.76	0.05	0.22	0.73	0.68	smiling face with heart-shaped eyes
😭	5,526	0.80	0.44	0.22	0.34	-0.09	loudly crying face
😘	3,648	0.85	0.05	0.19	0.75	0.70	face throwing a kiss
😊	3,186	0.81	0.06	0.24	0.70	0.64	smiling face with smiling eyes
👌	2,925	0.80	0.09	0.25	0.66	0.56	ok hand sign
💕	2,400	0.76	0.04	0.29	0.67	0.63	two hearts
👏	2,336	0.78	0.10	0.27	0.62	0.52	clapping hands sign

Fig 1. Top 10 emojis. Emojis are ordered by the number of occurrences *N*. The average *position* ranges from 0 (the beginning of the tweets) to 1 (the end of the tweets). $p_c, c \in \{-1, 0, +1\}$, are the negativity, neutrality, and positivity, respectively. $\overline{s}$ is the sentiment score.

doi:10.1371/journal.pone.0144296.g001

Π11. Emojis και συναισθηματικός προσανατολισμός κειμένου

Στην εργασία τους χρησιμοποιούν ένα μεγάλο δείγμα (1.6 εκατομμύρια tweets 15 διαφορετικών ευρωπαϊκών γλωσσών εκ των οποίων 4% περιέχουν emojis) για να αναδείξουν την σχέση που συνδέει τα emojis με την αρνητική, θετική ή ουδέτερη φόρτιση του κειμένου. Χρησιμοποιώντας διάφορα στατιστικά τεστ όπως το Welch's t-test καταλήγουν στο συμπέρασμα ότι τα tweets που περιλαμβάνουν emojis είναι σε μεγαλύτερο ποσοστό θετικού προσανατολισμού (53,9%) σε σχέση με αυτά που δεν περιλαμβάνουν emojis (36,6%). Η σχέση αυτή εμφανίζεται ανεξάρτητη από την γλώσσα στην οποία είναι γραμμένα τα tweets.

Επίσης τα emojis που χρησιμοποιούνται συχνότερα συναντώνται σε θετικά σχόλια ή αλλιώς τα θετικά emojis είναι τα πλέον δημοφιλή. Με την εργασία τους δημιούργησαν λίστα με τα δημοφιλέστερα emojis (2015) , τα ποσοστά που αντιστοιχούν σε θετικά, αρνητικά και ουδέτερα tweets, scores και διαστήματα εμπιστοσύνης.

SEMEVAL

Παράλληλα με την ανάπτυξη και διεύρυνση του κλάδου της επιστήμης που ασχολείται με την ανάλυση του λόγου, γραπτού και προφορικού, έχουν καθιερωθεί στην επιστημονική κοινότητα τα εργαστήρια-συνέδρια **semeval** στα οποία δίνεται και η δυνατότητα αξιολόγησης και σύγκρισης διαφόρων μεθόδων και τεχνικών για την βελτιστοποίηση συγκεκριμένων διαδικασιών. Φυσικά η ανάλυση κλίματος ή sentiment analysis είναι ενεργό πεδίο για το Semeval που διοργανώνεται κάθε έτος σε γενικότερα ή και ειδικότερα πλαίσια.

Η ημερίδα που διοργανώθηκε το 2015, μεταξύ άλλων περιελάμβανε την σύγκριση διαφορετικών μεθόδων για την συναισθηματική κατηγοριοποίηση κειμένων Twitter και κειμένων Twitter με έντονο μεταφορικό/αλληγορικό περιεχόμενο. Στην Δεύτερη κατηγορία διαγωνίστηκε και η ομάδα DsUniPi με το σύστημα της Μαρίας Καρανάσου που περιγράφεται στην διπλωματική εργασία **«Ανάλυση συναισθήματος σε Κοινωνικά δίκτυα, ο μεταφορικός λόγος στο Twitter».** Η εφαρμογή είναι εξ ολοκλήρου σε python και περιλαμβάνει καθαρισμό των κειμένων, εξαγωγή χαρακτηριστικών, αντιστοίχιση σκορ με βάση την πολικότητα κάποιων χαρακτηριστικών πχ. #hastags, υπολογισμό της εννοιολογικής ομοιομορφίας των κειμένων και ανάθεση σκορ καθώς και σκορ με βάση λεξικά όπως το Sentiwordnet προκειμένου να τροφοδοτήσει τους αλγόριθμους κατηγοριοποίησης. Οι αλγόριθμοι που χρησιμοποιεί είναι μεταξύ άλλων οι SVM με Linear Kernel, SVM με Radial Base Kernel, Naïve Bayes, Decision Tree κλπ. Τα καλύτερα αποτελέσματα στην φάση της προετοιμασίας τα πέτυχε ο αλγόριθμος SVM με linear kernel αλλά στα δεδομένα της δοκιμασίας ο SVR είχε την καλύτερη απόδοση.

Το συγκεκριμένο σύστημα ήρθε 10^{ο} στα αποτελέσματα του διαγωνισμού σε σύνολο 15 συμμετοχών.

Κ Ε Φ Α Λ Α Ι Ο 3

Εφαρμογή

3.1 Εισαγωγή

Στα πλαίσια της παρούσας εργασίας δημιουργήθηκε ένα αυτοτελές σύστημα (supervised machine learning) συναισθηματικής ανάλυσης (ανάλυσης κλίματος) γραπτού λόγου σε δεδομένα Twitter χρησιμοποιώντας κυρίως το πακέτο στατιστικού προγραμματισμού R.

Το σύστημα αυτό αναλύεται διεξοδικά στην συνέχεια μαζί με κάποια βοηθητικά συστήματα που δημιουργήθηκαν ώστε να διευρύνουν το πεδίο εφαρμογών του. Μερικές από αυτές τις δυνατότητες είναι η εισαγωγή δεδομένων (tweets) κατευθείαν από το Twitter API είτε για ανάλυση, είτε για κατηγοριοποίηση καθώς και η σύνδεση της R με την βάση δεδομένων MySQL και το Interface της εφαρμογής MySQL Workbench.

Ως σημείο αναφοράς για να στηρίξουμε την ανάλυση και τα αποτελέσματα αξιοποιείται η διπλωματική εργασία της **Μαρίας Καρανάσου: «Ανάλυση συναισθήματος σε Κοινωνικά δίκτυα, ο μεταφορικός λόγος στο Twitter»** στα πλαίσια του Π.Μ.Σ. «Ψηφιακά Συστήματα & Υπηρεσίες» του τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς. Η συγκεκριμένη εφαρμογή ήταν εξ ολοκλήρου σε γλώσσα Python και τα δεδομένα παρέχονται σε μορφή κώδικα MySQL. Περιλαμβάνουν την βάση που χρησιμοποιήθηκε για την εκπαίδευση και τα πειράματα αλλά και την βάση της τελικής δοκιμής-δοκιμασίας στα πλαίσια του **SEMEVAL 2015.**

Το σύστημα που έχουμε δημιουργήσει δοκιμάζεται στα συγκεκριμένα δεδομένα για λόγους σύγκρισης της αποτελεσματικότητας. Το παρόν σύστημα αξιοποιεί στην ανάλυση την πληροφορία που προέρχεται από την χρήση emojis σε αντίθεση με το σύστημα της **Μαρίας Καρανάσου.** Καθώς όμως στις συγκεκριμένες βάσεις δεδομένων δεν περιλαμβάνονται emojis η απόδοση του συστήματος δεν μπορεί να συγκριθεί στο μέγιστο βαθμό.

Προκειμένου να αξιοποιήσουμε στην ανάλυση την αντιστοίχηση των λέξεων με τα μέρη του λόγου όπου ανήκουν, κρίθηκε σκόπιμο να χρησιμοποιήσουμε την γλώσσα προγραμματισμού Python 2.7 καθώς η R δεν παρέχει ικανοποιητική βιβλιοθήκη για την συγκεκριμένη διαδικασία ιδιαίτερα σε περιβάλλον windows.

Το κεφάλαιο που ακολουθεί είναι χωρισμένο σε 3 μέρη, το πρώτο μέρος αφορά τα δεδομένα, το δεύτερο μέρος την δομή του συστήματος και τις φάσεις της ανάλυσης και το 3ο τα διάφορα πειράματα που πραγματοποιήθηκαν,

3.2 Δεδομένα

Γενικά τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Δεδομένα εκπαίδευσης

Τα δεδομένα εκπαίδευσης είναι αυτά που χρησιμοποιούμε στη πρώτη φάση προκειμένου να εκπαιδεύσουμε τους αλγόριθμους του συστήματος μετά την απαραίτητη επεξεργασία τους. Τα δεδομένα εκπαίδευσης θα πρέπει να είναι σε μορφή πίνακα. Κάθε γραμμή αποτελεί ξεχωριστή παρατήρηση. Για παράδειγμα, στο παρακάτω πλαίσιο δεδομένων η πρώτη στήλη περιέχει το κείμενο και η δεύτερη στήλη το σκορ ή την κατηγορία που ανήκει το κάθε κείμενο στην προκειμένη περίπτωση Αρνητικό, Ουδέτερο, Θετικό. Θα μπορούσαν τα κείμενα να διαχωρίζονται σε αρνητικά, θετικά ή να λάβουμε υπόψη και την ένταση του συναισθήματος και να διαχωρίζονται σε πολύ αρνητικό, αρνητικό, ουδέτερο, θετικό, πολύ θετικό κλπ

tweet	sentiment
Gas by my house hit $3.39!!!! I'm going to Chapel Hill on Sat. :)	positive
Iranian general says Israel's Iron Dome can't deal with their missiles (keep talking like that and we may end up finding out)	negative
with J Davlar 11th. Main rivals are team Poland. Hopefully we an make it a successful end to a tough week of training tomorrow.	positive
Talking about ACT's && SAT's, deciding where I want to go to college, applying to colleges and everything about college stresses me out.	negative
They may have a SuperBowl in Dallas, but Dallas ain't winning a SuperBowl. Not with that quarterback and owner. @S4NYC @RasmussenPoll	negative
Im bringing the monster load of candy tomorrow, I just hope it doesn't get all squiched	objective-OR-neutral
Apple software, retail chiefs out in overhaul: SAN FRANCISCO Apple Inc CEO Tim Cook on Monday replaced the heads... bit.ly/XQEhJU	objective-OR-neutral
@oluoch @victor_otti @kunjand I just watched it! Sridevi's comeback.... U remember her from the 90s?? Sun mornings on NTA ;)	positive
#Livewire Nadal confirmed for Mexican Open in February: Rafael Nadal is set to play at the Me... bit.ly/WY4Vjy #LiveWireAthletics	objective-OR-neutral
@MsSheLahY I didnt want to just pop up... but yep we have chapel hill next wednesday you should come.. and shes great ill tell her you asked	positive
@Alyoup005 @addicted2haley hmmmm November is an odd release date if true but if it becomes big enough maybe she could sing it at Grammys	objective-OR-neutral
#Iran US delisting MKO from global terrorists list in line with Iran campaign: Tehran, Oct 30, IRNA -- Secretary... bit.ly/XSTGtd	objective-OR-neutral
Good Morning Becky ! Thursday is going to be Fantastic ! @SwedenG @DJ4JG @Grdina @Paverlayer @FSBull @RevkahJC @DicksTrash @borderfox116	positive
Expect light-moderate rains over E. Visayas; Cebu, Bohol, Samar & Leyte have 30-70% chance of rains tonight! Expect fair weather tomorrow!:)	objective-OR-neutral
One ticket left for the @49ers game tomorrow! Don't miss the rematch of the NFC Championship game against the NY Giants! Hit me up!	positive
AFC away fans on Saturday. All this stuff about the 'she said no' chant. It's bollocks. When he has the ball, just turn your back on him.	negative

Π12. παράδειγμα δεδομένων με 3 κατηγορίες

Δεδομένα Ελέγχου

Τα δεδομένα ελέγχου μπορεί να προέρχονται από το ίδιο πλαίσιο δεδομένων ή από κάποιο άλλο. Μπορεί να περιλαμβάνουν και σκορ για να ελέγξουμε την αποτελεσματικότητα του αλγόριθμου ή να μην περιλαμβάνουν σκορ.

Τροφοδοτούνται στον αλγόριθμο κατηγοριοποίησης και με βάση την προηγούμενη εκπαίδευσή του ο αλγόριθμος εκχωρεί τα κείμενα στις επί μέρους κατηγορίες.

Πίνακας Χαρακτηριστικών

Η δημιουργία ενός μεγάλου πίνακα που περιλαμβάνει όλα τα χαρακτηριστικά με βάση τα οποία θα γίνει η κατηγοριοποίηση και ο εκ των υστέρων, διαχωρισμός του πίνακα σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου δίνει καλύτερα αποτελέσματα. Ρεαλιστικότερη είναι όμως η περίπτωση που δημιουργείτε ένας πίνακας με χαρακτηριστικά (στήλες) που προκύπτουν από τα δεδομένα εκπαίδευσης και η ανάλυση γίνεται στα χαρακτηριστικά των

δεδομένων ελέγχου που είναι κοινά με τα δεδομένα εκπαίδευσης. Στην συγκεκριμένη εργασία έχουμε εξετάσει και τις 2 περιπτώσεις.

Τα δεδομένα που χρησιμοποιούνται για την ανάλυση αποτελούνται από 2 δεξαμενές κειμένων-tweets (χαρακτηρισμένων) που χρησιμοποιούνται στη διπλωματική εργασία της **Μαρίας Καρανάσου: «Ανάλυση συναισθήματος σε Κοινωνικά δίκτυα, ο μεταφορικός λόγος στο Twitter».** Είναι κείμενα που περιλαμβάνουν έντονο μεταφορικό λόγο, σαρκασμό και γενικά μπορούμε να πούμε ότι ο βαθμός δυσκολίας είναι υψηλός για ένα αυτοματοποιημένο σύστημα. Το πρώτο σετ (8529 κείμενα tweets) αποτελούν τα δεδομένα στα οποία έγιναν οι δοκιμές-πειράματα προκειμένου να υλοποιηθεί και να βελτιστοποιηθεί το σύστημα για τον διαγωνισμό **SEMEVAL 2015**. Τα ίδια δεδομένα χρησιμοποιήθηκαν στο σύνολό τους σαν δεδομένα εκπαίδευσης στην φάση της τελικής δοκιμασίας στα πλαίσια του **SEMEVAL 2015.** Το δεύτερο σετ (4000 κείμενα tweets) περιλαμβάνει τα δεδομένα που χρησιμοποιήθηκαν αποκλειστικά σαν δεδομένα ελέγχου κατά την δοκιμή.

Στην παρούσα εργασία ακολουθήθηκε η ίδια διαδικασία. Οι δοκιμές και βελτιώσεις πραγματοποιήθηκαν με το πρώτο σετ των 8529 κειμένων το οποίο χωρίστηκε σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Το τελικό σετ των 4000 κειμένων χρησιμοποιήθηκε αποκλειστικά για τον έλεγχο του συστήματος που εκπαιδεύτηκε με το σύνολο των 8529 κειμένων. Στον παρακάτω πίνακα παρουσιάζεται η κατανομή των κειμένων των δύο σετ στις 3 ομάδες

ΚΑΤΗΓΟΡΙΑ	ΔΕΔ. ΠΕΙΡΑΜΑΤΩΝ	ΔΕΔ. ΤΕΛΙΚΗΣ ΔΟΚΙΜΗΣ
NEGATIVE	7264	3062
NEUTRAL	565	298
POSITIVE	700	640
TOTAL	**8529**	**4000**

Π13.Κατανομή των σετ δεδομένων με βάση τον συναισθηματικό τους προσανατολισμό

3.3 Σύστημα- μηχανισμός

Τα δεδομένα τροφοδοτούνται στο σύστημα αφού πρώτα έχουμε δημιουργήσει μια στήλη που αναλύεται το κάθε κείμενο στα μέρη του λόγου που το αποτελούν. Για την κατηγοριοποίηση των λέξεων σε μέρη του λόγου έχουν χρησιμοποιηθεί πολλές προσεγγίσεις. Κάποιες στηρίζονται σε γραμματικούς κανόνες που καθορίζουν τις κατηγορίες, άλλες στην εκπαίδευση αλγορίθμου με κατηγοριοποιημένα δεδομένα κλπ. Στην συγκεκριμένη εργασία αξιοποιήσαμε την κατηγοριοποίηση που κατέληξε ο αλγόριθμος Gate Pos Tagger στην εργασία της Μαρίας Καρανάσου και είναι εξειδικευμένος στα δεδομένα twitter. Οι σημαντικότερες κατηγορίες στις οποίες χωρίζεται ο αγγλικός λόγος παρουσιάζονται στον παρακάτω πίνακα.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Η σημασία του POS Tagging που έχει αναδειχθεί και στο κεφάλαιο της ιστορικής ανασκόπησης επιβεβαιώνεται και από την παρούσα εργασία.

Φάση Ι: Προπαρασκευή Δεδομένων

Στο πρώτο στάδιο τα δεδομένα αφού έχουν εισαχθεί στο σύστημα μετατρέπονται σε πλαίσιο δεδομένων (data frame) με συγκεκριμένη γραμματοσειρά και κωδικοποίηση (ASCII). Τα δεδομένα σαρώνονται και αφαιρούνται οι διπλές παρατηρήσεις από το δείγμα. Συχνά υπάρχουν tweets πανομοιότυπα που διαφέρουν σε κάποια μικρή λεπτομέρεια όπως για παράδειγμα ένα σημείο στίξης ή την ένδειξη RT. Αυτό μπορεί να συμβαίνει σκόπιμα και συνήθως πρόκειται για spam tweets η πρόκειται για αναδημοσίευση (Retweet). Η αφαίρεση των διπλών παρατηρήσεων από το δείγμα θα μειώσει τον θόρυβο από την ανάλυση αλλά και η αποτελεσματικότητά του θα είναι πλέον σε ρεαλιστικότερη βάση. Η τροφοδότηση του συστήματος μπορεί να γίνει και με το κείμενο ή με το κείμενο που προκύπτει από τον Pos Tagger, τις λέξεις του κειμένου μαζί με την ένδειξη για το μέρος του λόγου όπου ανήκουν.

Text	The combination of obese and bikini is so beautiful.
Pos Tagged Text	u'beautiful': 'JJ', u'obese': 'VBD', u'bikini': 'NN', u'so': 'IN'....

Π14. Παράδειγμα Pos Tagging

Φάση ΙΙ Αναγνώριση emojis

Σε αυτή την φάση σαρώνουμε τα δεδομένα για την ύπαρξη emojis. Όπως αναφέραμε και παραπάνω η χρήση των emojis γίνεται συνήθως για να φορτίσουμε συναισθηματικά το κείμενο, ή (λιγότερο συχνά), για να περιγράψουμε κάτι με εικόνες αντί για λέξεις. Μερικά από τα emojis, κάποια πολύ συνηθισμένα και κάποια μάλλον σπάνια, παρουσιάζονται στον παρακάτω πίνακα:

Π15. Παραδείγματα Emojis

Καθώς η χρήση τους είναι διαδεδομένη δεν θα μπορούσαμε να μην τα χρησιμοποιήσουμε στην ανάλυση. Η **R** όμως δεν αναγνωρίζει τα emojis σαν εικονίδια, αλλά τα αναπαράγει μέσα στο κείμενο κωδικοποιημένα. Χάρις την εργασία **του Kirill Pomogajko [Emoticons Decoder for Social media sentiment analysis in R]** έχουμε στην διάθεσή μας την αντιστοίχιση των emojis με την κωδικοποίηση της R για 840 διαφορετικά emojis. Για την ανίχνευση των εικονιδίων, χρησιμοποιούμε ακριβώς αυτή την αντιστοίχιση κατευθείαν από το διαδίκτυο και από την διεύθυνση:
https://raw.githubusercontent.com/today-is-a-good-day/Emoticons/master/emDict.csv.

Η τελευταία στήλη όπως φαίνεται και παρακάτω περιλαμβάνει την κωδικοποίηση των εικονιδίων στην R.

```
"Description";"Native";"Bytes";"R-encoding"
"AERIAL TRAMWAY";"🚡";"\xF0\x9F\x9A\xA1";"<ed><a0><bd><ed><ba><a1>"
"AIRPLANE";"✈";"\xE2\x9C\x88";"<e2><9c><88>"
"ALARM CLOCK";"⏰";"\xE2\x8F\xB0";"<e2><8f><b0>"
"ALIEN MONSTER";"👾";"\xF0\x9F\x91\xBE";"<ed><a0><bd><ed><b1><be>"
"AMBULANCE";"🚑";"\xF0\x9F\x9A\x91";"<ed><a0><bd><ed><ba><91>"
"AMERICAN FOOTBALL";"🏈";"\xF0\x9F\x8F\x88";"<ed><a0><bc><ed><bf><88>"
"ANCHOR";"⚓";"\xE2\x9A\x93";"<e2><9a><93>"
"ANGER SYMBOL";"💢";"\xF0\x9F\x92\xA2";"<ed><a0><bd><ed><b2><a2>"
"ANGRY FACE";"😠";"\xF0\x9F\x98\xA0";"<ed><a0><bd><ed><b8><a0>"
"ANGUISHED FACE";"😧";"\xF0\x9F\x98\xA7";"<ed><a0><bd><ed><b8><a7>"
"ANT";"🐜";"\xF0\x9F\x90\x9C";"<ed><a0><bd><ed><b0><9c>"
"ANTENNA WITH BARS";"📶";"\xF0\x9F\x93\xB6";"<ed><a0><bd><ed><b3><b6>"
"ANTICLOCKWISE DOWNWARDS AND UPWARDS OPEN CIRCLE ARROWS";"🔄";"\xF0\x9F\x94\x84";"<ed><a0><bd><ed><b4><84>"
```

Π16. Παράδειγμα βάσης Δεδομένων κωδικοποίησης Emojis

Χρησιμοποιούμε την grepl() ρουτίνα της R και ανιχνεύουμε στο κάθε κείμενο την ύπαρξη των εικονιδίων emojis.

Για να αξιοποιήσουμε την πληροφορία, από την ύπαρξη ενός ή πολλών εικονιδίων στο κείμενο, χρησιμοποιούμε απ' ευθείας από το διαδίκτυο μέσω της διεύθυνσης http://kt.ijs.si/data/Emoji_sentiment_rankin τον πίνακα της εργασία των **P. Kralj Novak, J. Smailovic, B. Sluban, I. Mozetic.** Χρησιμοποιώντας ως κλειδί την στήλη Unicode Name

ώστε να συνδέσουμε τους 2 πίνακες αντιστοιχούμε στο κάθε κείμενο τον υψηλότερο μέσω όρο της πιθανότητας των τριών κατηγοριών για τα emojis που ανιχνεύσαμε. Η πιθανότητα αυτή θα αξιοποιηθεί στο τελευταίο στάδιο επικουρικά στα αποτελέσματα του αλγόριθμου κατηγοριοποίησης. Στον πίνακα που ακολουθεί παρουσιάζεται μέρος της πληροφορίας διαθέσιμης στο διαδίκτυο όπου στο κάθε εικονίδιο emoji αντιστοιχείται μεταξύ άλλων το Unicode, η συχνότητα εμφάνισης, η ταξινόμηση, η πιθανότητα να είναι αρνητικό, η πιθανότητα να είναι ουδέτερο, η πιθανότητα να είναι θετικό.

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	3648	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

Π17. Πληροφορία διαθέσιμη στη σελίδα http://kt.ijs.si/data/Emoji_sentiment_rankin

Πλέον αφού έχουμε τελειώσει με την αναγνώριση των εικονιδίων στα κείμενα και την αντιστοίχιση πιθανοτήτων προχωρούμε στη επόμενη φάση που αφορά το δεύτερο μέρος των χαρακτηριστικών που θα χρησιμοποιήσουμε για την εκπαίδευση του αλγόριθμου.

Φάση ΙΙΙ: Εξαγωγή και επεξεργασία επιμέρους χαρακτηριστικών

Στην φάση αυτή ανιχνεύουμε την παρουσία επιμέρους χαρακτηριστικών που θα μπορούσαν να έχουν κάποια ιδιαίτερη αξία για την εκπαίδευση του αλγόριθμου και θα πρέπει να την ενισχύσουμε κατάλληλα. Τα χαρακτηριστικά αυτά θα πρέπει να πληρούν κάποιες προϋποθέσεις. Να έχουν μια ειδική βαρύτητα για την καταχώρηση του κειμένου στις διάφορες κατηγορίες και να μην εξειδικεύουν, ώστε να αφορούν αρκετές περιπτώσεις κειμένων. Ένα τέτοιο χαρακτηριστικό είναι η άρνηση. Μια πρακτική είναι η αντικατάσταση των λέξεων που δηλώνουν άρνηση με το NOT των **Das και Chen (2001)**. Σκοπός τους είναι να δημιουργήσουν μια γενικότερη-ισχυρότερη ομάδα NOT συνενώνοντας τις επιμέρους ομάδες των διαφόρων λέξεων που εκφράζουν άρνηση. Ένα άλλο χαρακτηριστικό είναι η ύπαρξη κεφαλαίων γραμμάτων που αρκετές φορές χρησιμοποιούνται για να δώσουν έμφαση στο κείμενο ή οι συνεχόμενοι χαρακτήρες που επίσης χρησιμοποιούνται για να δώσουν έμφαση σε κάποια λέξη ή φράση. Τα τελευταία χαρακτηριστικά μπορεί να μην βοηθούν ιδιαίτερα στον διαχωρισμό των κειμένων από θετικά σε αρνητικά, όπως η ύπαρξη θετικών

emoticons, αλλά βοηθούν στον διαχωρισμό ουδέτερων-αντικειμενικών με υποκειμενικά-συναισθηματικά φορτισμένα κείμενα. Τα χαρακτηριστικά αυτά εφόσον ανιχνευθούν θα τροφοδοτηθούν στο σύστημα αντιπροσωπευμένα είτε από binary μεταβλητές που θα εκφράζουν την παρουσία τους ή όχι στο κείμενο είτε από count μεταβλητές που αποτυπώνουν την συχνότητά εμφάνισης τους στο κείμενο. Στο συγκεκριμένο σύστημα τα επιμέρους χαρακτηριστικά που έχουμε συμπεριλάβει στην ανάλυση είναι τα παρακάτω:

- Σημεία στίξης (Count)
- Θετικά Emoticons (Binary)
- Αρνητικά Emoticons (Binary)
- Αν περιλαμβάνει περισσότερους από 2 συνεχόμενους ίδιους χαρακτήρες. (Binary)
- Χαρακτήρες που εκφράζουν γέλιο όπως LOL, lol, haha… (Binary)
- Αν το κείμενο αποτελεί απάντηση του χρήστη σε άλλο χρήστη, ή αναφέρει/κοινοποιεί άλλον χρήστη. (Binary)
- Αρνήσεις (Binary)
- Αν το tweet είναι Retweet ή όχι (Binary)
- Αν το tweet έχει συνδεθεί με κάποιο hash tag (Binary)
- Αν το κείμενο περιλαμβάνει URL (Binary)
- Τον αριθμό των κεφαλαίων χαρακτήρων. (Count)
- Τα διάφορα μέρη του λόγου που ανήκουν οι λέξεις POS Tags. (Count)

Για την ανίχνευση των παραπάνω χαρακτηριστικών χρησιμοποιήθηκαν οι ρουτίνες της R gregexpr() και grepl() σε επίπεδο γραμμής. Για παράδειγμα με τον παρακάτω κώδικα ανιχνεύουμε σε κάθε γραμμή-tweet την ύπαρξη θετικών emoticons. Τα αποτελέσματα της αναζήτησης επιστρέφονται σε ένα data frame και αποτυπώνουν την συχνότητα εμφάνισης του χαρακτηριστικού. Στην προκειμένη περίπτωση μπορούμε να κρατήσουμε 12 διαφορετικές στήλες που να αντιπροσωπεύουν τα αποτελέσματα των 12 emoticons. Εφόσον όμως μπορούμε να θεωρήσουμε ότι θα έχουν την ίδια επίδραση στην κατηγοριοποίηση, πιο σωστή προσέγγιση είναι η δημιουργία μιας μεγάλης ομάδας που θα αφορά περισσότερες παρατηρήσεις. Επίσης, όπως αναφέρθηκε και στο κεφάλαιο 2 σε αρκετές περιπτώσεις η συχνότητα δεν βοηθάει περισσότερο από την παρουσία ενός χαρακτηριστικού. Πρακτικά μπορούμε να καταλάβουμε τον λόγο, αφού η συχνότητα δημιουργεί υποπεριπτώσεις που δεν βοηθούν τον αλγόριθμο της κατηγοριοποίηση.

```
emoticons_simple_positive<-c(":)", ":-)", ":-D", ":))", ":-))", "\\(:", ":D", ";)", ":-)","XD",
";D",":]")                                          # define  symbols#

gemoticons_simple_positive<-lapply(emoticons_simple_positive, function (x) j<-
gregexpr(x,tweets_encod))                  ### search tweets for symbols##

emoticons_simple_positive_df=as.data.frame(matrix(data=NA, nrow=length(tweets_encod),
ncol=length(emoticons_simple_positive)))  ### create an empty matrix###
for (i in 1:12) {emoticons_simple_positive_df[[i]]<-sapply(gemoticons_simple_positive[[i]],
function(x) if(x[[1]]!=-1) length(x) else 0)}    #### populate matrix ####
names(emoticons_simple_positive_df)<-c(":)", ":-)", ":-D", ":))", ":-))", "(:", ":D", ";)", ":-
)","XD", ";D",":]")                                  ##### assign column names #####
```

Στο πρώτο κομμάτι του κώδικα ορίζουμε τα σύμβολα που θα αναζητήσουμε. Στο συγκεκριμένο παράδειγμα είναι 12 θετικά emoticons που χρησιμοποιούνται πιο συχνά. Στο δεύτερο κομμάτι αναζητούμε τα σύμβολα εντός των κειμένων και στο τέταρτο κομμάτι δημιουργούμε ένα πίνακα που περιέχει τα αποτελέσματα. Η εφαρμογή του κώδικα στο 80% των δεδομένων πειραμάτων έχει ανιχνεύσει 147 περιπτώσεις, όπως φαίνεται και στον παρακάτω πίνακα που προκύπτει από την εντολή colSums:

colSums(emoticons_simple_positive_df)

:)	:-)	:-D	:))	:-))	(:	:D	;)	:-)	XD	;D	:]
68	18	2	8	0	2	8	20	18	3	0	0

Αν σταματήσουμε εδώ, θα έχουμε καταγράψει την συχνότητα των εμφανίσεων των 12 emoticons για κάθε tweet σε ένα πίνακα με 12 στήλες και γραμμές όσες είναι τα δεδομένα Καλύτερα είναι όμως να προχωρήσουμε και να συμπτύξουμε τον πίνακα σε μια στήλη που θα αποτυπώνει απλώς την παρουσία θετικών emoticons.

```
smile<-rowSums(emoticons_simple_positive_df)
smile_binary <- ifelse(smile>0 ,1,0)
#  colapse all positive emoticons into 1 column and keep occurance not frequence 0 1#
```

Παρόμοια κινούμαστε και με τα υπόλοιπα χαρακτηριστικά και δημιουργούμε ένα πίνακα του οποίου οι γραμμές αντιπροσωπεύουν τα κείμενα και οι στήλες την παρουσία ή την συχνότητα των χαρακτηριστικών. Ο συγκεκριμένος πίνακας μας δίνει την δυνατότητα να ελέγξουμε στην συνέχεια και την αλληλεπίδραση μεταξύ των χαρακτηριστικών. Πολλαπλασιάζοντας κατάλληλα ανά 2 τις μεταβλητές μπορούμε να πάρουμε τις αλληλεπιδράσεις $2^{ης}$ τάξης.

Αυτός ο πίνακας δεν είναι ο βασικός πίνακας, για την ακρίβεια είναι οι πρώτες στήλες ενός πολύ μεγαλύτερου πίνακα με χιλιάδες στήλες που θα δημιουργήσουμε στην επόμενη φάση.

Φάση IV: εξαγωγή και επεξεργασία των κυρίως χαρακτηριστικών.

Στην Φάση αυτή γίνεται η βασική ανάλυση των κειμένων σε επίπεδο λέξης. Για την ανάλυση χρησιμοποιούμε την βιβλιοθήκη **RtextTools** η οποία παρέχει κάποια σημαντικά

εργαλεία για ανάλυση κειμένου εκμεταλλευόμενη και άλλες βιβλιοθήκες της R. Με την παρακάτω εντολή δημιουργούμε ένα πίνακα με την συχνότητα των διαφορετικών όρων που υπάρχουν στα κείμενα που θα αναλύσουμε.

```
matrix= create_matrix(tweets, language="english", removeStopwords=TRUE,
removeNumbers=TRUE,stemWords=FALSE, removePunctuation = TRUE,
tm::weightTfIdf)
```

Η εντολή μας δίνει τη δυνατότητα, αφού ορίσουμε την γλώσσα των κειμένων να επεξεργαστούμε τα κείμενα με τα ακόλουθα ορίσματα μεταξύ άλλων:

- να αφαιρέσουμε τις Stop words που αποτελούν τις κοινές λέξεις που χρησιμοποιούνται δομικά σε ένα κείμενο και η παρουσία τους δεν προσφέρει πληροφορία για το διαχωρισμό των κειμένων. Μερικές τέτοιες λέξεις στην αγγλική γλώσσα είναι : the, is, at, which, and, on κλπ.
- Να αφαιρέσουμε χαρακτήρες αριθμών
- Να αφαιρέσουμε τα σημεία στίξης
- Να περικόψουμε τις λέξεις στις ρίζες τους που θεωρητικά αποτελούν συμπυκνωμένη πληροφορία και γενικεύουν αντί να εξειδικεύουν. Για παράδειγμα οι λέξεις fishing, fished, fisher μετατρέπονται σε fish θεωρώντας ότι εμπεριέχει ένα μεγάλο κομμάτι της πληροφορίας η ρίζα fish αρκετό για την ανάλυση συναισθήματος.
- Να χρησιμοποιήσουμε σαν όρους τις ίδιες τις λέξεις ή φράσεις των 2 λέξεων ή των τριών κλπ.
- Μέσω της βιβλιοθήκης tm μας παρέχει 2 επιλογές για την εκπροσώπηση κάθε όρου (λέξης-διαφορετικού συνδυασμού χαρακτήρων ή αλλιώς n-gramm) στον πίνακα. Η πρώτη επιλογή είναι ***tf*** που σημαίνει term frequency και αποτυπώνει την συχνότητα του κάθε όρου. Η δεύτερη είναι ***tfidf*** που σημαίνει term frequency-inverse document frequency και αποτυπώνει την συχνότητα εμφάνισης του όρου αλλά προσαρμοσμένη σε σχέση με την εμφάνιση του όρου στο σύνολο των κειμένων. Μαθηματικά ορίζεται ως:

$$tf(t,d) = 0,5 + 0,5\frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

$$idf(t,D) = \log\frac{N}{1+|\{d \in D : t \in d\}|}$$

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

Όπου N είναι ο συνολικός αριθμός κειμένων (tweets) ,με t συμβολίζετε τον όρο και d το κείμενο. $|\{d \in D : t \in d\}|$ είναι ο αριθμός των κειμένων που εμφανίζεται ο όρος (προσθέτουμε 1 για να αποφύγουμε την διαίρεση με το 0).

- Επίσης ένα σημαντικό όρισμα που συμπεριλήφθηκε αργότερα στην εντολή είναι το originalmatrix το οποίο είναι απαραίτητο προκειμένου να χρησιμοποιήσουμε τους αλγόριθμους κατηγοριοποίησης σε ένα νέο σετ δεδομένων των οποίων τα χαρακτηριστικά του βασικού πίνακα δεν θα ταυτίζονται. Όταν δηλαδή ο πίνακας βασικών χαρακτηριστικών που δημιουργείτε κατά την εκπαίδευση δεν έχει λάβει υπόψη του τα χαρακτηριστικά των δεδομένων ελέγχου αλλά η κατηγοριοποίηση εκμεταλλεύεται μόνο τα κοινά χαρακτηριστικά μεταξύ των δύο σετ.

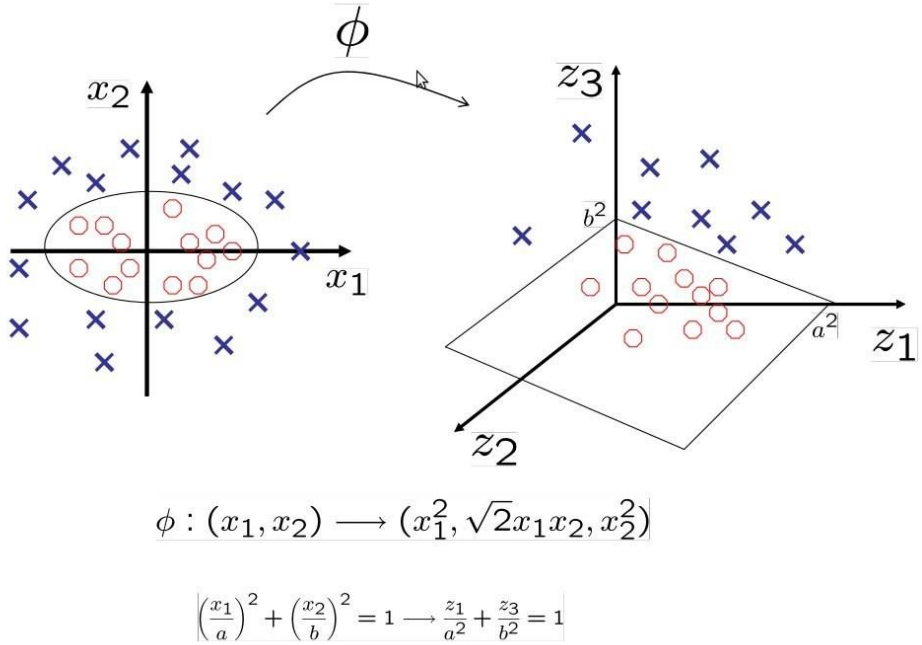
Ο πίνακας που προκύπτει με την παραπάνω εντολή συνενώνεται με τον πίνακα της Φάσης ΙΙΙ και τροφοδοτούνται στους αλγόριθμους για εκπαίδευση. Η εκπαίδευση είναι η V φάση του συστήματος.

Φάση V: Αλγόριθμοι κατηγοριοποίησης, εκπαίδευση και εφαρμογή τους.

Για την ανάλυση συναισθήματος κατά καιρούς έχουν εφαρμοστεί πάρα πολλοί αλγόριθμοι και μεθοδολογίες. Οι αλγόριθμοι που αναφέραμε στο κεφάλαιο 2 είναι ένα μέρος του συνόλου που έχουν χρησιμοποιηθεί και δοκιμαστεί. Στον τομέα του supervised learning οι πιο δημοφιλείς είναι σίγουρα οι SVM, Naïve Bayes και Maximum Entropy με τις διάφορες παραλλαγές τους.

SVM (Support Vector Machines)

Ο αλγόριθμος SVM (**Vapnik, 1979)** δίνει συχνά τα καλύτερα αποτελέσματα από τους 3. Είναι ένας αλγόριθμος, όπως αναφέραμε και στο δεύτερο κεφάλαιο, που δεν βασίζεται σε πιθανοθεωρητικά μοντέλα αλλά κατηγοριοποιεί με βάση την καλύτερη δυνατή διαχώριση των δεδομένων σε διάφορα υπερεπίπεδα. Για παράδειγμα στο σχήμα που ακολουθεί **[27]** η μέθοδος χρησιμοποιεί μια Kernel function για να προβάλλει τα δεδομένα σε μεγαλύτερες διαστάσεις ερευνώντας για την καλύτερη κατηγοριοποίηση.



Π18. Παράδειγμα κατηγοριοποίησης με SVM

Η βιβλιοθήκη **RtextTools** της R μας δίνει την δυνατότητα να χρησιμοποιήσουμε τον αλγόριθμο αξιοποιώντας την βιβλιοθήκη της R **e1071** (Meyer et al., 2012). Η συγκεκριμένη βιβλιοθήκη μας δίνει την δυνατότητα να επιλέξουμε μεταξύ διαφορετικών Kurnel functions όπως: linear, polynomial, radial basis κλπ.

Στο σύστημα μας εφαρμόζουμε τον συγκεκριμένο αλγόριθμο με radial kernel function. Ο μαθηματικός τύπος του radial kernel function για 2 διαφορετικές ομάδες όπου ο αριθμητής εκφράζει την τετραγωνική ευκλείδεια απόσταση και το σ είναι ελεύθερη παράμετρος είναι:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

3.4 Ανάλυση Πιθανοθεωρητικών Μοντέλων

Στην φάση αυτή κρίνεται σκόπιμο να εστιάσουμε στους αλγόριθμους που στηρίζονται σε πιθανοθεωρητικά μοντέλα.

Τα μοντέλα αυτά είναι το Naïve Bayes και το Maximum entropy και ανήκουν σε 2 διαφορετικές οικογένειες. Το Naïve Bayes μοντέλο ανήκει στην κατηγορία των Generative (Γενικευτικά) Μοντέλων και το μοντέλο Maximum Entropy ανήκει στην κατηγορία των Discriminative (Διαχωριστικών) μοντέλων.

Generative Models-Μοντέλα Γενίκευσης Naïve Bayes

Μοντέλα Γενίκευσης είναι τα μοντέλα τα οποία κατηγοριοποιούν τα δεδομένα αφού πρώτα έχουν προσπαθήσει στο στάδιο της εκπαίδευσης να «κατανοήσουν» τις κατανομές που γεννούν τις διαφορετικές ομάδες-κατηγορίες. Σε αυτά ανήκει και το μοντέλο Naïve Bayes. Στην περίπτωση που εξετάζουμε την κατηγοριοποίηση του κειμένου με βάση την συναισθηματική φόρτιση του:

Το μοντέλο Naïve Bayes για να καταλήξει στην πιθανότητα $P(C=c/d)$ όπου c είναι η κλάση και d είναι το μεμονωμένο κείμενο (π.χ. Tweet) λαμβάνει υπόψη του την κατανομή των κειμένων στις κατηγορίες (π.χ. αρνητικά, ουδέτερα, θετικά) $P(C=c)$ και την πιθανότητα δεδομένης της κλάσης να υπάρχει το χαρακτηριστικό f_i (λέξη, σημείο στίξης κλπ) $P\ (f_i / C=c)^{n_i(d)}$.

$$P_{NB}(c / d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i / c)^{n_i(d)})}{P(d)}$$

Όπως αναφέρεται και στο κεφάλαιο 2 ένα βασικό μειονέκτημα του συγκεκριμένου αλγόριθμου είναι η υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών που χρησιμοποιούμε προκειμένου να διαχωρίσουμε τα κείμενα στις επιμέρους ομάδες.

Το γεγονός ότι ο συγκεκριμένος αλγόριθμος προσπαθεί να κατανοήσει την φύση των ομάδων και όχι απλά να τα διαχωρίσει μπορεί να θεωρηθεί ως μειονέκτημα καθώς σκοπός μας είναι να διαχωρίσουμε με τον καλύτερο και πιο αποτελεσματικό τρόπο τα δεδομένα χρησιμοποιώντας την πληροφορία που έχουμε στην διάθεσή μας. Αν για παράδειγμα, η κατανομή των κειμένων στις επιμέρους ομάδες P(C=c) δεν είναι η ίδια στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου ο αλγόριθμος εξ ορισμού δεν θα είναι αποτελεσματικός.

Discriminative Models- Διαχωριστικά Μοντέλα Maximum Entropy

Στην κατηγορία των Discriminative models ανήκει και το μοντέλο Maximum Entropy. Η διαφορά σε σχέση με τα Μοντέλα Γενίκευσης είναι ότι δεν χρησιμοποιεί τη κατανομή των επιμέρους ομάδων για να διαχωρίσει τα δεδομένα. Ο Γενικός τύπος που παρουσιάστηκε στο δεύτερο κεφάλαιο από την εργασία των **Bo Pang, Lillian Lee και Shivakumar Vaithyanathan [6]** είναι ο παρακάτω:

$$P_{ME}(c/d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

$$F_{i,c}(d,c') := \begin{cases} 1, n_i(d) > 0 and c' = c \\ 0, otherwise \end{cases}$$

Ο Αλγόριθμος δεν κάνει υποθέσεις για την ανεξαρτησία των μοντέλων, υπολογίζει την πιθανότητα P(c/d) απ' ευθείας. Η συγκεκριμένη μαθηματική έκφραση εδώ χρησιμοποιεί την ύπαρξη των χαρακτηριστικών και όχι την συχνότητα τους αλλά αυτό δεν είναι απαραίτητο. Με τον όρο λ_i δίνεται η δυνατότητα να αναθέσουμε διαφορετική βαρύτητα στο κάθε χαρακτηριστικό που χρησιμοποιούμε. Τέλος ο όρος Z(d) είναι ίσος με $\sum_c \sum_i \lambda_{i,c} F_{i,c}(d,c)$ και χρησιμοποιείται προκειμένου να διασφαλίσουμε ότι οι πιθανότητες P(C=c/d) αθροίζουν στην μονάδα, δηλαδή αποτελούν κατανομή πιθανότητας.

Ο Αλγόριθμος Maximum Entropy ταυτίζεται με την Multinomial Λογιστική Παλινδρόμηση (Multinomial Logistic Regression). Τα λ_i του παραπάνω τύπου ταυτίζονται με τις παραμέτρους των χαρακτηριστικών που τροφοδοτούμε το μοντέλο. Αν χρησιμοποιήσουμε συμβολισμό διανυσμάτων και θεωρήσουμε ότι επίπεδο αναφοράς είναι το 0 ο παραπάνω τύπος μπορεί να γραφεί και ως:

$$P(C_i) = \frac{e^{a_i + \vec{\beta}_i \vec{f}}}{\sum_{j=2}^{k} e^{a_j + \vec{\beta}_j \vec{f}}}$$

Στον νέο συμβολισμό το i παριστάνει την κατηγορία (π.χ. αρνητική, ουδέτερη, θετική) και όχι το χαρακτηριστικό. Το α_i και το $\boldsymbol{\beta}_i$ είναι οι παράμετροι (το β_i είναι διάνυσμα) του μοντέλου. Ουσιαστικά πρόκειται για i-1 μοντέλα με i-1 σετ παραμέτρων για i-1 διαφορετικές κατηγορίες και μια κατηγορία ως επίπεδο αναφοράς.

Η χρήση ενός Γενικευμένου Γραμμικού μοντέλου όπως είναι το μοντέλο της multinomial λογιστικής παλινδρόμησης μας δίνει την δυνατότητα να εξετάσουμε την αλληλεπίδραση των χαρακτηριστικών εκτός από την κύρια επίδραση τους. Σε σχέση λοιπόν με τον αλγόριθμο Naïve Bayes, όχι μόνο δεν υποθέτει την ανεξαρτησία μεταξύ των χαρακτηριστικών αλλά μας δίνει την δυνατότητα να αξιοποιήσουμε τις αλληλεπιδράσεις τους προς το συμφέρον μας.

Στην ανάλυση προκειμένου να ελέγξουμε την αξία των αλληλεπιδράσεων στην κατηγοριοποίηση θα χρησιμοποιήσουμε μόνο τον αλγόριθμο Maximum Entropy που περιλαμβάνει η βιβλιοθήκη maxent (Jurka, 2012) και ενσωματώνεται στην βιβλιοθήκη **RtextTools.**

Με τις παρακάτω εντολές μετατρέπουμε τους πίνακες σε container. Ένα αντικείμενο που αποθηκεύει τα κείμενα διαιρώντας τα στις επιμέρους αλληλουχίες χαρακτήρων (n grammes) και τους αντιστοιχεί την κατηγορία όπου ανήκουν. Στην μορφή αυτή μπορούμε να εφαρμόσουμε τους αλγόριθμους κατηγοριοποίησης.

```
container = create_container(mega_matrix,
as.numeric(as.factor(dataset_sentiment$sentiment)),trainSize=1:NofTrainData,
testSize=(NofTrainData+1):NofData,virgin=FALSE)
models = train_models(container, algorithms=c("MAXENT" , "SVM"), kernel =
"radial", l1_regularizer = 0, l2_regularizer = 0, use_sgd = FALSE,
set_heldout = 0, verbose = FALSE)
results = classify_models(container, models)
```

Όπου NofTrainData έχουμε ορίσει το πλήθος των κειμένων που έχουμε επιλέξει ενώ τα υπόλοιπα κείμενα θα χρησιμοποιηθούν για έλεγχο.

Η πρώτη εντολή δημιουργεί ένα container μέσω της βιβλιοθήκης **tm** προκειμένου να αποθηκεύσει τον πίνακα των χαρακτηριστικών. Εντός της εντολής ορίζεται ποια κείμενα θα χρησιμοποιηθούν για εκπαίδευση και ποια για έλεγχο καθώς και οι κατηγορίες που αντιστοιχούν στο κάθε κείμενο. Η δεύτερη εντολή ορίζει τα μοντέλα που θα

χρησιμοποιηθούν στα δεδομένα και κατηγοριοποιεί τα κείμενα. Στα μοντέλα που επιλέγουμε μας δίνετε και η δυνατότητα να τα τροποποιήσουμε αξιοποιώντας ορίσματα που σχετίζονται π.χ. με τον Kernel του SVM ή regulizer του MAXENT.

Η τελευταία εντολή αποθηκεύει τα αποτελέσματα. Μετά από διαμόρφωση, τα αποτελέσματα παρουσιάζονται από το σύστημα με την παρακάτω μορφή.

```
Total number of tweets: 6773
Total number of Tweets used for training: 5418
Total number of Tweets tested: 1355
The accuracy of the SVM classifier is: 0.8457565
The accuracy of the maxentropy classifier is: 0.7911439

Svm results table

                predicted N     predicted Neutral     predicted P
actual N          1120                8                  4
actual Neutral    128                 16                 2
actual P          65                  2                  10


maxentropy results table

                predicted N     predicted Neutral     predicted P
actual N          1041                66                 25
actual Neutral    127                 12                 7
actual P          53                  5                  19


n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
n >= 1          1.00          0.79
n >= 2          0.91          0.85


ALGORITHM PERFORMANCE

SVM_PRECISION       SVM_RECALL       SVM_FSCORE
      0.6966667         0.4100000         0.4366667

MAXENTROPY_PRECISION   MAXENTROPY_RECALL   MAXENTROPY_FSCORE
                0.4533333           0.4166667           0.4266667
```

Π19. Μορφή αποτελεσμάτων της εφαρμογής

Δημιουργείται ένας πίνακας διπλής εισόδου για τον κάθε αλγόριθμο όπου αντιπαραβάλλονται οι κατηγορίες που αναθέτει ο αλγόριθμος στα κείμενα, με τις πραγματικές κατηγορίες που ανήκουν τα δεδομένα. Έχουν χρησιμοποιηθεί 5418 tweets για να εκπαιδεύσουν τον αλγόριθμο και 1355 για να τον ελέγξουν.

Η ακρίβεια (accuracy) του κάθε αλγόριθμου προκύπτει από την διαίρεση των σωστών καταχωρήσεων δια των συνολικών. Είναι το άθροισμα της διαγωνίου του πίνακα δια το πλήθος των κειμένων ελέγχου.

$$Accuracy = \frac{Correct}{Total}$$

Ως Precision, ορίζεται το ποσοστό των σωστά καταχωρημένων κειμένων εντός μιας κατηγορίας δια του συνόλου της κατηγορίας που κατέληξε ο αλγόριθμος. Για την κατηγορία των θετικών κειμένων του Maxentropy είναι:

$$precision = \frac{true\_positive}{(true\_positive + false\_positive)} = \frac{19}{(19+25+7)} = 0.3725$$

Με όρους πιθανοτήτων μπορούμε να διατυπώσουμε τα παραπάνω ως εξής : Δεδομένου ότι έχει καταχωρηθεί στα θετικά κείμενα από τον αλγόριθμο, είναι η πιθανότητα να είναι θετικό στην πραγματικότητα. Τιμές κοντά στην μονάδα μας δείχνουν ότι ο αλγόριθμος επιστρέφει περισσότερα σωστά αποτελέσματα στην κατηγορά παρά λανθασμένα.

Ως Recall (η αλλιώς sensitivity) ορίζεται το ποσοστό των σωστά καταχωρημένων κειμένων εντός των πραγματικών ομάδων κατηγοριοποίησης. Για την κατηγορία των θετικών κειμένων του αλγόριθμου Maxentropy είναι:

$$recall = \frac{true\_positive}{(true\_positive + false\_non\_positive)} = \frac{19}{(19+53+5)} = 0.2468$$

Με όρους πιθανοτήτων ορίζεται ως: Δεδομένου ότι το κείμενο είναι θετικό στην πραγματικότητα είναι το ποσοστό της σωστής κατηγοριοποίησης. Επίσης θα πρέπει να πλησιάζει την μονάδα.

Το F-score ορίζεται μαθηματικά με τον παρακάτω τύπο:

$$F_{score} = \frac{precision * recall}{precision + recall}$$

Και μπορεί να ερμηνευθεί ως ένας σταθμισμένος μέσος όρος του Precision και του Recall.

Τα ποσοστά που εμφανίζονται από τον αλγόριθμο είναι οι μέσοι όροι των κατηγοριών μιας και πρόκειται για πίνακα 3x3 κα όχι 2x2.

3.5 Βοηθητικά συστήματα

Όπως αναφέρθηκε και παραπάνω η πλατφόρμα Twitter μέσω του Twitter API (Application Programming Interface) παρέχει την δυνατότητα σύνδεσης της με άλλες εφαρμογές όπως είναι η R.

Για να συνδέσουμε οποιαδήποτε εφαρμογή με το Twitter πρέπει πρώτα να εγγραφούμε στο αντίστοιχο site https://apps.twitter.com/ προκειμένου να αποκτήσουμε κωδικούς πρόσβασης και την εξουσιοδότηση.
Για να συνδέσουμε την R με την πλατφόρμα του twitter ο πιο εύκολος τρόπος είναι να αξιοποιήσουμε την βιβλιοθήκη twitteR.

```
#handshake twitter#
consumer_key<-"nbJ59GWA5bgi5Ah4……."
consumer_secret<-"XeKvS28W8eG83GfEwkF5zEuVVlhz2T0……r"
access_token_secret<-"glTujsXaLkLlBVQeg5WHd82s6VEdbD……..4"
access_token<-"3352843239-qITgFcV8UFjTDg2rNhKTrNdO2tp1t9z……i"
setup_twitter_oauth(consumer_key,consumer_secret,access_token,access_token_secret)
```

Στον παραπάνω κώδικα αντικαθιστούμε τους κωδικούς πρόσβασης που μας παρέχει το Twitter και τρέχοντας τον κώδικα ζητούμε έγκριση. Εφόσον όλα είναι σωστά έχουμε συνδεθεί επιτυχημένα με την εφαρμογή του Twitter, μπορούμε να δημοσιεύσουμε κείμενα, να επικοινωνήσουμε με άλλους χρήστες και γενικά να διαχειριστούμε την εφαρμογή μέσω της R. Στην προκειμένη περίπτωση αυτό που μας ενδιαφέρει είναι η πρόσβαση σε κείμενα-δημοσιεύσεις ανάλογα με κάποια κριτήρια που ορίζουμε εμείς και φυσικά μας επιτρέπει η Twitter API. Αυτό γίνεται εφικτό με την εντολή που ακολουθεί:

```
some_tweets<-searchTwitter(searchterm, n=sample, lang=language, since=NULL, until=NULL,locale=NULL, geocode=NULL, sinceID=NULL, maxID=NULL,resultType=NULL, retryOnRateLimit=120)
```

Με την συγκεκριμένη εντολή ζητούμε από το Twitter να μας φέρει κείμενα που περιλαμβάνουν συγκεκριμένη λέξη (searchterm) ή λέξεις που ενώνονται με + όπως πχ Greece+Athens. Επίσης ορίζουμε το μέγεθος του δείγματος που επιθυμούμε (n=) Με όριο τα

1500 ημερησίως. Με το όρισμα lang καθορίζουμε την γλώσσα που επιθυμούμε να είναι γραμμένα τα κείμενα. Με τα ορίσματα since Και until Θέτουμε χρονικά όρια στις δημοσιεύσεις. Με το όρισμα geocode θέτουμε χωρικά πλαίσια χρησιμοποιώντας συντεταγμένες. Τέλος με τα ορίσματα sinceID και maxID χρησιμοποιούμε τον αριθμό-κωδικό συγκεκριμένης δημοσίευσης για να καθορίσουμε χρονικά όρια.

Το αποτέλεσμα της εντολής είναι η εισαγωγή των tweets σε μορφή λίστας. Η κάθε δημοσίευση περιλαμβάνει εκτός από το κείμενο και άλλες επιμέρους πληροφορίες, όπως γεωγραφικά δεδομένα, το όνομα του χρήστη, χρόνο δημοσίευσης, και άλλα δεδομένα.

Για να χρησιμοποιήσουμε στην συγκεκριμένη εργασία τα σύνολα δεδομένων του SEMEVAL 2015 που παρέχονταν σε μορφή κώδικα sql αφού εγκαταστήσαμε την MySQL και το Interface της εφαρμογής Workbench συνδέσαμε την R μέσω της βιβλιοθήκης RMySQL.

Η σύνδεση της R έγινε με την παρακάτω εντολή όπου χρησιμοποιούμε τα διαπιστευτήρια της MySQL και το όνομα της βάσης δεδομένων.

```
mydb = dbConnect(MySQL(), user='root', password='…', dbname='Sentifeed', host='localhost')
```

Με τις παρακάτω εντολές διαχειριζόμαστε τους πίνακες που έχουμε αποθηκεύσει στην βάση.

```
dbListTables(mydb) # show tables of database#
dbListFields(mydb, 'classificationreport')  # show variables of table .....#

rs = dbSendQuery(mydb, "SELECT * FROM tweettestdata;") # select* #
sql_data = fetch(rs, n=-1)
```

Συγκεκριμένα με την εντολή dbListTables(mydb) εμφανίζουμε όλους τους πίνακες που περιλαμβάνει η βάση ενώ με την dbListFields(mydb, 'classificationreport') εμφανίζουμε τις μεταβλητές που περιλαμβάνει ο πίνακας “classification report”.

Τέλος με την εντολή dbSendQuery γράφουμε sql queries μέσω της R προκειμένου να ανασύρουμε τα δεδομένα που θέλουμε ή να τα τροποποιήσουμε. Η τελευταία γραμμή κώδικά φέρνει πλέον στην R τον πίνακα που έχουμε επιλέξει με την προηγούμενη εντολή.

Κ Ε Φ Α Λ Α Ι Ο 4

4.1 Αποτελέσματα

Έχοντας ολοκληρώσει την περιγραφή του συστήματος που δημιουργήθηκε για την ανάλυση κλίματος ή ανάλυση συναισθήματος σε δεδομένα twitter συνεχίζουμε με την παρουσίαση των πειραμάτων και των αποτελεσμάτων που καταλήξαμε.

Τα πειράματα αυτά χωρίζονται σε 3 διαφορετικές φάσεις κατά τις οποίες διαφοροποιούνται οι συνθήκες των πειραμάτων σε σχέση με την δημιουργία του πίνακα χαρακτηριστικών και τις κατηγορίες συναισθηματικού προσανατολισμού.

Φάση Ι: κοινός πίνακας χαρακτηριστικών για δεδομένα εκπαίδευσης και ελέγχου. Τρείς κατηγορίες συναισθηματικού προσανατολισμού.

Σε πρώτη φάση το σύστημα δοκιμάζεται σε ένα σύνολο δεδομένων 8529 tweets. Το συγκεκριμένο σετ είχε χρησιμοποιηθεί και από την Μαρία Καρανάσου και την ομάδα DSUNIPI προκειμένου να γίνουν τα πρώτα πειράματα, οι απαραίτητες τροποποιήσεις και δοκιμές σε πρώτη φάση αλλά και σαν δεδομένα εκπαίδευσης κατά την τελική δοκιμή. Τα κείμενα αυτά, όπως αναφέραμε και παραπάνω περιλαμβάνουν έντονο μεταφορικό-αλληγορικό λόγο αυξάνοντας τον βαθμό δυσκολίας για ένα αυτοματοποιημένο μηχανισμό ανάλυσης.

Τα κείμενα έχουν χωριστεί σε 3 κατηγορίες ανάλογα με το περιεχόμενό τους. Αρνητικά, Ουδέτερα και Θετικά. Το σύστημα που δημιουργήσαμε καλείται να διαχωρίσει τα δεδομένα σωστά στις 3 ομάδες. Για να επιτύχουμε αυτό τον σκοπό και αφού έχουμε καταλήξει στα χαρακτηριστικά που περιγράψαμε στο προηγούμενο κεφάλαιο δοκιμάζουμε 4 διαφορετικές προσεγγίσεις που προκύπτουν από τους συνδυασμούς των παρακάτω στοιχείων.

- Την παρουσία ή μη στο μοντέλο των αλληλεπιδράσεων δευτέρου βαθμού των επιμέρους χαρακτηριστικών που τροφοδοτούμε το σύστημα όπως είναι η παρουσία Κεφαλαίων γραμμάτων, οι αρνήσεις, τα emoticons, η συχνότητα των μερών του λόγου (Pos Tags) κλπ.

- Την τροφοδότηση του πίνακα των κυρίως χαρακτηριστικών με κείμενα στην αρχική τους μορφή ή με κείμενα στα οποία έχουμε επισημάνει τα μέρη του λόγου που τα αποτελούν.

Στον πίνακα που ακολουθεί παρουσιάζονται τα αποτελέσματα των 20 δοκιμών χρησιμοποιώντας διάφορα ποσοστά του σετ για εκπαίδευση και έλεγχο. Η δεύτερη στήλη αναφέρει την μορφή των δεδομένων που χρησιμοποιήσαμε για την ανάλυση, η Τρίτη στήλη το ποσοστό των διαφορετικών κειμένων που χρησιμοποιήθηκε για εκπαίδευση. Η πέμπτη στήλη παρουσιάζει την εξέταση η μη αλληλεπιδράσεων δευτέρου βαθμού.

Από τον πίνακα μπορούμε να παρατηρήσουμε ότι ο αλγόριθμος SVM πετυχαίνει τα καλύτερα αποτελέσματα. Αν μάλιστα τον τροφοδοτήσουμε με κείμενα POS tagged βελτιώνουμε ακόμα περισσότερο την αποτελεσματικότητά του. Από την άλλη πλευρά ο MAXENT αλγόριθμος φαίνεται να αποδίδει καλύτερα στις περιπτώσεις που το μοντέλο περιλαμβάνει αλληλεπίδραση.

Η αλληλεπίδραση φαίνεται να αυξάνει την αποτελεσματικότητα του MAXENT αλλά δεν είναι ξεκάθαρα τα οφέλη για τον SVM. Τέλος αξιοσημείωτη είναι η σχέση της αναλογίας των κειμένων που χρησιμοποιούμε για εκπαίδευση/έλεγχο με τα αποτελέσματα της κατηγοριοποίησης.

	Input	Train DataSet (%)	Test DataSet (%)	2 Way-Interractions	SVM accuracy	SVM precision	SVM recall	SVM fscore	MAXENT accuracy	MAXENT precision	MAXENT recall	MAXENT fscore
1	text	50	50	OFF	0,851	0,727	0,433	0,473	0,791	0,453	0,413	0,427
2	text	60	40	OFF	0,850	0,737	0,437	0,480	0,794	0,470	0,413	0,433
3	text	70	30	OFF	0,852	0,767	0,437	0,480	0,789	0,470	0,420	0,433
4	text	80	20	OFF	0,852	0,743	0,467	0,517	0,786	0,463	0,413	0,430
5	text	90	10	OFF	0,848	0,737	0,467	0,520	0,797	0,490	0,413	0,437
6	text	50	50	ON	0,857	0,727	0,463	0,513	0,820	0,543	0,473	0,497
7	text	60	40	ON	0,856	0,733	0,473	0,527	0,815	0,537	0,483	0,507
8	text	70	30	ON	0,854	0,730	0,470	0,523	0,819	0,560	0,497	0,520
9	text	80	20	ON	0,851	0,723	0,480	0,533	0,804	0,523	0,473	0,490
10	text	90	10	ON	0,848	0,727	0,477	0,530	0,803	0,527	0,477	0,490
11	pos tagged text	50	50	OFF	0,870	0,740	0,537	0,593	0,791	0,457	0,410	0,423
12	pos tagged text	60	40	OFF	0,872	0,767	0,540	0,603	0,792	0,457	0,410	0,423
13	pos tagged text	70	30	OFF	0,872	0,790	0,537	0,607	0,791	0,463	0,420	0,437
14	pos tagged text	80	20	OFF	0,873	0,780	0,557	0,623	0,791	0,470	0,420	0,437
15	pos tagged text	90	10	OFF	0,857	0,757	0,527	0,583	0,798	0,527	0,467	0,487
16	pos tagged text	50	50	ON	0,867	0,750	0,517	0,577	0,833	0,573	0,473	0,503
17	pos tagged text	60	40	ON	0,865	0,767	0,507	0,570	0,809	0,520	0,470	0,487
18	pos tagged text	70	30	ON	0,864	0,753	0,513	0,573	0,825	0,573	0,497	0,523
19	pos tagged text	80	20	ON	0,861	0,747	0,520	0,580	0,815	0,550	0,550	0,510
20	pos tagged text	90	10	ON	0,858	0,770	0,523	0,587	0,815	0,583	0,507	0,530

Π20. Πίνακας αποτελεσμάτων φάσης Ι

Για να τεκμηριώσουμε τα παραπάνω και στατιστικά, μπορούμε να χρησιμοποιήσουμε ένα μοντέλο λογιστικής παλινδρόμησης. Σαν μεταβλητή απόκρισης ορίζουμε την μεταβλητή Accuracy και σαν ερμηνευτικές μεταβλητές:

α) την αξιοποίηση ή μη της αλληλεπίδρασης (μεταβλητή @2Wayinterractions) μεταξύ των επιμέρους χαρακτηριστικών και

β) την τροφοδότηση του αλγόριθμου με απλό κείμενο ή με Pos tagged κείμενο (μεταβλητή input) για τον κυρίως πίνακα χαρακτηριστικών.

Τα αποτελέσματα του Wald Chi-Square Test για την στατιστική σημαντικότητα των παραγόντων στην ακρίβεια των δύο αλγορίθμων παρουσιάζονται στους παρακάτω πίνακες σε επίπεδο σημαντικότητας 5%.

Για τον αλγόριθμο SVM η αλληλεπίδραση των χαρακτηριστικών δεν επηρεάζει στατιστικά σημαντικά την ακρίβεια του, ενώ η τροφοδότηση του με Pos Tagged κείμενο επηρεάζει σημαντικά την αποτελεσματικότητα του αλγόριθμου.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	1012113,183	1	,000
@2WayInterractions	,864	1	,353
Input	68,611	1	,000
@2WayInterractions * Input	5,708	1	,017

Dependent Variable: SVM accuracy
Model: (Intercept), @2WayInterractions, Input, @2WayInterractions * Input

Π21. Έλεγχος Wald γι την σημαντικότητα των παραγόντων στο μοντέλο SVM

Τα αντίθετα αποτελέσματα εμφανίζονται για τον αλγόριθμο Maxentropy όπου ο παράγοντας αλληλεπίδραση είναι στατιστικά σημαντικός ενώ οι δυο κατηγορίες κειμένου που τροφοδοτούμε τον αλγόριθμο δεν μεταβάλλουν στατιστικά σημαντικά την ακρίβεια της κατηγοριοποίησης.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	346748,765	1	,000
@2WayInterractions	75,618	1	,000
Input	2,605	1	,106
@2WayInterractions * Input	1,146	1	,284

Dependent Variable: MAXENT accuracy
Model: (Intercept), @2WayInterractions, Input, @2WayInterractions * Input

Π22. Έλεγχος Wald γι την σημαντικότητα των παραγόντων στο μοντέλο MAXENT

Και τα 2 μοντέλα λογιστικής παλινδρόμησης διαφέρουν στατιστικά σημαντικά από το NULL Μοντέλο μόνο με τον σταθερό όρο.

Φάση ΙΙ: πίνακας χαρακτηριστικών με βάση τα δεδομένα εκπαίδευσης. Τρείς κατηγορίες συναισθηματικού προσανατολισμού.

Στα παραπάνω πειράματα δημιουργήθηκε ένας πίνακας χαρακτηριστικών ο οποίος στην συνέχεια χωρίστηκε σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου. Στην συνέχεια θα δημιουργήσουμε τον πίνακα χαρακτηριστικών μόνο από τα δεδομένα εκπαίδευσης, ο έλεγχος θα γίνει με βάση τα χαρακτηριστικά που είναι κοινά στις 2 βάσεις δεδομένων. Ως δεδομένα εκπαίδευσης θα χρησιμοποιήσουμε το σύνολο των 8529 κειμένων-tweets που

χρησιμοποιήσαμε για τα προηγούμενα πειράματα και ως δεδομένα ελέγχου θα χρησιμοποιήσουμε τα 4000 tweets που χρησιμοποιήθηκαν σαν τελική δοκιμασία στον διαγωνισμό Semeval 2015. Η βαρύτητα του επιμέρους πίνακα θα είναι μεγαλύτερη από πριν σε αυτά τα πειράματα αλλά η γενικότερη αποτελεσματικότητα θα είναι χαμηλότερη.

Από τα 8529 δεδομένα εκπαίδευσης, μετά την πρώτη φάση της λειτουργίας του συστήματος παραμένουν 8466 μοναδικά κείμενα για εκπαίδευση. Επίσης αξιοποιώντας την δυνατότητα συντονισμού των παραμέτρων του MAXENT (tune) θέτουμε το όρισμα «l2_regularizer = 1» για να βελτιώσουμε την απόδοση του αλγόριθμου.
Τα αποτελέσματα των δοκιμών παρουσιάζονται στον παρακάτω πίνακα.

	Input	2 Way-Interractions	SVM accuracy	SVM precision	SVM recall	SVM fscore	MAXENT accuracy	MAXENT precision	MAXENT recall	MAXENT fscore
1	text	OFF	0,705	0,453	0,480	0,417	0,672	0,443	0,450	0,443
2	text	ON	0,717	0,423	0,410	0,417	0,680	0,443	0,447	0,443
3	pos tagged text	OFF	0,687	0,443	0,477	0,453	0,666	0,450	0,467	0,450
4	pos tagged text	ON	0,717	0,443	0,460	0,450	0,665	0,443	0,460	0,447

Π23. Πίνακας αποτελεσμάτων φάσης ΙΙ

Όπως παρατηρούμε στον πίνακα ο αλγόριθμος SVM πετυχαίνει ξανά, καλύτερη κατηγοριοποίηση σε σχέση τον MAXENT αλγόριθμο. Η ακρίβεια του SVM κυμαίνονται γύρω από το 71% και του MAXENT γύρω από το 67%. Την καλύτερη απόδοση την πετυχαίνει ο MAXENT όταν ο κύριος πίνακας χαρακτηριστικών τροφοδοτείται με απλό κείμενο και στον επιμέρους πίνακα λαμβάνουμε υπόψη την αλληλεπίδραση των χαρακτηριστικών. Φαίνεται ότι η αλληλεπίδραση βοηθάει και την κατηγοριοποίηση του SVM.

Φάση ΙΙΙ: πίνακας χαρακτηριστικών με βάση τα δεδομένα εκπαίδευσης. Έντεκα κατηγορίες συναισθηματικού προσανατολισμού.

Τέλος για λόγους σύγκρισης θα επαναλάβουμε την παραπάνω δοκιμή διαχωρίζοντας αυτή την φορά τα δεδομένα σε 11 ομάδες (αντί τριών)! Η κατηγοριοποίηση γίνεται με βάση το αρχικό σκορ που τους έχει ανατεθεί και εκφράζει διαφορετική ένταση του συναισθήματος. Με βάση αυτό το σκορ έγινε η εκπαίδευση αλλά και η τελική δοκιμασία των συμμετεχόντων στο Semeval 2015. Στον πίνακα που ακολουθεί παρουσιάζεται η κατανομή των δεδομένων εκπαίδευσης και ελέγχου στις διάφορες κατηγορίες.

	-5	-4	-3	-2	-1	0	1	2	3	4	5	Total
Train Data	4	403	2907	2895	855	892	196	202	122	50	3	8529
Test Data	4	100	737	1541	680	298	169	155	201	111	4	4000

Π24. Πίνακας κατανομής δεδομένων της φάσης ΙΙΙ

Για το συγκεκριμένο πείραμα χρησιμοποιούμε τις ρυθμίσεις που στις προηγούμενες δοκιμές (φάση 2) απέφεραν την υψηλότερη ακρίβεια (accuracy). Σαν πρώτη ύλη θα χρησιμοποιηθούν τα Pos Tagged κείμενα και στον επιμέρους πίνακα θα αξιοποιήσουμε την αλληλεπίδραση των

χαρακτηριστικών. Τα αποτελέσματα είναι κοντά στα αντίστοιχα της Μαρίας Καρανάσου. Ο αλγόριθμος SVM κατηγοριοποίησε σωστά το 23,5% (29% στο σύστημα της Μαρίας Καρανάσου) και ο MAXENT κατηγοριοποίησε σωστά το 21,9% των κειμένων.

Input	2 Way-Interractions	SVM accuracy	SVM precision	SVM recall	SVM fscore	MAXENT accuracy	MAXENT precision	MAXENT recall	MAXENT fscore
pos tagged text	ON	0,235	0,060	0,110	0,053	0,219	0,098	0,115	0,080

Π25. Αποτελέσματα Φάσης ΙΙΙ

Τέλος προκειμένου να αποτυπώσουμε την συνεισφορά των επιμέρους χαρακτηριστικών στην σωστή κατηγοριοποίηση εφαρμόζουμε ένα τελευταίο πείραμα. Εφαρμόζουμε πολυωνυμική λογιστική παλινδρόμηση χρησιμοποιώντας μόνο τον επιμέρους πίνακα (κεφάλαιο 3 φάση ΙΙΙ) με τα χαρακτηριστικά που έχουμε απομονώσει και τις αλληλεπιδράσεις δευτέρου βαθμού. Για το πείραμα χρησιμοποιούμε το σετ δεδομένων των 8529 tweets (αφού αφαιρέσουμε τις διπλές παρατηρήσεις).

Πρέπει να σημειώσουμε, ότι τα αποτελέσματα είναι ενδεικτικά, καθώς η συνεισφορά του επιμέρους πίνακα συμπεριλαμβανομένων και των αλληλεπιδράσεων είναι κατά πολύ μικρότερη από την αντίστοιχη του κυρίως πίνακα, έχουμε εξαναγκάσει όλες τις κύριες επιδράσεις και αλληλεπιδράσεις στο μοντέλο και ο αλγόριθμος της R δεν συγκλίνει.

Σαν κατηγορία αναφοράς ορίζουμε τα αρνητικά κείμενα. Στον πίνακα που ακολουθεί παρουσιάζονται οι συντελεστές των κύριων επιδράσεων και το p-value του Wald test.

Όσον αφορά τις κύριες επιδράσεις παρατηρούμε ότι τα Pos Tags στην πλειοψηφία τους εμφανίζονται στατιστικά σημαντικά. Μάλιστα η χρήση κύριων και κοινών ουσιαστικών **NNS** και ρημάτων στο 3 πρόσωπο **VBZ** διαχωρίζει (στατιστικά σημαντικά) τα ουδέτερα κείμενα (αντικειμενικά) από τα αρνητικά (υποκειμενικά). Χρήσιμες επίσης για τον διαχωρισμό είναι οι μεταβλητές: **reply**, **hastag** και **negation**.

Τα κείμενα που δημοσιεύονται ως απάντηση σε άλλα κείμενα ή τα αναφέρουν (reply) συνδέονται περισσότερο με ουδέτερο περιεχόμενο παρά με αρνητικό περιεχόμενο. Αντίθετα τα κείμενα που περιλαμβάνουν hastag*, άρνηση είναι πιθανότερο να έχουν αρνητικό περιεχόμενο παρά ουδέτερο.

Τα θετικά κείμενα διαχωρίζονται από τα αρνητικά με τα **RB** (ρήματα στο παρόν), **JJ** (επίθετα) και **VB** (επιρρήματα) με θετικό πρόσημο συντελεστή. Δηλαδή όσο αυξάνεται η συχνότητα των συγκεκριμένων οικογενειών λέξεων αυξάνεται η πιθανότητα να είναι θετικά τα κείμενα . Πέρα από τα μέρη του λόγου, στατιστικά σημαντικές επιδράσεις παρουσιάζουν οι μεταβλητές reply, hastag*, URL και Capital. Όπως και για τα ουδέτερα κείμενα έτσι και για τα θετικά φορτισμένα η μεταβλητή reply έχει θετικό συντελεστή, όπως και η URL επιβεβαιώνοντας τους Go et al (2009) [**18**]. Η μεταβλητή hastag όπως και πριν έχει αρνητικό πρόσημο κατά συνέπεια υποδηλώνει αρνητικό συναίσθημα αντί για θετικό.

	(Intercept)	NN	NNS	RB	JJ	VBG	VBP	DT	VB	VBN	VBZ	CD	JJS	CC	`!`	`%`
neutral coefficient	-1,060	-0,220	-0,635	-0,375	####	-1,241	-0,709	-1,259	0,033	-0,487	-2,022	-0,579	1,150	-0,581	0,008	-1,141
positive coefficient	-3,639	0,140	-0,167	0,632	0,724	0,174	-0,550	-0,570	0,866	-0,495	0,846	-0,362	-0,748	0,362	0,095	-0,292
neutral significance	0,038	0,007	0,010	0,055	0,001	0,002	0,073	0,001	0,887	0,342	0,003	0,580	0,400	0,682	0,877	0,288
positive significance	0,000	0,095	0,522	0,001	0,001	0,646	0,179	0,197	0,000	0,369	0,160	0,750	0,661	0,770	0,041	0,654

	`&`	`;`	`?`	`@`	`-`	`*`	smile emoticons	frown emoticons	laughing	reply	RT	hastag	URL	Capital	negation	consec. letters
neutral coefficient	0,096	0,621	0,108	-0,145	0,006	0,112	-1,600	-0,569	0,768	2,364	0,837	-2,278	-1,535	0,011	-2,442	0,668
positive coefficient	0,176	-0,282	-0,210	0,049	0,093	-0,163	-1,052	-0,173	0,725	1,447	0,743	-1,849	1,071	0,019	-0,359	0,712
neutral significance	0,550	0,035	0,235	0,203	0,955	0,628	0,536	0,897	0,357	0,000	0,081	0,000	0,155	0,088	0,000	0,241
positive significance	0,393	0,598	0,089	0,586	0,364	0,538	0,580	0,916	0,416	0,001	0,105	0,000	0,015	0,002	0,539	0,237

Π26. Πίνακας κύριων επιδράσεων

*Το συγκεκριμένο σετ δεδομένων έχει επιλεγεί κατά μεγάλο ποσοστό από κείμενα που περιέχουν συγκεκριμένα hastags όπως #sarcasm κλπ και συνδέονται εξ ορισμού κυρίως με αρνητικά σχόλια.

Αξίζει να σημειώσουμε ότι τα emoticons(smile emoticons, frown emoticons) και η ύπαρξη χαρακτήρων που υποδηλώνουν γέλιο (laughing) δεν είναι στατιστικά σημαντικά γεγονός που μπορεί να οφείλεται στην περιορισμένη εμφάνισή τους στα δεδομένα.

Στον επόμενο πίνακα παρουσιάζονται οι αλληλεπιδράσεις των παραγόντων που βρέθηκαν στατιστικά σημαντικές. Για την ουδέτερη κατηγορία ξεχωρίζουν οι αλληλεπιδράσεις των PosTags με τις μεταβλητές hastag, negation reply. Για τα κείμενα που αποτελούν απαντήσεις σε άλλες δημοσιεύσεις (reply) η αυξημένη συχνότητα ουσιαστικών NN χαρακτηρίζει περισσότερο το αρνητικό περιεχόμενο. Αντίστοιχα η συνύπαρξη hastag και αυξημένης συχνότητας κάποιων POS tags όπως τα ουσιαστικά NN, συνδέονται με τα ουδέτερα κείμενα. Η συνύπαρξη ρημάτων κα άρνησης (VB:negation) συνδέεται με τα ουδέτερα κείμενα. Ενδιαφέρον επίσης παρουσιάζουν οι διαφοροποιήσεις των κειμένων που ανήκουν στην κατηγορία reply όταν περιέχουν αυξημένη συχνότητα ρημάτων VB και σχετίζονται περισσότερο με τα αρνητικά παρά με τα ουδέτερα. Στατιστικά σημαντική είναι η αλληλεπίδραση των θετικών emoticons με τις αρνήσεις που παραπέμπουν σε ουδέτερο παρά σε αρνητικό κείμενο.

Για την κατηγορία των θετικών κειμένων στατιστικά σημαντικές είναι πάλι οι αλληλεπιδράσεις της μεταβλητής hastag με την RB που συμβολίζει την συχνότητα των επιρρημάτων και συνδέει το κείμενο με αρνητικό συναίσθημα παρά με θετικό. Επίσης η συνύπαρξη θετικού emoticon με hastag συνδέεται έντονα με τα αρνητικά κείμενα ανιχνεύοντας έτσι σαρκασμό-ειρωνεία.

	`NN:reply	`NN:hastag	`RB:hastag	`RB:negation	`JJ:VBG`	`JJ:DT`	`JJ:VB`	`VBP:VB`	`VBG:VB`	`VBG:hastag	`DT:CC`
neutral coefficient	-0,184	0,265	-0,234	0,307	0,243	0,178	0,001	0,137	0,081	0,590	0,408
positive coefficient	-0,097	0,121	-0,375	-0,146	0,089	0,053	-0,148	0,035	-0,153	0,368	0,952
neutral significance	0,000	0,000	0,070	0,007	0,006	0,039	0,984	0,035	0,134	0,010	0,331
positive significance	0,048	0,030	0,001	0,248	0,382	0,617	0,042	0,643	0,028	0,072	0,022

	`VB:VBZ`	`VB:reply	`VB:negation	`VBN:hastag	`VBN:consec. letters`	`CD:hastag	`smile emoticons:hastag	`smile emoticons: negation	`laughing: hastag	`DT:hastag
neutral coefficient	0,259	-0,264	0,306	0,723	-0,036	-1,669	-0,371	2,617	-0,520	-0,105
positive coefficient	-0,076	-0,128	-0,259	0,226	-0,843	-0,961	-2,795	0,512	-1,188	0,480
neutral significance	0,027	0,031	0,022	0,013	0,894	0,001	0,818	0,015	0,328	0,630
positive significance	0,611	0,314	0,110	0,401	0,016	0,064	0,016	0,575	0,017	0,031

Π27. Πίνακας αλληλεπιδράσεων

4.2 Συμπεράσματα

Με την παρούσα εργασία παρουσιάζεται ένα αυτοτελές σύστημα κατηγοριοποίησης γραπτού κειμένου με βάση τον συναισθηματικό του προσανατολισμό. Σε αντίθεση με άλλα συστήματα που στηρίζονται σε μεγάλο βαθμό σε εξωτερικές βιβλιοθήκες απ' όπου τροφοδοτούνται με πληροφορίες όπως είναι η πολικότητα των επιμέρους λέξεων, ο βαθμός συγγένειας των λέξεων μεταξύ τους κλπ, το συγκεκριμένο σύστημα στηρίζεται σε μεγάλο βαθμό στην πληροφορία που παρέχεται από το ίδιο το κείμενο. Μόνη εξαίρεση είναι η κατηγοριοποίηση των λέξεων στα διάφορα μέρη του λόγου για την οποία μπορούμε να αξιοποιήσουμε εύκολα τις δυνατότητες που μας παρέχει η βιβλιοθήκη NLTK της Python [31] και η πολικότητα των emojis. Το κέρδος μας είναι η δυνατότητα εφαρμογής του συγκεκριμένου συστήματος σε πολλές διαφορετικές γλώσσες. Το συγκεκριμένο σύστημα μπορεί να συγκριθεί με πιο περίπλοκα συστήματα όπως αυτό της Μαρίας Καρανάσου καθώς, παρότι είναι απλό αξιοποιεί σε μεγαλύτερο βαθμό την πληροφορία που εμπεριέχεται στα ίδια τα κείμενα.

Με την συγκεκριμένη εργασία μελετήσαμε την αλληλεπίδραση των χαρακτηριστικών και προσπαθήσαμε να την αξιοποιήσουμε προς όφελος της κατηγοριοποίησης. Είναι σαφές ότι οι αλληλεπιδράσεις των χαρακτηριστικών βελτιώνουν σημαντικά την απόδοση του αλγόριθμου MAXENT ενώ για τον αλγόριθμο SVM τα αποτελέσματα είναι αντιφατικά. Στις περιπτώσεις που εφαρμόζουμε ένα ήδη εκπαιδευμένο μοντέλο σε δεδομένα ελέγχου, όπως στην δεύτερη φάση των παραπάνω πειραμάτων η αξιοποίηση των αλληλεπιδράσεων φαίνεται να συνεισφέρει στην σωστή κατηγοριοποίηση.

Προκειμένου να υπάρχει ένα μέτρο σύγκρισης για την αποτελεσματικότητα του συστήματος το συγκρίναμε με το αντίστοιχο της Μαρίας Καρανάσιου-DSUNIPI [23]. Καθώς ο αλγόριθμος που χρησιμοποιήθηκε για την τελική δοκιμή ήταν ο SVM μπορούμε να πούμε

ότι τα 2 συστήματα είναι σχετικά κοντά σε αποτελεσματικότητα τουλάχιστον σε όρους του συγκεκριμένου αλγόριθμου.

Στο τελευταίο μέρος των πειραμάτων προσπαθήσαμε να αποτυπώσουμε την συμβολή της κάθε μεταβλητής του επιμέρους πίνακα στον βαθμό που αυτό είναι εφικτό. Τα μέρη του λόγου που αποτελούν το κείμενο και συγκεκριμένα η συχνότητα τους φαίνεται να συμβάλλει σημαντικά στην σωστή κατηγοριοποίηση. Επίσης στατιστικά σημαντική είναι και η συμβολή χαρακτηριστικών που συνδέονται με το συγκεκριμένο μέσω όπως οι μεταβλητές: reply, hastag, URL ή γενικότερα με την ηλεκτρονική γραφή Capital letters. Με την μελέτη και των αλληλεπιδράσεων δευτέρου βαθμού καταφέραμε να αξιοποιήσουμε και χαρακτηριστικά που καταγράψαμε αλλά δεν εμφανίστηκαν στατιστικά σημαντικά σαν κύριες επιδράσεις. Για παράδειγμα η σχετική πιθανότητα να είναι θετικό ένα κείμενο αν συνυπάρχουν hastag και θετικό emoticon είναι 2,795 φορές μικρότερη από την σχετική πιθανότητα να είναι αρνητικό.

Η μελέτη και τα αποτελέσματα αυτά καταδεικνύουν τις δυνατότητες επιτυχίας καλύτερων αποτελεσμάτων κατηγοριοποίησης με την αξιοποίηση και τον ακριβέστερο συντονισμό των παραμέτρων των αλγορίθμων που χρησιμοποιούνται κατά κόρον στις ημέρες μας.

Προς αυτή την κατεύθυνση ένα επόμενο βήμα θα μπορούσε να είναι η βελτίωση του συστήματος ώστε να ενσωματώνει στην ανάλυση τις κύριες επιδράσεις και τις αλληλεπιδράσεις που είναι στατιστικά σημαντικές και πραγματικά βελτιώνουν το αποτέλεσμα της κατηγοριοποίησης.

Κατάλογος Πινάκων-Σχημάτων

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Federica Giummolè, Salvatore Orlando, Gabriele Tolomei: «**Trending Topics on Twitter Improve the Prediction of Google Hot Queries**»

[2] Jeff Gentry R package «**Package 'twitteR'**» (2015) https://cran.r-project.org/web/packages/twitteR

[3] Kurt Hornik [aut, cre], Johannes Rauch [aut], Christian Buchta [aut], Ingo Feinerer [aut] R package « **Textcat**»

[4] Vasileios Hatzivassiloglou και Kathleen R. McKeown 1997 « **Predicting the Semantic Orientation of Adjectives**» (**1997)**

[5] Peter D. Turney «**Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews»** (2002)

[6] Bo Pang, Lillian Lee και Shivakumar Vaithyanathan **«Thumbs up? Sentiment Classification using Machine Learning Techniques»**

*[7] Das και Chen «**Yahoo! for amazon: Sentiment extraction from small talk on the web»** (2001)*

*[8] Pimwadee Chaovalit και Lina Zhou **«Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches»** (2005)*

[9] Bo Pang, Lillian Lee «**A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts**» (2005)

[10] Ana-Maria Popescu, Oren Etzioni «**Extracting Product Features and Opinions from Reviews**» (2005)

[11] Minqing Hu and Bing Liu «**Mining and Summarizing Customer Reviews**» (2004)

[12] Alyssa Liang «**Rotten Tomatoes: Sentiment Classification in Movie Reviews**» (2006)

[13] Neil O'Hare1 , Michael Davy2 , Adam Bermingham1 , Paul Ferguson1 , Páraic Sheridan2 , Cathal Gurrin1 , Alan F. Smeaton «**Topic-Dependent Sentiment Analysis of Financial Blogs**» (2009)

[14] Edoardo Airoldi, Xue Bai, Rema Padman «**Markov Blankets and Meta-Heuristics Search: Sentiment Extraction from Unstructured Texts»**

[15] Alec Go, Richa Bhayani Lei Huang, «**Twitter Sentiment Classification using Distant Supervision**» (2009)

[16] J. Read «**Using emoticons to reduce dependency in machine learning techniques for sentiment classification**»

[17] Alexander Pak, Patrick Paroubek **«Twitter as a Corpus for Sentiment Analysis and Opinion Mining»** (2010)

[18] Go et al **«Twitter Sentiment Classification using Distant Supervision»** (2009)

[19] SidaWang Christopher D. Manning «**Baselines and Bigrams: Simple, Good Sentiment and Topic Classification**»

[20] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau «**Sentiment analysis of twitter Data**»

[21] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore «**Twitter Sentiment Analysis: The Good the Bad and the OMG!**»

[22] Petra Kralj Novak, Jasmina Smailović, Borut Sluban και Igor Mozetič «**Sentiment of Emojis**»

[23] Μαρία Καρανάσου **«Ανάλυση συναισθήματος σε Κοινωνικά δίκτυα, ο μεταφορικός λόγος στο Twitter»** (2015)

[24] Kirill Pomogajko **«Emoticons Decoder for Social media sentiment analysis in R»**

[25] Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, Wouter van Atteveldt R package « **Rtexttools**» https://cran.r-project.org/web/packages/RTextTools (2014)

[26] V Vapnik «**Support Vector Machines**» (1979)

[27] Sussex university http://users.sussex.ac.uk/~christ/crs/ml/lec08a.html

[28] Meyer et al. R package «e1071» (2015)

[29] Timothy P. Jurka, Yoshimasa Tsuruoka R Package «**maxent**» (2013)

[30] Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] R package «**tm**». (2015)

[31] Steven Bird, Edward Loper, Ewan Klein **«Natural Language Toolkit»** (2001)

[32] Νίκος Πελέκης, ***Σημειώσεις για το μάθημα «Στατιστικές Μέθοδοι Εξόρυξης Δεδομένων»****.(2015)*

[33] *Κων/νος Πολίτης. Σημειώσεις* ***για το μάθημα «Γενικευμένα Γραμμικά Μοντέλλα»*** *(2016)*

[34] Nello Cristianini John Shawe-Taylor « **Support Vector Machines and other kernel Based learning methods**» (2000)