ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Μελέτη τεχνικών και εργαλείων ανάλυσης συναισθήματος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΝΑΓΙΩΤΗ ΠΕΤΡΟΠΟΥΛΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Μελέτη τεχνικών και εργαλείων ανάλυσης συναισθήματος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΠΑΝΑΓΙΩΤΗ ΠΕΤΡΟΠΟΥΛΟΥ

Επιβλέπων : Δημήτριος Ασκούνης
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Οκτωβρίου 2015.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Βασίλειος Ασημακόπουλος
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

(Υπογραφή)

.....

ΠΑΝΑΓΙΩΤΗΣ ΠΕΤΡΟΠΟΥΛΟΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα τελευταία χρόνια, με την αυξανόμενη χρήση Web 2.0 εφαρμογών έχουν δημιουργηθεί καινούργιοι τρόποι επικοινωνίας. Τα μέσα κοινωνικής δικτύωσης περιέχουν ένα θησαυρό πληροφοριών, μεταξύ των οποίων γνώμες και συναισθήματα των χρηστών τους και ο σκοπός της Ανάλυσης Συναισθήματος είναι να βγάλει χρήσιμα συμπεράσματα από αυτά. Ειδικότερα, οι επιχειρήσεις μπορούν να αξιοποιήσουν αυτές τις πληροφορίες ώστε να γίνουν πιο ανταγωνιστικές και να βελτιώσουν το επίπεδο των παρεχόμενων προϊόντων και υπηρεσιών τους. Πιο συγκεκριμένα, η παρούσα διπλωματική εργασία ασχολείται με τις τεχνικές και τα εργαλεία της ανάλυσης συναισθήματος. Παρουσιάζονται κατηγορίες στις οποίες μπορεί να ενταχθεί η ανάλυση συναισθήματος, ως προς τις χρήσεις της, και παραδείγματα αυτών. Έπειτα, σε τεχνικό επίπεδο, παρατίθενται οι κατηγορίες σε σχέση με την προσέγγιση κειμένου, και σε σχέση με την προσέγγιση της τεχνικής που ακολουθούμε κατά την διαδικασία. Ύστερα, εξετάζονται εκτεταμένα οι κατηγορίες των τεχνικών οι οποίες είναι οι τεχνικές με λεξικά, τεχνικές με επιβλεπόμενη μηχανική μάθηση, τεχνικές μη-επιβλεπόμενη μηχανική μάθηση, υβριδικές τεχνικές. Παρουσιάζονται παραδείγματα από μελέτες πάνω στην κάθε κατηγορία και συγκεντρωτικοί πίνακες που βοηθούν στην επισκόπηση της τεχνολογίας αιχμής και εξάγονται χρήσιμα συμπεράσματα. Έπειτα, με γνώμονα τη χρησιμότητα της ανάλυσης συναισθήματος γενικά, αλλά και ειδικότερα στις επιχειρήσεις, γίνεται επισκόπηση των εφαρμογών και δίνονται παραδείγματα αυτών. Τέλος, συμπεραίνονται κάποιες χρήσιμες παρατηρήσεις για την παρούσα κατάσταση που βρίσκεται η ανάλυση συναισθήματος, αλλά και για το μέλλον αυτής.

Λέξεις Κλειδιά: <<ανάλυση συναισθήματος, κοινωνικά δίκτυα, (μη) επιβλεπόμενη μάθηση, λεξικά, μικρο-ιστολόγια, ταξινομητής>>

Abstract

In recent years, the widespread use of Internet and Web 2.0, has revolutionized the computer and communication world like nothing before, giving birth to new ways of communication and social interactivity. Social networks contain a great amount of information, including user sentiments and opinions, Sentiment Analysis helps to make useful conclusions out of them. Companies take into consideration and integrate into their business planning these conclusions, in order to become more competitive and increase their overall performance in today's market. In particular, this diploma thesis is about the various techniques and tools of sentiment analysis. Sentiment analysis is categorized in regard of its use, where examples are featured and, on a technical aspect, in relation to the text processing and the applied analysis technique. Then, those techniques are being extensively examined and grouped under the following categories: the lexicon based technique, the supervised machine learning, the unsupervised machine learning, and the combination of the above. Proceeding, examples of studies in each category and tables that help review the state of the art technology are presented and useful conclusions are drawn. Also, according to the general usefulness of the sentiment analysis especially in business, applications are reviewed along some practical examples of their use. Finally, useful observations on the current level of sentiment analysis and concideration for its future are deduced.

Keywords: <<Sentiment analysis, social networks, (un)supervised learning, lexicon, microblogs, classifier>>

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτριο Ασκούνη που είχε την επίβλεψη της συγκεκριμένης διπλωματικής εργασίας για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και σύγχρονο αντικείμενο.

Θέλω να πω ένα μεγάλο ευχαριστώ στην κ. Ευμορφία Μπιλίρη για την αμέριστη βοήθεια της και το ενδιαφέρον της καθ' όλη την διάρκεια της παρούσας μελέτης.

Πίνακας περιεχομένων

1 Εισαγωγή

Τη σημερινή εποχή, η ανάπτυξη του Διαδικτύου έχει αλλάξει σε μεγάλο βαθμό την καθημερινότητά μας προσφέροντας νέους τρόπους επικοινωνίας, ενημέρωσης και αλληλεπίδρασης μεταξύ των ανθρώπων. Οι χρήστες του Internet δεν είναι πλέον παθητικοί αποδέκτες πληροφοριών αλλά συμμετέχουν σε κοινωνικά δίκτυα και έχουν τη δυνατότητα να συζητήσουν με άλλους χρήστες, να ανταλλάξουν απόψεις και ιδέες. Ταυτόχρονα βλέπουμε να απομακρύνονται από τις κλασικές υπηρεσίες όπως τα e-mails. Η άποψη της κοινής γνώμης γύρω από ποικίλα θέματα είναι πάρα πολύ σημαντική και για αυτό γίνεται προσπάθεια να την κατανοήσουμε μέσω ερωτηματολογίων και δημοσκοπήσεων. Καθώς ολοένα και περισσότεροι χρήστες αναρτούν κριτικές σχετικά με τα προϊόντα ή υπηρεσίες που χρησιμοποιούν, οι πλατφόρμες κοινωνικής δικτύωσης αποτελούν σημαντικές πηγές πληροφοριών όσον αφορά την άποψη και τα συναισθήματα των ανθρώπων οι οποίες για πρώτη φορά στην ιστορία είναι καταγεγραμμένες απευθείας σε ηλεκτρονική μορφή. Η ανάγκη ανάλυσης και αξιοποίησης αυτής της πληροφορίας με αυτοματοποιημένο τρόπο οδήγησε στην εμφάνιση του επιστημονικού πεδίου της Ανάλυσης Συναισθήματος (Sentiment Analysis). Η Ανάλυση Συναισθήματος σε δεδομένα κοινωνικών δικτύων αποκτά ολοένα και περισσότερο έδαφος στον ακαδημαϊκό χώρο αλλά και στον επιχειρηματικό χώρο χάρη στις πολλά υποσχόμενες προοπτικές της.

1.1 Αντικείμενο διπλωματικής

1.1.1 Ανάλυση Συναισθήματος

Η Ανάλυση Συναισθήματος (Sentiment Analysis) ή αλλιώς η Εξόρυξη Γνώμης (Opinion Mining) είναι ο τομέας της Επεξεργασίας Φυσικής Γλώσσας που ασχολείται με την υπολογιστική ανάλυση των απόψεων, των συναισθημάτων, των εκτιμήσεων, των αξιολογήσεων και των στάσεων των ανθρώπων προς οντότητες, όπως άτομα, προϊόντα, υπηρεσίες, θέματα, γεγονότα και τα χαρακτηριστικά τους [Liu12]. Η ραγδαία ανάπτυξη της τεχνολογίας τα τελευταία χρόνια έχει οδηγήσει σε ευρεία χρήση του διαδικτύου. Καθημερινά, όλο και περισσότεροι χρήστες χρησιμοποιούν το διαδίκτυο για να εκφράσουν τη γνώμη τους

και για να μοιραστούν τις εμπειρίες και τα συναισθήματά τους με άλλους χρήστες. Με την εμφάνιση του Web 2.0 δίνεται η δυνατότητα στο χρήστη, εκτός από καταναλωτής, να είναι και παραγωγός πληροφορίας. Μ' άλλα λόγια, ο χρήστης έχει τη δυνατότητα να αλληλεπιδρά και να μοιράζεται δεδομένα μ' άλλους χρήστες. Παρά το γεγονός ότι πρόκειται για σχετικά πρόσφατο τομέα της υπολογιστικής- γλωσσολογικής έρευνας, έχει αναδειχθεί σε ένα αρκετά δραστήριο ερευνητικό τομέα, λόγω των τεχνικών προκλήσεων που θέτει αλλά κυρίως λόγω της πληθώρας των εφαρμογών που προσφέρει σχεδόν σε κάθε πεδίο (βιομηχανία προϊόντων, marketing, οικονομικές και πολιτικές επιστήμες, Μέσα Μαζικής Επικοινωνίας, ψυχολογία κ.α.), αλλά και των ποικίλλων ερευνητικών ζητημάτων που έχει θέσει.

Τα συναισθήματα που κρύβουν τα δεδομένα του διαδικτύου έχουν τεράστια σημασία τόσο σε παρόχους προϊόντων και υπηρεσιών, όσο και σε καταναλωτές. Οι κύριες μέθοδοι για την εξαγωγή του συναισθήματος από τα δεδομένα γίνεται με τη χρήση λεξικών και χρήση αλγορίθμων ταξινόμησης, οι οποίοι κατατάσσουν αυτόματα τα κείμενα σε κατηγορίες. Δηλαδή, χρησιμοποιώντας τη γνώση που αποκτούν κατά την εκπαίδευση, εξάγουν το συναίσθημα για τα επόμενα κείμενα. Πολύ σημαντικό ρόλο σε αυτό παίζει η ακρίβεια με την οποία θα εξαχθεί το συναίσθημα. Άλλος ένας πολύ σημαντικός παράγοντας που αφορά στις σύγχρονες εφαρμογές είναι η χρήση μεγάλου όγκου δεδομένων τόσο κατά την εκπαίδευση, όσο και κατά τη χρήση του ταξινομητή στην κατάταξη νέων κειμένων σε κατηγορίες. Αυτό οφείλεται στη ραγδαία αύξηση των δεδομένων που υπάρχουν στο διαδίκτυο. Στην προσπάθεια να αξιοποιήσουμε αποτελεσματικά τον τεράστιο όγκο δεδομένων που παράγονται καθημερινά από απλούς χρήστες (user-generated content) στα μέσα κοινωνικής δικτύωσης, έχουν πραγματοποιηθεί σημαντικές έρευνες εφαρμόζοντας διαφορετικές τεχνικές και προσεγγίσεις.

Από τις πιο δημοφιλείς εφαρμογές του Web 2.0 είναι τα κοινωνικά δίκτυα, όπως το Facebook και το Twitter. Σύμφωνα με επίσημα στατιστικά ("Twitter Statistics," Available: [1]), το Twitter διαθέτει πάνω από 288 εκατομμύρια ενεργούς χρήστες που δημοσιεύουν περισσότερα από 500 εκατομμύρια tweets συνολικά την ημέρα. Κρίνεται σκόπιμο να αναφερθεί, πως το Facebook και το Twitter αποτελεί μόνο ένα μικρό μέρος των Web 2.0 εφαρμογών, γεγονός που υποδηλώνει τον τεράστιο όγκο δεδομένων που διακινείται στο διαδίκτυο και από άλλες εφαρμογές όπως Facebook, blogs, forum, Google+, Vine, Instagram, κτλ. Η πληροφορία που κρύβεται ανάμεσα σε ένα τεράστιο όγκο δεδομένων είναι καθοριστικής σημασίας για τις σύγχρονες εφαρμογές. Ωστόσο, πρέπει να τονιστεί ότι η απαίτηση των σύγχρονων εφαρμογών για χρήση μεγάλου όγκου δεδομένων είναι δύσκολο να ικανοποιηθεί με τη χρήση κλασικών αλγορίθμων ταξινόμησης. Δηλαδή, η χρήση μεγάλου όγκου δεδομένων από κλασικούς ταξινομητές έχει ως αποτέλεσμα αρκετά μεγάλους χρόνους εκπαίδευσης και κατάταξης νέων εγγράφων σε κατηγορίες, με αποτέλεσμα να μην είναι δύσκολο να καλυφθούν οι χρονικές απαιτήσεις που έχουν οι σύγχρονες εφαρμογές.

1.1.2 Ανάλυση Συναισθήματος και Κατηγοριοποίηση Κειμένου

Από τις αρχικές εργασίες πάνω στο πρόβλημα της Ανάλυσης Συναισθήματος [Tur02] έγινε κατανοητό ότι η κατηγοριοποίηση συναισθήματος (sentiment classification) διαφέρει από το κλασικό πρόβλημα κατηγοριοποίησης κειμένου. Η κατηγοριοποίηση κείμενου ή ανίχνευση θέματος, αναφέρεται στην αντιστοίχηση κειμένων φυσικής γλώσσας σε θεματικές κατηγορίες ή κλάσεις οι οποίες ανήκουν σε ένα προκαθορισμένο σύνολο [Seb02].

Οι κατηγορίες στις οποίες εντάσσεται το κείμενο καθορίζονται με βάση τα θέματα-στόχους του εκάστοτε προβλήματος. Επομένως, διαφορετικά προβλήματα ταξινόμησης κειμένου βασίζονται σε διαφορετικά σύνολα κατηγοριών. Το πλήθος των κατηγοριών σε ένα σύνολο ποικίλει˙ μπορεί να εκτείνεται από ένα μικρό σύνολο δύο μόνο κατηγοριών έως σύνολα με δεκάδες κατηγορίες. Παράλληλα, ανάλογα με το πρόβλημα και το σύνολο κατηγοριών, ένα κείμενο μπορεί να ανήκει σε μια ή περισσότερες επικαλυπτόμενες κατηγορίες π.χ. ένα άρθρο να αντιστοιχηθεί με τις κατηγορίες “πολιτική”, “οικονομία” και “επικαιρότητα”.

Αντίθετα, η Ανάλυση Συναισθήματος αναφέρεται σε ένα μικρό σύνολο κατηγοριών (π.χ. θετικό, αρνητικό, ουδέτερο). Συγκεκριμένα, λόγω του ότι επικεντρώνεται στην κατάταξη ενός κειμένου ως προς την πολικότητα του, οι κατηγορίες είναι ανεξάρτητες της θεματολογίας του προβλήματος και μεταξύ τους αμοιβαία αποκλειόμενες, κάνοντας έτσι το πρόβλημα που προσπαθεί να επιλύσει η Ανάλυση Συναισθήματος ένα πιο απλό πρόβλημα με το οποίο ασχολείται η Επεξεργασία Φυσικής Γλώσσας [Liu12].

Ο υπολογιστής δεν χρειάζεται να αντιλαμβάνεται πλήρως τη σημασιολογία της κάθε πρότασης, αλλά να εντοπίζει τη συνολική στάση του συγγραφέα και να την ταξινομεί ως προς την πολικότητά της. Βέβαια, αυτή είναι μια διαδικασία η οποία αρκετές φορές παρουσιάζει δυσκολίες ακόμη και για τον άνθρωπο.

1.1.3 Δυσκολίες & Προκλήσεις

Αναγνωρίζοντας ένα συγκεκριμένο σύνολο λέξεων-κλειδιών (keywords) θα μπορούσαμε να προσδιορίσουμε τη συνολική πολικότητα της άποψης που εκφράζεται στο κείμενο, καθώς η πολικότητα ενός κειμένου προκύπτει από την πολικότητα των μεμονωμένων λέξεων από τις οποίες απαρτίζεται. Η παραπάνω διαδικασία είναι μία από τις πρώτες μεθόδους που χρησιμοποιήθηκαν και υιοθετεί μία πολύ δημοφιλή και αποτελεσματική τεχνική για την ανίχνευση θέματος. Ωστόσο, η προσέγγιση μέσω λέξεων κλειδιών στην ανίχνευση συναισθήματος δεν εμφανίζει υψηλά ποσοστά ακρίβειας και έχει αποδειχθεί ελλιπής σε ορισμένες περιπτώσεις.

Έτσι λοιπόν, αν δούμε αναλυτικά τις διαφορές που έχουν οι δύο μέθοδοι στην προσπάθεια τους να παράγουν αποτέλεσμα, μπορούμε να πούμε ότι το πρόβλημα της κατηγοριοποίησης συναισθήματος είναι πιο δύσκολο σε σχέση με την ανίχνευση θεματολογίας. Μία από τις πιο σημαντικές διαφορές με την κατηγοριοποίηση θεματολογίας και τις δυσκολίες στην περιοχή της Ανάλυσης Συναισθήματος είναι, ότι "το συναίσθημα/άποψη μπορεί πολλές φορές να εκφρασθεί με πιο λεπτό/έμμεσο τρόπο χωρίς τη χρήση συναισθηματικά φορτισμένων λέξεων με αποτέλεσμα να είναι δύσκολο να αναγνωρισθεί από τους επιμέρους όρους του κειμένου όταν αυτοί εξετάζονται μεμονωμένα" [PL08]. Επιπρόσθετα, πέρα από τον προσδιορισμό της πολικότητας, όταν απουσιάζουν συναισθηματικά φορτισμένες λέξεις, ιδιαίτερα απαιτητικός είναι και ο διαχωρισμός των υποκειμενικών και αντικειμενικών λέξεων και φράσεων ενός κειμένου. Όπως αναφέρεται από τους Kim και Hovy στο [KH06] "πολλές φορές ακόμη και άνθρωποι διαφωνούν για το αν μία δήλωση αποτελεί άποψη ή όχι". Ένα άλλο ζήτημα που απασχολεί ιδιαίτερα την Ανάλυση Συναισθήματος είναι ο προσδιορισμός του κατόχου - εκφραστή της άποψης (opinion holder) που διατυπώνεται στο κείμενο. Το συγκεκριμένο θέμα έχει μελετηθεί εκτενώς, κυρίως σε αναλύσεις σε πολιτικά debates, εξετάζοντας αν η γνώμη ανήκει στο συγγραφέα/δημιουργό ή στον σχολιαστή. Η γενικότερη αντίληψη της θετικής ή αρνητικής άποψης δεν εξαρτάται άμεσα από το εκάστοτε θέμα συζήτησης. Ωστόσο, το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετούνται [PL08]. Άλλος ένας παράγοντας που επηρεάζει την πολικότητα είναι η σειρά των λέξεων και φράσεων στο κείμενο. Οι ίδιες λέξεις με διαφορετική σειρά μπορεί να οδηγήσουν σε τελείως διαφορετική συνολική πολικότητα. Τέλος, στις δυσκολίες που συναντά η Ανάλυση Συναισθήματος πρέπει να συμπεριληφθούν και οι προκλήσεις της ευρύτερης περιοχής της Επεξεργασίας Φυσικής Γλώσσας, όπως η αμφισημία, ο χειρισμός της άρνησης, η ειρωνεία και ο σαρκασμός.

1.1.4 Ανάλυση Συναισθήματος σε Μικρο-Ιστολόγια ,Twitter

Η ανάλυση συναισθήματος αποκτά ιδιαίτερο ενδιαφέρον με την διείσδυση των κοινωνικών δικτύων στην καθημερινότητα του ανθρώπου. Πλέον ο κάθε χρήστης έχει την δυνατότητα να επικοινωνεί με πολλά διαφορετικά μέσα, από τα πιο παραδοσιακά, όπως το email ή τα blogs, μέχρι τα πιο σύγχρονα, όπως το Facebook και το Twitter. Τα τελευταία ονομάζονται και μικρο-ιστολόγια (microblogs), στα οποία οι χρήστες μπορούν καθημερινά να μοιράζονται τις προσωπικές απόψεις τους για διάφορα ζητήματα, αλλά και να ανταλλάσσουν απόψεις με άλλους χρήστες. Επίσης, τους δίνεται η δυνατότητα, πέρα από την συγγραφή κειμένου, να εκφράζονται και με διαφορετικούς τρόπους όπως με τη χρήση του «like» και το «share» στο Facebook. Γίνεται φανερό ότι τα κοινωνικά δίκτυα μεταμορφώνονται σε πηγές

συναισθήματος και τα δεδομένα που εξάγονται από αυτά αποκτούν ολοένα και μεγαλύτερη σημασία.

Το Twitter είναι μια πλατφόρμα κοινωνικής δικτύωσης που δημιουργήθηκε το 2006 και επιτρέπει στους χρήστες να ανταλλάσσουν μηνύματα μέχρι 140 χαρακτήρων˙ τα μηνύματα αυτά είναι γνωστά ως "tweets". Επιπλέον αποτελεί, μια διαρκώς αναπτυσσόμενη υπηρεσία, που σήμερα έχει πάνω από 500 εκατομμύρια χρήστες, οι οποίοι παράγουν συνολικά 340 εκατομμύρια μηνύματα κάθε μέρα.

Μέσα στους 140 χαρακτήρες που έχει στην διάθεση του ένας χρήστης, εκτός από το κυρίως κείμενο (text) και συνδέσμους (URLs), υπάρχουν και κάποια ειδικά σύμβολα του Twitter που μπορεί να χρησιμοποιηθούν. Το πρώτο είναι το "@" ακολουθούμενο από ένα όνομα χρήστη (username), το οποίο χρησιμοποιείται για ειδική αναφορά σε κάποιον άλλον χρήστη. Ένα άλλο ειδικό σύμβολο είναι το "#", ακολουθούμενο από μια λέξη. Αυτός ο συμβολισμός αναφέρεται σαν "hashtag" και χρησιμοποιείται για να δείξει ο χρήστης σε ποιο θέμα αναφέρεται το tweet του. Έτσι, κατά κάποιο τρόπο ομαδοποιεί tweets που αναφέρονται σε κοινά ζητήματα. Τέλος, υπάρχει περίπτωση ένας χρήστης να αναμεταδώσει (ReTweet) ένα μήνυμα ενός άλλου χρήστη, οπότε μπροστά από το μήνυμα μπαίνει αυτόματα το σύμβολο "RT". Γενικότερα τα tweets δεν είναι τόσο στοχαστικά, με την έννοια ότι δεν έχουν μια σαφώς εκφρασμένη άποψη σχετικά με ένα αντικείμενο, όπως συνηθίζεται σε forums και blogs. Μπορούν όμως να εκφράζουν άμεσα και με σαφήνεια την άποψη του ατόμου για κάποιο συγκεκριμένο θέμα.

Τα κύρια χαρακτηριστικά που διαφοροποιούν το Twitter, έναντι των άλλων κοινωνικών δικτύων, που το καθιστούν ένα εργαλείο που υπερτερεί σε εφαρμογές εντοπισμού συναισθήματος είναι:

- Ο μοναδικός τρόπος αλληλεπίδρασης: Ο κάθε χρήστης μπορεί να δημοσιεύσει (post) μόνο σύντομα μηνύματα, μέχρι 140 χαρακτήρες. Έτσι, όλο το συναίσθημα «εγκλωβίζεται» στο κείμενο, σε αντίθεση με άλλες πλατφόρμες όπου το συναίσθημα αποκτά και άλλες μορφές (όπως το "Like" στο Facebook ή το "+1" στο Google+).
- Κοινωνικό γράφημα (Social graph): Οι χρήστες του Twitter δημιουργούν ένα κοινωνικό γράφημα με ιδιαίτερη δομή. Όταν λέμε ότι ένας χρήστης «ακολουθεί» (following) έναν άλλον, τότε εννοούμε ότι παρακολουθεί τα tweets του. Αντίθετα αυτοί που παρακολουθούνται από άλλους χρήστες έχουν «οπαδούς» (followers).
- Το μεγαλύτερο μέρος των δεδομένων του Twitter είναι ελεύθερο στο κοινό (μέσω του Streaming API) και συνεπώς είναι δυνατή η συλλογή ενός επαρκή αριθμού δεδομένων.
- Τα tweets περιέχουν χρονοσφραγίδα (timestamp) δείχνοντας έτσι την αλληλουχία των γεγονότων.

Η Ανάλυση Συναισθήματος όταν εφαρμόζεται σε δεδομένα από μικρο-ιστολόγια και κοινωνικά δίκτυα, καλείται να αντιμετωπίσει περαιτέρω δυσκολίες, οι οποίες οφείλονται στην ιδιαίτερη φύση των κειμένων :

- Μήκος Κειμένου: τα μηνύματα είναι συνήθως σύντομα (π.χ. μέγιστο όριο 140 χαρακτήρες στο Twitter). Αν και ο περιορισμός μήκους μπορεί να οδηγήσει σε περιεκτικές και επί του θέματος τοποθετήσεις, πολλές φορές απουσιάζει το ευρύτερο εννοιολογικό πλαίσιο με αποτέλεσμα να μην είναι σαφής η πολικότητα του κειμένου [BS10].
- Λεξιλόγιο: τα περισσότερα κείμενα διατυπώνονται σε ανεπίσημη, καθομιλούμενη γλώσσα, και εμφανίζουν πολύ μεγαλύτερη ποικιλομορφία σε σχέση με άλλα είδη κειμένου. Περιλαμβάνουν αργκό, νεολογισμούς, εσκεμμένες παραλλαγές λέξεων για έμφαση (επιμήκυνση φθόγγων, χρήση κεφαλαίων γραμμάτων), συντομογραφίες (π.χ. "gr8"-"great") που καθιστούν δύσκολη την εφαρμογή λεκτικών αναλυτών ή άλλων εργαλείων που στηρίζονται στη γραπτή και πιο επίσημη μορφή της γλώσσας.
- Θόρυβος: οι πλατφόρμες κοινωνικής δικτύωσης επιτρέπουν μία αυθόρμητη επικοινωνία σε πραγματικό χρόνο, όπου πολλές φορές οι χρήστες αναρτούν μηνύματα χωρίς να ελέγχουν για συντακτικά ή γραμματικά λάθη. Ένα μεγάλο ποσοστό από τα δεδομένα που παράγονται περιέχει ακούσια ορθογραφικά λάθη και ακατανόητες εκφράσεις, τα οποία συνιστούν ουσιαστικά θόρυβο. Η αναγνώριση και αποκλεισμός τους αποτελεί ιδιαίτερη πρόκληση για τα σύγχρονα συστήματα ανίχνευσης συναισθήματος.
- Πολυγλωσσικό Περιεχόμενο: τα μέσα κοινωνικής δικτύωσης εξαπλώνονται σε μη αγγλόφωνες χώρες, αποκτώντας χρήστες που χρησιμοποιούν και γράφουν σε διαφορετικές γλώσσες, αρκετές φορές ακόμη και σε επίπεδο πρότασης ή μηνύματος. Το φαινόμενο αυτό έχει ως αποτέλεσμα, ιδιαίτερα διαδεδομένες τεχνικές στοχευμένες σε συγκεκριμένες γλώσσες (language-specific), να καθίστανται πρακτικά μη εφαρμόσιμες.

1.1.5 Αξιοποίηση της ανάλυσης συναισθήματος από επιχειρήσεις

Τα τελευταία χρόνια η Ανάλυση Συναισθήματος προσελκύει όλο και περισσότερο το ενδιαφέρον της ακαδημαϊκής κοινότητας, αλλά και των επιχειρήσεων χάρη στις πιθανές εφαρμογές της, κυρίως στον τομέα της Επιχειρηματικής Ευφυίας (Business Intelligence). Πρωτοπόρες εταιρείες επενδύουν στην εξόρυξη γνώμης από τα μέσα κοινωνικής δικτύωσης,

χρησιμοποιώντας τεχνικές ανάλυσης συναισθήματος. Ιδιαίτερα κρίσιμη είναι η έγκαιρη ανάλυση της γνώμης των καταναλωτών στα κοινωνικά δίκτυα και η κατάλληλη προσαρμογή στις ανάγκες τους, καθώς αυξάνεται συνεχώς ο αριθμός των ανθρώπων που στηρίζονται σε αξιολογήσεις άλλων καταναλωτών πριν λάβουν την τελική απόφαση αγοράς. Η ανάλυση συναισθήματος μπορεί να βοηθήσει σε πολλούς τομείς τις επιχειρήσεις.

Συγκεκριμένα μπορεί να βελτιώσει την εξυπηρέτηση πελατών. Στην περίπτωση αυτή η ανάλυση συναισθήματος δίνει καλές πληροφορίες για τις παρούσες και για τις μελλοντικές προτιμήσεις των πελατών, τα θέματα ενδιαφέροντος, τις απόψεις, την αρέσκεια για προϊόντα, και την εξυπηρέτηση. Αυτό δίνει την δυνατότητα να χαράξει η επιχείρηση μια στρατηγική για να επωφεληθεί από τα θετικά συναισθήματα και να αντιπαλέψει τα αρνητικά συναισθήματα για το προϊόν ή την υπηρεσία. Επίσης, η ανάλυση συναισθήματος δύναται να ποσοτικοποιεί της αντιλήψεις για την επιχείρηση, το προϊόν ή την υπηρεσία, τις διαφημιστικές καμπάνιες, κτλ. Επιπλέον, οργανισμοί μπορούν να χρησιμοποιήσουν αυτή την πληροφορία από τη ανάλυση συναισθήματος για καλύτερο σχεδιασμό μάρκετινγκ ώστε να βελτιωθεί η φήμη τη επιχείρησης, έχοντας γνώση για τα συναισθήματα και τις προτιμήσεις των ανταγωνιστών. Επίσης, επιτρέπει στην επιχείρηση να συγκρίνει τις επιδόσεις της με αυτές των ανταγωνιστών. Μια επιχείρηση, εκμεταλλευόμενη τα αποτελέσματα της ανάλυσης συναισθήματος θα μπορεί να προβλέπει την μόδα και να σχεδιάσει κατάλληλα προϊόντα ώστε να είναι στην κορυφή του ανταγωνισμού. Τέλος, η ανάλυση συναισθήματος εξουσιοδοτεί επιχειρήσεις παρέχοντάς τους εκτεταμένες και διορατικές πληροφορίες για τις προτιμήσεις του κοινού. Κάνοντας σωστή χρήση αυτών, δίνεται η δυνατότητα για νέες επιχειρηματικές κινήσεις και ευκαιρίες. Παρέχει λοιπόν την επιχειρηματική ευφυΐα με την οποία θα παρθούν εύστοχες αποφάσεις για την ανάπτυξη της επιχείρησης.

1.2 Οργάνωση κειμένου

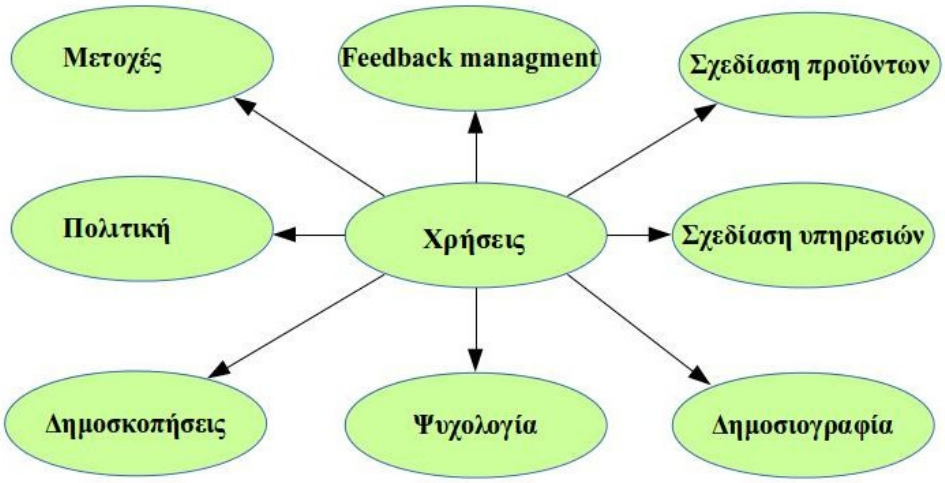
Στην συνέχεια της διπλωματικής εργασίας το κείμενο οργανώνεται ως εξής:

Κατηγορίες με βάση τις χρήσεις, την προσέγγιση κειμένου και τις τεχνικές ανάλυσης συναισθήματος παρουσιάζονται στο Κεφάλαιο 2 . Το Κεφάλαιο 3 συζητά τις τεχνικές με λεξικά. Στο Κεφάλαιο 4 αναπτύσσουμε τις τεχνικές με επιβλεπόμενη μηχανική μάθηση. Στο Κεφάλαιο 5 ασχολούμαστε με τις υβριδικές μεθόδους και αναφερόμαστε σε συγκεκριμένες αντιπροσωπευτικές εργασίες. Στη συνέχεια, στο Κεφάλαιο 6 αναφέρονται εφαρμογές ελεύθερες και εμπορικές καθώς και παραδείγματα χρήσης των εφαρμογών αυτών. Τέλος, στο Κεφάλαιο 7 διατυπώνονται τα συμπεράσματα και οι μελλοντικές προοπτικές για το αντικείμενο της ανάλυσης συναισθήματος.

2 Κατηγοριοποίηση

2.1 Κατηγοριοποίηση με βάση τις χρήσεις

Οι χρήσεις της ανάλυσης συναισθήματος είναι πάρα πολλές και αυξάνονται με ραγδαίους ρυθμούς καθώς οι άνθρωποι ανακαλύπτουν περισσότερα για τη χρησιμότητά της σε διάφορους τομείς, τόσο στις επιχειρήσεις όσο και σε διαφορετικά πεδία της καθημερινότητας και των ενδιαφερόντων του ανθρώπου. Μερικές από τις κατηγορίες χρήσης παρουσιάζονται παρακάτω.



Σχήμα 2.1: Κατηγορίες χρήσης

- **Ψυχολογία**

  Τα δεδομένα που προέρχονται από τα κοινωνικά δίκτυα είναι πάρα πολύ χρήσιμα τόσο σε κοινωνικό όσο και σε ψυχολογικό επίπεδο. Για παράδειγμα, έχει αντικαταστήσει τις παραδοσιακές μορφές που γινόταν η έρευνα πάνω στο τομέα της ψυχολογίας· ενώ παλιά χρησιμοποιούσαν ερωτηματολόγια και ακαδημαϊκές συνεντεύξεις, τώρα όλο και περισσότερο στρέφονται στις τεχνικές ανάλυσης συναισθήματος για αναλύσεις που αφορούν στην ψυχολογία. Αυτό μπορεί να βοηθήσει πολύ την επιστήμη καθώς πολύ μεγάλος πλούτος πληροφορίας μπορεί να

είναι στη διάθεσή τους άμεσα. Μπορούμε να βρούμε παραδείγματα εφαρμογών, όπως οι [WZJ+13], που εντοπίζουν την κατάθλιψη, που είναι μια από της σοβαρότερες ασθένειες της εποχής μας. Μια ακόμα πολύτιμη εφαρμογή θα μπορούσε να είναι η αναγνώριση και λήψη μέτρων σε περιπτώσεις που υπάρχει κίνδυνος αυτοκτονίας.

- **Δημοσιογραφία**

  Λόγω του ότι η ανάλυση συναισθήματος στοχεύει να βγάλει ένα συμπέρασμα για το πώς σκέφτεται ο κόσμος, ένας άλλος τομέας που μπορούμε να δούμε μια πολύ διαδεδομένη χρήση της ανάλυσης συναισθήματος είναι η δημοσιογραφία. Για αυτό πολλές εφαρμογές ενημέρωσης, όπως οι Politico, Pew, NBC, CNN, Current TV, Twitter, χρησιμοποιούν εργαλεία ανάλυσης συναισθήματος με σκοπό την ενημέρωση, αναφέρει ο Sam Petulla στο άρθρο [2].

- **Feedback managment, επιχειρήσεις**

  Η ανάλυση συναισθήματος βοηθά στην κατανόηση των προτιμήσεων του κόσμου. Μπορούμε να πούμε ότι τη σύγχρονη εποχή είναι το μέσο με το οποίο επικοινωνεί μια επιχείρηση με τους πελάτες της καθώς έχει αντικαταστήσει παλιά μέσα Customer relationship management (CRM), όπως ερωτηματολόγια, δημοσκοπήσεις κτλ. Έτσι βοηθά την επιχείρηση να πετύχει καλύτερη εξυπηρέτησή των πελατών και να βελτιώσει τη φήμη της σε σχέση με αυτή των ανταγωνιστών της. Εν γένει βελτιώνει την επιχειρηματική ευφυία της επιχείρησης. Υπάρχουν επίσης εταιρείες, όπως η Kia, η Best Buy (BBY), η Viacom (VIA.B) Paramount Pictures, η Cisco Systems (CSCO), και Intuit (INTU) που χρησιμοποιούν ανάλυση συναισθήματος για να καθορίσουν τον τρόπο που οι πελάτες, οι εργαζόμενοι και οι επενδυτές αισθάνονται, αναφέρεται στο άρθρο του bloomberg [3]. Κάποιες άλλες χρησιμοποιούν ακόμη και λογισμικό για να ελέγχει τον τόνο των μηνυμάτων ηλεκτρονικού ταχυδρομείου και τις άλλες επικοινωνίες.

  Χαρακτηριστικό παράδειγμα είναι στις 21 Ιουνίου 2005 ένας δημοσιογράφος έγραψε ένα απλό blog post για την εμπειρία του από τη χρήση προϊόντος από μια από τις μεγαλύτερες τεχνολογικές εταιρείες του κόσμου την DELL. Το κείμενό του με εκφράσεις όπως “DELL SUCKS. DELL LIES. Put that in your Google and smoke it, Dell,” προσέλκυσε το ενδιαφέρον πάρα πολλών χρηστών του διαδικτύου που ενδιαφέρονταν να αγοράσουν υπολογιστή. Έπειτα από αυτό η DELL έγινε γνωστή ως “Dell Hell,” και αυτό επέφερε ένα φαινόμενο ντόμινο με κακές κριτικές για την εταιρεία και δραματικές μειώσεις στην επιτυχία της και τις πωλήσεις της. Παρότι η εταιρεία είχε κτίσει μια πολύ ισχυρή φήμη κατά τη διάρκεια του 1990 και αρχές 21ου αιώνα, η εμπειρία ενός πελάτη έμελε να λειτουργήσει ως καταλύτης στη φήμη της

DELL για τα δύο επόμενα χρόνια κάνοντας τεράστια ζημία στην επιχειρηματική της φήμη. Η ρίζα του προβλήματος ήταν ότι η ίδια η εταιρεία είχε βάλει τον εαυτό της σε αυτή τη δύσκολη θέση διότι, αποτυγχάνοντας να επικοινωνήσει σωστά με τους ανθρώπους που υπήρχαν στη βάση της και τη στήριζαν˙ τους πελάτες της. Καταρχάς, θα έπρεπε να είχε συνυπολογίσει την διαδικτυακή παρουσία της επιχείρησής, έτσι ώστε να έχει μια καλύτερη κατανόηση για το κατά πόσο οι προτιμήσεις των πελατών της ήταν σε αντιστοιχία με αυτό που η ίδια προσέφερε.

Δισεκατομμύρια δολάρια δαπανώνται ετησίως στις μελέτες για μάρκετινγκ, για να προσδιοριστούν οι ανάγκες και οι επιθυμίες των καταναλωτών καθώς και ο σωστός τρόπος απεύθυνσης σε αυτούς. Αντίθετα, ανησυχητικό φαινόμενο είναι η έλλειψη της Online παρουσίας μιας επιχείρησης, γιατί αυτό σημαίνει έλλειψη ενδιαφέροντος για την επιχείρησή αυτή. Η ενασχόληση με την διαδικτυακή διαχείριση φήμης επιτρέπει να παρθούν ενεργητικά βήματα στην κατεύθυνση των στόχων που θέτει η εταιρεία, επιτρέποντάς της να αναφέρει ενδεχόμενο πρόβλημά της στο ευρύ κοινό, προτού κάποιος ανταγωνιστής προβεί σε δυσφήμισή της. Αναφερόμενοι στο παραπάνω παράδειγμα, η DELL να κατάλαβε τη σημασία της επικοινωνίας μέσα από το διαδίκτυο με τους πελάτες με σκοπό να προσδώσει αξία σε αυτούς. Έτσι προσπάθησε να συμμετέχει σε συζητήσεις ούτως ώστε να αντιστρέψει το αρνητικό κλίμα.

Η φήμη χτίζεται κατά τη διάρκεια του χρόνου αλλά μπορεί εύκολα να τραυματιστεί πολύ γρήγορα. Όμως, πάντα υπάρχουν τρόποι να χτιστεί μια γερή φήμη που δύσκολα θα επηρεάζεται.

- **Σχεδίαση προϊόντων και υπηρεσιών**

Μια από τις πιο ζωτικής σημασίας δραστηριότητες των επιχειρήσεων είναι η σχεδίαση προϊόντων και υπηρεσιών. Αυτό προϋποθέτει καλή ανάγνωση της αγοράς, των καταναλωτών, και καλή πρόβλεψη για το μέλλον. Ο τομέας που ασχολείται με αυτό το αντικείμενο είναι η Επιχειρηματική ευφυΐα (Business intelligence, BI), η οποία ορίζεται ως το σύνολο των τεχνικών και εργαλείων που χρησιμεύει στη μετατροπή δεδομένων σε ουσιαστικές και χρήσιμες πληροφορίες για τους σκοπούς της ανάλυσης των επιχειρήσεων. Η ανάλυση συναισθήματος είναι μια εφαρμογή που αποδεικνύεται ότι βοηθά πολύ στην αναγνώριση συμπάθειας, προτιμήσεων και προβλέπει μελλοντικές ανάγκες και απαιτήσεις και επομένως είναι ένα πολύ σημαντικό συμβουλευτικό εργαλείο σε όλες τις σύγχρονες επιχειρήσεις.

Για παράδειγμα, όταν η Εθνική Συνέλευση Ομοσπονδία Λιανικής (National Retail Federation Convention), με βάση την ανάλυση συναισθήματος περισσότερων από μισό εκατομμύριο δημόσιων αναρτήσεων σε μέσα κοινωνικής δικτύωσης χρησιμοποιώντας εργαλεία της IBM [4], προβλέπει ότι το “steampunk” (υπό-είδος ένδυσης εμπνευσμένο από τα ρούχα, την τεχνολογία και τα κοινωνικά ήθη της

βικτοριανής κοινωνίας), θα είναι μια σημαντική τάση την επόμενη περίοδο. Οι μεγάλες εταιρείες μόδας, οι σχεδιαστές αξεσουάρ και κοσμημάτων και κάθε ένας που ασχολείται με το ρουχισμό αναμένεται να ενσωματώσει μια αισθητική "steampunk" στα σχέδιά τους για το επόμενο έτος.

- **Μετοχές**

  Οι έρευνες γύρω από τον οικονομικό τομέα δείχνουν ότι η αγορά μετοχών μπορεί να επηρεαστεί άμεσα από άρθρα και απόψεις στα κοινωνικά δίκτυα. Η πληροφόρηση, όσο και οι συναισθηματικές πτυχές των ειδήσεων ή των απόψεων στα κοινωνικά δίκτυα, μπορεί να έχουν επίπτωση στις τιμές των μετοχών, στον όγκο των συναλλαγών, στην ευστάθεια της αγοράς και ακόμη και στα μελλοντικά κέρδη της εταιρείας. Όλο και περισσότερα στοιχεία δείχνουν ότι το συναίσθημα μπορεί να βοηθήσει στην πρόβλεψη της μεταγενέστερης δραστηριότητας της αγοράς. Η επίδραση των ειδήσεων σχετικά με το εμπόριο των τιμών είναι ασύμμετρη ως προς το χρόνο. Ειδήσεις που προξενούν θετικό συναίσθημα έχει αποδειχθεί ότι σχετίζονται με μεγάλες αυξήσεις τιμών, για ένα σχετικά σύντομο χρονικό διάστημα. Αντίθετα, ειδήσεις που προξενούν αρνητικό συναίσθημα συνδέονται με μειώσεις των τιμών, για ένα πιο παρατεταμένο χρονικό διάστημα αναφέρουν οι δημιουργοί του TheySay [5].

  Η σχέση μεταξύ μετοχών και ανάλυσης συναισθήματος γίνεται πολύ εμφανής στο εξής παράδειγμα.

  Το 2013 ένα ψεύτικο tweet οδήγησε τον Dow Jones σε κατακόρυφη πτώση σε μόλις δύο λεπτά και αφού η φάρσα αποκαλύφθηκε η επακόλουθη ανάκαμψη πήρε μόλις τρία λεπτά. Η πρόβλεψη του δείκτη τιμών των μετοχών (Dow Jones Industrial Average) γίνεται επίσης πιο ακριβής όταν λαμβάνεται υπόψη το κλίμα σε ολόκληρη τη πλατφόρμα του Twitter.

  Η χρήση αισιόδοξης ή απαισιόδοξης γλώσσας στις ειδήσεις επηρεάζει τον αναγνώστη διακριτικά. Εξαντλητικές μελέτες ανακάλυψαν ότι οι αγορές τείνουν να αντιδρούν υπερβολικά σε σχέση με τις κακές ειδήσεις. Όλοι οι συντάκτες των ειδήσεων συνειδητά ή ασυνείδητα εισάγουν τις δικές τους προκαταλήψεις και προσωπικές γνώμες στο κείμενό τους. Οι αναγνώστες των ειδήσεων λαμβάνουν πληροφορία τόσο για τα πραγματικά όσο και για τα συναισθηματικά στοιχεία των ειδήσεων και αντιδρούν έντονα σε περιεχόμενο που περιέχει μεγαλύτερο αρνητικό συναίσθημα. Επιπλέον, πολλοί συγγραφείς δεν είναι αμερόληπτοι σε σχέση με κάποια συγκεκριμένη μετοχή και την προκατάληψη αυτή την εισάγουν σκόπιμα για επηρεάσουν τον αναγνώστη. Η Υπόθεση των Αποτελεσματικών Αγορών (ΕΜΗ, efficient market hypothesis) αναφέρει ότι οι επενδυτές που είναι σε θέση να αποκτήσουν πρόσβαση σε κοινωνικά δίκτυα και σε μέσα μαζικής ενημέρωσης

γρήγορα είναι σε θέση να κάνουν καλύτερα τις συναλλαγές και να μεγιστοποιήσουν της απόδοσης των μετοχών τους.

- **Πολιτική**

  Συστήματα που ασχολούνται με τις υπηρεσίες μικρο-ιστολογίων (Microblogging), όπως το Twitter και άλλες πλατφόρμες κοινωνικής δικτύωσης, δείχνουν πολύ μεγάλη ανάπτυξη σε πολλές εφαρμογές, συμπεριλαμβανομένης της πολιτικής και των εκστρατειών μάρκετινγκ. Για παράδειγμα, οι τεχνικές με απευθείας σύνδεση μέσω διαδικτύου και η οργάνωσή τους έπαιξε ένα πολύ σημαντικό ρόλο στην εκστρατεία του προέδρου των ΗΠΑ Μπαράκ Ομπάμα το 2008. Η Current TV έκανε ένα πρόγραμμα κατά τη διάρκεια της συζήτηση μεταξύ Τζον Μακέιν και Μπαράκ Ομπάμα που ονομαζόταν "Hack the Debate", ζητώντας από το κοινό να δημοσιεύσει σχόλια στο Twitter. Με αυτό τον τρόπο μπόρεσαν να βγάλουν πάρα πολλά συμπεράσματα οι πολιτικοί ακούγοντας τις απόψεις του κόσμου. Η ανάλυση συναισθήματος και το μάρκετινγκ της πολιτικής εκστρατείας έπαιξε ένα καθοριστικό ρόλο στο τελικό αποτέλεσμα των εκλογών. Μετά την επιτυχία του Μπαράκ Ομπάμα, το Twitter έχει γίνει ένα μόνιμο κανάλι επικοινωνίας πάνω στην πολιτική σκηνή [6]. Ως εκ τούτου, μπορεί να αναμένεται ότι η ανάλυση συναισθήματος θα είναι μέρος της κάθε εκστρατείας στο μέλλον, αλλά και χρήσιμο εργαλείο για τον αφουγκρασμό της κοινωνίας από τα πολιτικά όργανα.

- **Δημοσκοπήσεις**

  Με τη εντυπωσιακή αύξηση των μέσων κοινωνικής δικτύωσης που έχουν βάση το κείμενο, εκατομμύρια ανθρώπων μεταδίδουν τις σκέψεις και τις απόψεις τους σχετικά με μια μεγάλη ποικιλία θεμάτων. Μπορούμε να αναλύσουμε τα διαθέσιμα δεδομένα και να συμπεράνουμε τις στάσεις-απόψεις του πληθυσμού με τον ίδιο τρόπο που δημοσκόποι κάνουν δημόσια ερωτήματα για να καταλάβουν την κοινή γνώμη. Έτσι η εξόρυξη της κοινής γνώμης από ελεύθερα διαθέσιμο περιεχόμενο κειμένου θα μπορούσε να είναι ταχύτερη και λιγότερο δαπανηρή από τις παραδοσιακές δημοσκοπήσεις. Για παράδειγμα, μια τυπική τηλεφωνική δημοσκόπηση των χιλίων ερωτηθέντων κοστίζει τουλάχιστον δεκάδες χιλιάδες δολάρια για να πραγματοποιηθεί. Από την άλλη μεριά, η ανάλυση αυτή από τα κοινωνικά δίκτυα θα επιτρέψει επίσης να λάβουμε υπόψη μια μεγαλύτερη ποικιλία δημοσκοπικών ερωτήσεων. Τη σημερινή εποχή μεγάλες εταιρείες δημοσκοπήσεων έχουν παραδεχθεί ότι χρησιμοποιούν εργαλεία της ανάλυσης συναισθήματος.

  Για παράδειγμα οι [BRS10], αναφέρουν ότι με τη χρήση ενός απλού ανιχνευτή συναισθήματος στο Twitter που αφορά στην εμπιστοσύνη των καταναλωτών αλλά και στις δημοσκοπήσεις για τις προεδρικές εκλογές έχουμε πολύ ενθαρρυντικά

αποτελέσματα σε σχέση με το χρόνο και το κόστος των αποτελεσμάτων. Τα αποτελέσματα τους δείχνουν ότι με περισσότερο προηγμένες NLP τεχνικές θα έχουμε ακόμα μεγαλύτερη βελτίωση της εκτίμησης της γνώμης που ήδη είναι σε πολύ ικανοποιητικό επίπεδο.

Υπάρχουν φυσικά και πολλές άλλες χρήσεις της ανάλυσης συναισθήματος στα κοινωνικά δίκτυα που έχουν κατακλύσει την καθημερινότητα του ανθρώπου. Εφαρμογές τόσο στις επιχειρήσεις όσο και γενικότερου σκοπού.

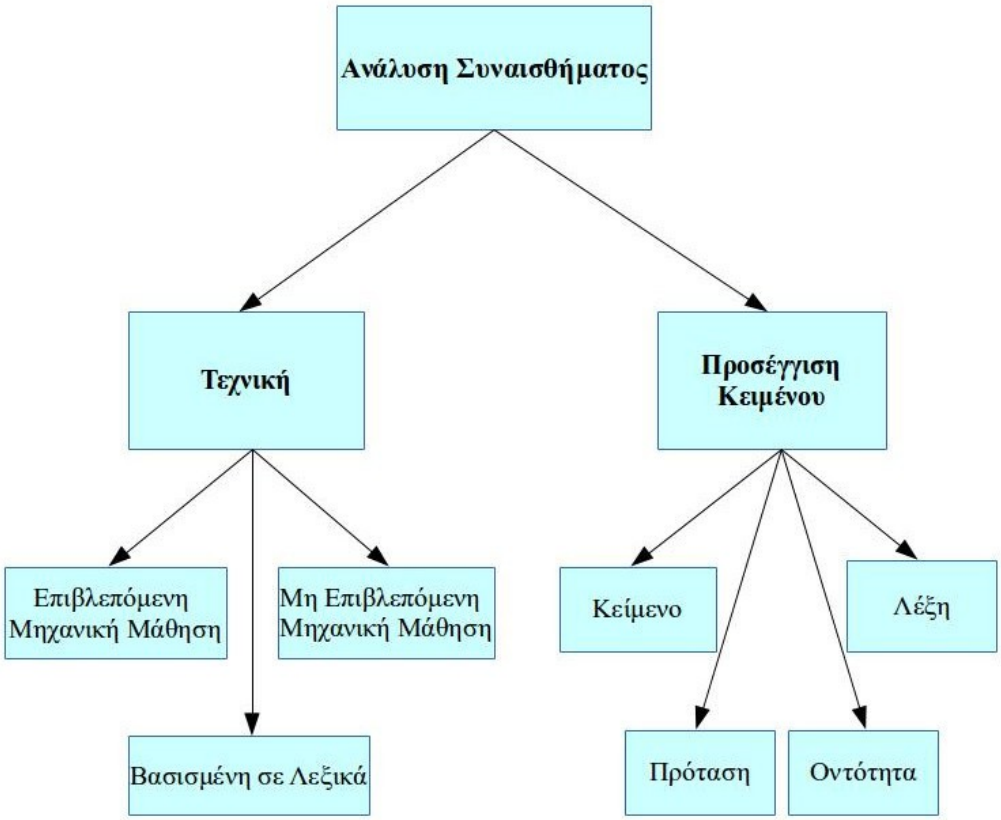
Για παράδειγμα όπως αναφέρεται εδώ [7], μη κερδοσκοπικές οργανώσεις, όπως η Αμερικανική Αντικαρκινική Εταιρεία, χρησιμοποιούν επίσης ανάλυση συναισθήματος. Η ACS χρησιμοποιεί ανάλυση συναισθήματος για να πάρει πληροφορίες σχετικά με τα προγράμματα και τις εκδηλώσεις της. Η ανάλυση συναισθήματος πρέπει να λάβει υπόψη της το ιδιάζων λεξιλόγιο του θέματος που θέλει να αναλύσει. Για παράδειγμα, επισημαίνεται ότι για άλλες εταιρείες, η λέξη "καρκίνος" και "σκοτώσει" θα μπορούσε να δείξει ένα αρνητικό σχόλιο, αλλά αυτό δεν ισχύει για την περίπτωση της ACS. Ενδεικτικά στοιχεία του μεγέθους της εταιρείας που κάνουν σαφή την ανάγκη για χρήση ανάλυσης συναισθήματος. Συγκεκριμένα η ACS παίρνει περίπου 6000 μηνύματα Twitter, Facebook, άρθρα σε blog, και σχόλια σχετικά με τα blogs ή άρθρα μέσα σε ένα μήνα. Έχει συγκεντρώσει περισσότερα από 4 δισεκατομμύρια δολάρια για την καταπολέμηση του καρκίνου. Την ACS “ακολουθούν” περίπου 62.000 άνθρωποι το χρόνο για στις εκδηλώσεις της. Και μέσω της αναλυσης συναισθήματος που έχουν κάνει στα στοιχεία που λαμβάνει έχουν καταλήξει στο συμπέρασμα ότι με τις εκδηλώσεις με τα αναμμένα κεριά (Luminaria Ceremony), όπου οι άνθρωποι τιμούν τους αγαπημένους που έχουν χάσει από καρκίνο ανάβοντας ένα κερί, είναι η πιο σημαντική στιγμή της εκδήλωσης για πολλούς.

Ένα άλλο παράδειγμα είναι η χρήση ανάλυσης συναισθήματος από την κυβέρνηση των ΗΠΑ για ζητήματα εθνικής ασφάλειας. Με βάση τους New York Times (2006), η Αμερικανική κυβέρνηση ξοδεύει περίπου $2.4 εκατομμύρια δολάρια χρηματοδοτώντας έρευνες που αναπτύσσουν λογισμικά κατασκευασμένα να παρακολουθούν τη διαδικτυακή δραστηριότητα.

Χαρακτηριστικά αναφέρεται “... η ανάλυση συναισθήματος προορίζεται να αναγνωρίζει πιθανές απειλές για τη χώρα. Θέλουμε να καταλαβαίνουμε τη ρητορική που εμφανίζεται στη δημοσιότητα και την ένταση της, καθώς και τη διαφορά μεταξύ αντιπάθειας και υπερβολικής καταγγελίας.” [8].

Σε κάθε περίπτωση, για να είναι δυνατή η εξαγωγή συναισθήματος, έχουμε κάποιες τεχνικές επεξεργασίας φυσικής γλώσσας και κάποιους αλγόριθμους που μπορούν να κατηγοριοποιηθούν με βάση τον τρόπο που προσεγγίζουμε το κείμενο αλλά και την ταξινόμηση συναισθήματος που θέλουμε να κάνουμε. Έτσι καθορίζονται κάποιοι τρόποι κατηγοριοποίησης της ανάλυσης συναισθήματος. Επίσης, αναλόγως με τη τεχνική και το βαθμό που παρεμβαίνει ο άνθρωπος στη διαδικασία, καθορίζονται επιπλέον κάποιες

κατηγορίες. Παραστατικά τα παρουσιάζουμε στο παρακάτω σχήμα και αναλύονται παρακάτω.



Σχήμα 2.2: Κατηγοριοποίηση Ανάλυσης Συναισθήματος

2.2 Κατηγορίες προσέγγισης κειμένου

2.2.1 Ταξινόμηση σε επίπεδο εγγράφου/κειμένου

Αυτή η προσέγγιση θεωρεί ότι κάθε έγγραφο περιέχει τις απόψεις ενός μόνο ατόμου γύρω από ένα συγκεκριμένο θέμα και έχει ως στόχο να χαρακτηρίσει το συναίσθημα που εκφράζεται μέσα από το κείμενο που περιλαμβάνει κρίσεις και απόψεις ως θετικό η αρνητικό. Το μειονέκτημα αυτής της προσέγγισης είναι ότι θεωρεί ως δεδομένο ότι σε ένα κείμενο η κριτική έχει μόνο ένα αντικείμενο αναφοράς, οπότε πρακτικά δεν είναι εφαρμόσιμη σε περιπτώσεις κειμένων που περιέχουν για παράδειγμα σύγκριση δύο διαφορετικών προϊόντων. Οι περισσότερες τεχνικές ανάλυσης συναισθήματος εγγράφων είναι επιβλεπόμενης μάθησης, ωστόσο υπάρχουν και τεχνικές μη επιβλεπόμενης μάθησης˙ έννοιες που θα αναλύσουμε παρακάτω με λεπτομέρεια.

2.2.2 Ταξινόμηση σε επίπεδο πρότασης

Ο τρόπος ανάλυσης σε αυτό το επίπεδο εστιάζει στην πρόταση και τον ακριβή προσδιορισμό της θετικής, αρνητικής ή ουδέτερης στάσης που εκφράζει. Γίνεται η παραδοχή ότι υπάρχει μόνο μια άποψη μέσα σε κάθε πρόταση. Οι προτάσεις μπορούν να χαρακτηριστούν απευθείας ως θετικές ή αρνητικές. Επιπλέον, η προσέγγιση αυτή συχνά συνδέεται με την ταξινόμηση υποκειμενικότητας (subjectivity classification), που διαχωρίζει τις προτάσεις που περιέχουν γεγονότα της πραγματικότητας-αντικειμενικά από αυτές που περιέχουν υποκειμενικές κρίσεις-προσωπικές απόψεις [Liu12], και στη συνέχεια αυτές οι οποίες περιέχουν κάποια στοιχεία υποκειμενικής πληροφορίας ταξινομούνται ως θετικές ή αρνητικές. Επίσης, συχνά συνυπολογίζονται πολλές παράμετροι, όπως το φαινόμενο της άρνησης (negation), π.χ. καθόλου καλός, το θέμα της τροπικότητας (modality), η αμφισημία των λέξεων, ο συντακτικός ρόλος των λέξεων στην πρόταση κ.α.

2.2.3 Ταξινόμηση σε επίπεδο λέξης

Το επίπεδο αυτό ουσιαστικά χρησιμοποιείται για ταξινόμηση επιπέδου πρότασης ή κειμένου και βασίζεται στην παραδοχή ότι οι πιο σημαντικοί δείκτες συναισθημάτων είναι οι λέξεις γνώμης (opinion words). Μια λίστα από τέτοιες λέξεις ονομάζεται λεξικό συναισθημάτων [Liu12]. Για την δημιουργία λεξικών συναισθημάτων χρησιμοποιούνται πληροφορίες που προκύπτουν από την επεξεργασία, είτε μεγάλων σωμάτων ηλεκτρονικών κειμένων (text corpora), είτε γλωσσολογικών πόρων, όπως θησαυροί και λεξικά, με σκοπό την επέκταση μιας αρχικής λίστας με λέξεις γνώμης (seed words). Στα λεξικά που προέρχονται από σώματα κειμένου (text corpora), η επέκταση της λίστας αυτής, μπορεί να γίνει με χρήση συντακτικών μοτίβων τα οποία ικανοποιούνται μέσα σε αυτά τα κείμενα. Ένας άλλος τρόπος επέκτασης είναι με τη χρήση πληροφοριών που προκύπτουν από τη συχνότητα διάφορων μοτίβων από λέξεις [Tur02]. Αντίθετα, τα λεξικά που βασίζονται σε γλωσσολογικούς πόρους προσπαθούν να πραγματοποιήσουν αυτή την επέκταση χρησιμοποιώντας τα συνώνυμα, τα αντώνυμα και την ιεραρχία αυτών των λέξεων μέσα σε γλωσσολογικούς θησαυρούς όπως το WordNet.

2.3 Ταξινόμηση σε επίπεδο οντότητας και χαρακτηριστικών

Η ταξινόμηση αυτού του επιπέδου εστιάζει στην ίδια την άποψη και όχι σε ανάλυση δομικών στοιχείων της γλώσσας (κείμενο, πρόταση, φράση). Παρατηρούμε ότι μερικές φορές η ταξινόμηση σε επίπεδο κειμένου ή πρότασης δεν είναι επαρκής για κάποιες εφαρμογές. Αυτό συμβαίνει διότι δεν είναι εφικτός, ούτε ο εντοπισμός των μεταβλητών (opinion targets) στις οποίες αναφέρεται μια γνώμη, ούτε και η ανάθεση ενός ξεχωριστού συναισθήματος σε κάθε μια από αυτές. Επιπλέον, αν υποθέσουμε ότι ένα υποκείμενο έχει μια άποψη (θετική ή αρνητική) για μια οντότητα, δεν σημαίνει ότι θα έχει την ίδια άποψη για κάθε επιμέρους χαρακτηριστικό της [Liu12]. Για αυτό χρειαζόμαστε μια λεπτομερή ανάλυση ούτως ώστε να είναι δυνατός ο διαχωρισμός όλων των χαρακτηριστικών στα οποία αναφέρεται μια άποψη,

όπως η εύρεση του συναισθηματικού τους φορτίου. Η εν λόγω ταξινόμηση βασίζεται κυρίως στη ιδέα ότι μια υποκειμενική κρίση αποτελείται από ένα συναίσθημα (semtiment) και έναν στόχο (target) στον οποίο απευθύνεται, ο οποίος στα περισσότερα συστήματα κειμενικής ανάλυσης αναπαρίσταται μέσω οντοτήτων (entities). Στόχος αυτής της ανάλυσης είναι να αναζητήσει τα εκφραζόμενα συναισθήματα και τις απόψεις προς τα αντικείμενα-στόχους, αλλά και τα επιμέρους χαρακτηριστικά τους (aspects).

Για παράδειγμα στην πρόταση "Η ποιότητα κλήσης του iPhone είναι καλή, αλλά η διάρκεια ζωής της μπαταρίας είναι μικρή", αξιολογούνται δύο διαφορετικές όψεις της ίδιας οντότητας (η ποιότητα κλήσης και η διάρκεια μπαταρίας).

Αυτού του είδους η ανάλυση συνεξετάζει και άλλους παράγοντες που σχετίζονται με την έκφραση της άποψης, όπως το πρόσωπο που εκφράζει την άποψη (opinion holder) αλλά και τον χρόνο της έκφρασης (time). Οι τεχνικές που συνήθως αξιοποιούνται για την ανάλυση αυτού του επιπέδου στοχεύουν στην εξόρυξη των χαρακτηριστικών των αξιολογούμενων οντοτήτων (feature extraction) και στον προσδιορισμό-κατηγοριοποίηση αυτών των χαρακτηριστικών (feature sentiment classification) ως προς το τρίπτυχο θετικό-αρνητικό-ουδέτερο, που υλοποιείται κυρίως μέσω των προσεγγίσεων που περιλαμβάνουν επιβλεπόμενη εκμάθηση μηχανής και δημιουργία λεξικών πόρων [Liu12].

2.4 Τεχνικές

Με βάση την τεχνική που χρησιμοποιούμε προσδιορίζονται κάποιες κατηγορίες για την ανάλυση συναισθήματος. Οι κυριότερες από αυτές, όπως δείχνει και το πιο πάνω σχήμα, είναι οι τεχνικές με Επιβλεπόμενη Μηχανική Μάθηση, Μη Επιβλεπόμενη Μηχανική Μάθηση, και οι τεχνικές Βασισμένες σε Λεξικά, καθώς και ο συνδυασμός αυτών. Επειδή αποτελούν τις σημαντικότερες κατηγορίες, τόσο σε ερευνητικό όσο και σε εμπορικό επίπεδο, θα αναφερθούμε με περισσότερη λεπτομέρεια στα επόμενα κεφάλαια.

3 Τεχνικές με λεξικά

Οι τεχνικές βασισμένες σε λεξικά (lexicon-based) χρησιμοποιούν προκατασκευασμένα λεξικά συναισθήματος, όπου μέσα σε αυτά χαρακτηρίζονται οι διάφοροι όροι του κειμένου και προκύπτει η συνολική πολικότητα.

Είναι τεχνικές που μπορούν να επιτυγχάνουν αρκετά καλά ποσοστά ακρίβειας όταν εφαρμόζονται σε γνωστά θεματικά πεδία όπου το λεξιλόγιο των κειμένων τους καλύπτεται από τα εκάστοτε λεξικά που χρησιμοποιούν. Στα πλεονεκτήματα τους εντάσσεται ότι δεν χρειάζονται σύνολα εκπαίδευσης και έτσι μπορούν να εφαρμοστούν σε πολύ μεγάλο εύρος θεμάτων. Ωστόσο έχουν και σημαντικά μειονεκτήματα, λόγω διαφόρων περιορισμών. Πρώτον, λόγο του περιορισμού του πλήθος των λέξεων στα λεξικά που χρησιμοποιούν οι μέθοδοι, δεν μπορεί να έχει ικανοποιητική απόδοση σε πολύ δυναμικά περιβάλλοντα, όπως το Twitter, όπου είναι γεμάτο νεολογισμούς και συντομογραφίες. Δεύτερον, τα λεξικά συναισθήματος αναθέτουν συνήθως σταθερό συναισθηματικό προσανατολισμό στις λέξεις χωρίς να εξετάζουν το πλαίσιο μέσα στο οποίο χρησιμοποιούνται και αυτό μπορεί να μας οδηγήσει σε εσφαλμένα συμπεράσματα.

3.1 Επισκόπηση τεχνικών

Με βάση τη προσέγγιση αυτή που κάνει χρήση λεξικών, όταν έχουμε ένα κείμενο στο οποίο θέλουμε να κάνουμε ανάλυση συναισθήματος, το επεξεργαζόμαστε ως ένα σύνολο από λέξεις ανεξάρτητων μεταξύ τους, των οποίων η γραμματική, η σύνταξη και η σειρά τους δεν μας απασχολεί. Η έκφραση που έχει αποδοθεί στην προσέγγιση αυτή είναι “σάκος από λέξεις” (bag of words).

Στη συνέχεια της διαδικασίας για τη ανάλυση συναισθήματος, θα πρέπει να δοθεί συναισθηματικό περιεχόμενο στις λέξεις του κειμένου. Αυτό γίνεται με χρήση των λεξικών συναισθήματος που θα αναλυθούν εκτενέστερα στη συνέχεια μιας και διαδραματίζουν πολύ σημαντικό ρόλο. Τα λεξικά κάνουν την απόδοση του συναισθήματος σύμφωνα με τις λέξεις που περιέχουν και εκφράζουν συναίσθημα (sentiment words) στις οποίες έχουν αποδοθεί βαθμολογίες που εκφράζουν κατά πόσο το νόημα της λέξης ταιριάζει σε συγκεκριμένες

κατηγορίες συναισθήματος. Το πιο σύνηθες είναι να κατηγοριοποιούνται σε θετικό, αρνητικό και ουδέτερο συναίσθημα. Ωστόσο, υπάρχουν και λεξικά με περισσότερες κατηγορίες όπως χαρά, ενθουσιασμός, λύπη κτλ, αλλά και διαβαθμίσεις όπως θετικό, αρνητικό, ουδέτερο, πολύ θετικό, πολύ αρνητικό, (positive, negative, neutral, high positive, high negative).

Για κάθε λέξη από το κείμενο, αναζητείται η αντίστοιχη λέξη στο λεξικό και κρατείται η βαθμολογία της. Στο τέλος το συνολικό συναίσθημα του κειμένου προσδιορίζεται από το άθροισμα των βαθμολογιών των επιμέρους λέξεων.

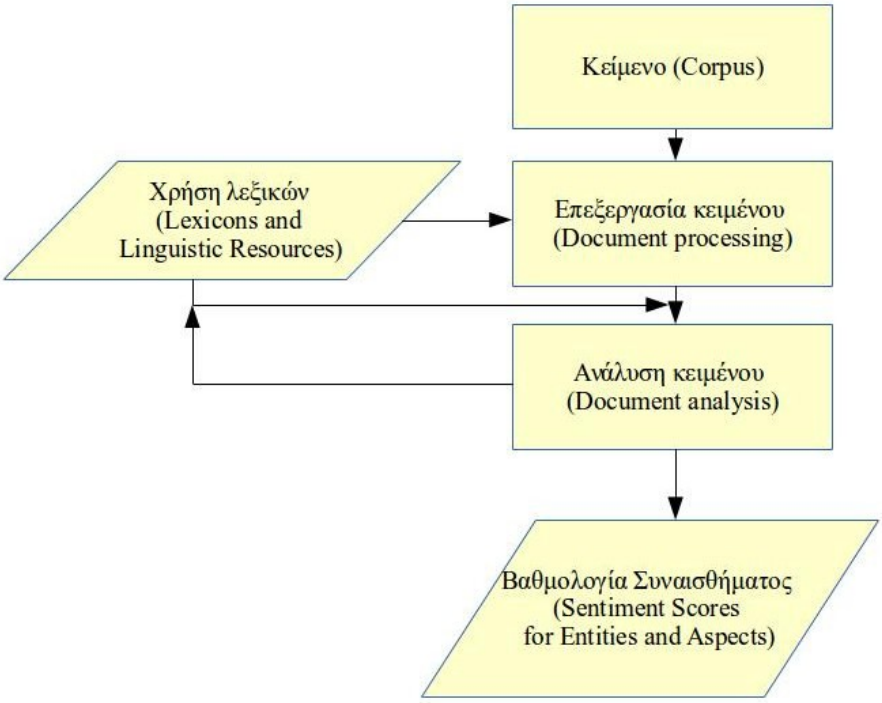
Αρκετά συστήματα διεκδικούν τον τίτλο δημιουργίας του πρώτου λεξικού συναισθήματος. Τα πρώτα παραδείγματα τέτοιων εργασιών είναι η [YH03], και η [HW00] όπου χρησιμοποιούν επίθετα για την πρόβλεψη της υποκειμενικότητας και εμφανίζουν πολύ ικανοποιητικά αποτελέσματα. Έδωσαν έμφαση στα βαθμωτά επίθετα (συγκριτικού – υπερθετικού βαθμού). Καταλήγοντας ότι τα λεξικολογικά χαρακτηριστικά, όπως η συναισθηματική κατεύθυνση και η βαθμίδα (συγκριτικού – υπερθετικού βαθμού), καθορίζουν σε μεγάλο βαθμό την υποκειμενικότητα την πρότασης. Επιπλέον έδωσαν έμφαση στα βαθμωτά επίθετα (συγκριτικού – υπερθετικού βαθμού).

Η εργασία του [Tur02], όπου θέλει να βαθμολογήσει κριτικές (reviews) με thumbs up ή down είναι ένα από τα πρώτα παραδείγματα χρήσης λεξικών συναισθήματος. Η μέθοδός του έχει τρία στάδια (1) εξαγωγή φράσεων που περιέχουν επίθετα ή επιρρήματα (2) υπολογισμός της συναισθηματικής κατεύθυνσης της κάθε φράσης και (3) κατηγοριοποίηση της κριτικής βασισμένη στο μέσο συναισθηματικό βάρος της φράσης.

Μερικές ενδιαφέρουσες παραλλαγές αυτών των γενικών μεθόδων είναι, η χρήση της πολικότητας από προηγούμενες προτάσεις σαν συνδετικός κρίκος όταν δεν μπορούμε να αποφανθούμε εύκολα, ή η ενσωμάτωση της πληροφορίας από επισημασμένα (labeled) δεδομένα. Ένα σημαντικό στοιχείο για την εφαρμογή αυτού του τύπου της τεχνικής είναι, η δημιουργία του λεξικού μέσω της επισήμανσης των λέξεων ή των φράσεων με συναισθηματική πολικότητα και η υποκειμενικότητα αυτών.

Στα πρώτα έργα, οι [HM97] παρουσιάζουν μια προσέγγιση βασισμένη σε λεξικολογικές ευριστικές τεχνικές. Η τεχνική τους είναι στηριγμένη στο γεγονός ότι για κατηγοριοποίηση συναισθήματος μπορούμε να βρούμε σχέσεις πολικότητας χρησιμοποιώντας συνδετικές λέξεις όπως το “και” (and) ή το “αλλά” (but). Σε άλλες εργασίες οι “λέξεις σπόροι” (seed words) των οποίων η πολικότητα είναι γνωστή παρέχονται. Έτσι μπορούμε με χρήση αυτών να βρίσκουμε σχέσεις μεταξύ λέξεων μέσα από τα λεξικά. Επίσης, για τη χρήση λεξικών συναισθήματος δύο σημαντικές εργασίες έγιναν από τους Esuli και Sebastiani 2006 [ES06], Taboada, Anthony, και Voll 2006 [TAV06]. Μέσω της λεξικολογικής προσέγγισης οι Hu and Liu 2004, Kim and Hovy, 2004; Ding et al., 2008; Taboada, et al., 2010 καθόρισαν τη συναισθηματική πολικότητα μέσω από κάποια συνάρτηση που χρησιμοποιεί λέξεις γνώμης (opinion words) που αφορούν σε κάποιο κείμενο ή σε κάποια πρόταση.

Παρακάτω παρουσιάζεται σε σχήμα ένα σύστημα ανάλυσης συναισθήματος με χρήση λεξικών:



Σχήμα 3.1: Σύστημα ανάλυσης συναισθήματος με χρήση λεξικού

3.2 Πίνακας χρήσεων λεξικών

Τα λεξικά συναισθήματος τα συναντάμε πάρα πολύ συχνά σε κάθε είδους προσπάθεια για ανάλυση συναισθήματος, τόσο σε μη επιβλεπόμενη μάθηση, όσο και σε επιβλεπόμενη μηχανική μάθηση και σε υβριδικές μεθόδους˙ έννοιες που θα εξεταστούν παρακάτω. Τα συναντάμε να χρησιμοποιούνται είτε αυτούσια είτε σε συνδυασμό με άλλα εργαλεία για ανάλυση συναισθήματος. Παρακάτω παρουσιάζεται ένας πίνακας που παρουσιάζει μερικά λεξικά που συναντήσαμε στη βιβλιογραφία, και το πώς χρησιμοποιήθηκαν. Πρέπει να τονιστεί σε αυτό το σημείο ότι, σε πολλές εργασίες διακρίνουμε τη δημιουργία εκ νέου λεξικών τα οποία είναι προσαρμοσμένα στις ανάγκες της εκάστοτε περίπτωσης, καθώς το πεδίο στο οποίο ενδιαφέρονται να εξάγουν το συναίσθημα καθορίζει και το τελικό λεξικό που θα χρειαστεί να χρησιμοποιηθεί.

Συγγραφείς/Άρθρο	**Λεξικό**	**Χρήση Λεξικού**
[KWM11]	MPQA	Χρησιμοποιούν την πολικότητα των λέξεων, δημιουργώντας 3 χαρακτηριστικά (features) βασισμένα στη παρουσία

		κάποιας λέξης στο λεξικό.
[BF10]	MPQA	Ο βαθμός υποκειμενικότητας και πολικότητας της εκάστοτε λέξης.
[OFM13]	WordNet, SentiWordNet	Υπολογίζουν τη συναισθηματική πολικότητα κάθε λέξης.
[JYZ+11]	General Inquirer	Υπολογίζουν και χαρακτηριστικά ανεξάρτητα του στόχου μέσω μονογραμμάτων και του λεξικού.
[MGL09]	IBM lexicon (India Research Labs)	Κατηγοριοποίηση με βάση το λεξικό.
[SHF+14]	Thelwall-Lexicon	Σαν πείραμα εφάρμοσαν την προσέγγιση τους χρησιμοποιώντας το συγκεκριμένο λεξικό που θεωρούν ένα από τα λεξικά αιχμής για κοινωνικά δίκτυα.
[HBB13]	SentiWord- Net, WordNet	Τα ρήματα και τα επίθετα τα λάβανε από το WordNet. Τα Senti- features εξάχθηκαν από το SentiWordNet.
[BBP+12]	WordNet-Affect, SentiWordNet, multi-lingual and Italian computational lexicons	Για να αναγνωρίσουν ποίες από τις ετικέτες (tags) έχουν μια χρήσιμη τιμή με συναισθηματικό περιεχόμενο. Αναγνώριση λέξεων με άμεσο και έμμεσο συναισθηματικό περιεχόμενο.
[SFH+14]	SentiWord- Net lexicon, MPQA subjectivity lexicon, Thelwall-Lexicon	Αναγνώριση συναισθηματικής πολικότητας λέξεων.
[HTG+13]	MPQA Opinion Corpus 7 (MPQA)	Συναισθηματική πολικότητα λέξεων που εμφανίζονται μαζί με άλλες και έχουν μεγάλη πιθανότητα να έχουν ίδια συναισθηματική κατεύθυνση.
[ZGD+11]	Ding et al., 2008 lexicon	Κράτησαν το λεξικό από τους συγγραφείς Ding et al., 2008 και το ενίσχυσαν με opinion hashtags από το Twitter.
[BS10]	SentiWord- Net	Σαν μια βάση για δυαδική κατηγοριοποίηση χρησιμοποίησαν την πολικότητα από το λεξικό (positve/negative).
[OBR+10]	OpinionFinder lexicon, (Wilson, Wiebe, and Hoffmann 2005), MPQA	Βαθμολογία συναισθήματος για κάθε λέξη που εξάγεται από το λεξικό.

Πίνακας 3.2: Λεξικά και χρήσεις τους

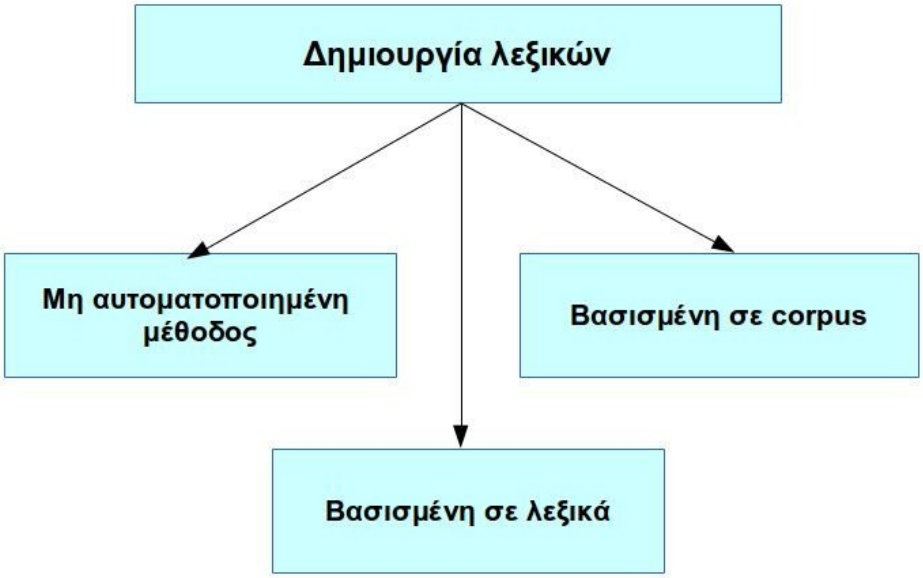
Το συμπέρασμα που βγαίνει είναι ότι μερικά από τα πιο συνηθισμένα λεξικά που χρησιμοποιούνται είναι το MPQA, OpinionFinder Lexicon, General Inquirer, και το SentiWordNet.

3.3 Δημιουργία λεξικών

Ο πιο αντιπροσωπευτικός τρόπος για να εκτελεστεί η μη επιβλεπόμενη μηχανική μάθηση για την ανάλυση συναισθήματος είναι η μέθοδος που βασίζεται σε λεξικό (lexicon-based method). Οι μέθοδοι αυτοί βασίζονται σε ένα προκαθορισμένο λεξικό συναισθημάτων για να καθορίσουν τη γενική πολικότητα συναισθήματος ενός συγκεκριμένου κειμένου.

Ένας από τους πιο κοινούς τρόπους ερευνητικής προσέγγισης στο πεδίο της ανάλυσης συναισθήματος (sentiment analysis), είναι η δημιουργία λεξικών πόρων αποτελούμενων από όρους που εκφράζουν άποψη, θετική ή αρνητική (opinion words). Πρόκειται για λέξεις-φράσεις προκαθορισμένου σημασιολογικού προσανατολισμού (semantic orientation) ως προς το υποκειμενικό τους περιεχόμενο, που λειτουργούν ως βασικοί πυρήνες κατά τις εφαρμογές κατά την ανάλυση κειμένου.

Μπορούμε να εντοπίσουμε κάποιες βασικές μεθοδολογικές προσεγγίσεις για τη δημιουργία τέτοιων λεξικών:



Σχήμα 3.3: Προσεγγίσεις για τη δημιουργία τέτοιων λεξικών

- Οι μέθοδοι βασισμένες σε λεξικό (dictionary-based methods), χρησιμοποιούν λεξικό για παράδειγμα το WordNet, για να προσδιορίσουν τον προσανατολισμό συναισθήματος μιας λέξης από σημασιολογικά / γλωσσικά σχετικές λέξεις. Αυτό λέγεται bootstrapping, και αφορά στη συλλογή όρων για τους οποίους γνωρίζουμε ήδη το συναισθηματικό τους προσανατολισμό. Η τεχνική αυτή παρουσιάζει το μειονέκτημα της αδυναμίας εντοπισμού των ιδιαίτερων σημασιολογικών αποχρώσεων που μπορεί να λαμβάνει ένας όρος κατά τη χρήση του.

- Η μέθοδος των σωμάτων κειμένων (corpus-based methods), οι οποίες συμπεραίνουν το συναισθηματικό προσανατολισμό για τις λέξεις από ένα συγκεκριμένο ηλεκτρονικό σώμα (corpus). Η μέθοδος αυτή διερευνά τη σχέση ανάμεσα στις λέξεις και κάποια παρατηρούμενη λέξη συναισθήματος που ονομάζεται (seed) σπόρος / πληροφορίας, και στη συνέχεια οικοδομεί ένα λεξικό συναισθήματος πάνω σε αυτή. Οι προσεγγίσεις αυτές για να εξάγουν τους κανόνες λαμβάνουν υπόψη το συντακτικό και τα συμφραζόμενα. Λειτουργεί βασισμένη σε μια αρχική ομάδα λέξεων, συνήθως επίθετα, προσπαθώντας να τις εμπλουτίσει με σχετιζόμενες λέξεις που συνήθως έχουν ανάλογο σημασιολογικό προσανατολισμό και συναισθηματική πολικότητα.
- Η μη αυτοματοποιημένη αναζήτηση και συλλογή όρων, η οποία, αν και χρονοβόρα, συνήθως συνδυάζεται με τις αυτοματοποιημένες μεθόδους για τη διόρθωση λαθών των τελευταίων. Πρόκειται για μία μέθοδο που χρησιμοποιεί μια ομάδα ανθρώπινων σχολιαστών να ονομάσει χειροκίνητα ένα σύνολο λέξεων για να χτίσει το λεξικό συναισθήματος, π.χ., General Inquirer και MPQA .

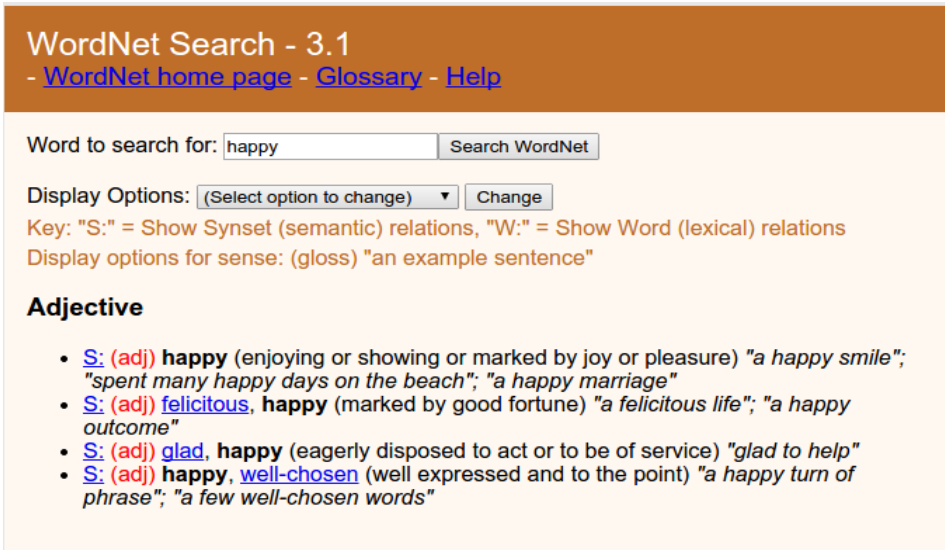
3.4 Λεξικά

Χαρακτηριστικά παραδείγματα λεξικών συναισθημάτων, και λεξιλογικών βάσεων είναι το WordNet, SentiWordNet, Harvard General Inquirer, το Linguistic Inquiry and Word Counts (LIWC), το Bing Liu's Opinion Lexicon, το MPQA Subjectivity Lexicon και το Affective Norms for English Words (ANEW) που παρουσιάζονται εν συντομία παρακάτω.

3.4.1 WordNet

Το WordNet [Fel98] είναι μια λεξικολογική βάση δεδομένων αγγλικών λέξεων. Δημιουργήθηκε το 1986 από το Πανεπιστήμιο του Princeton στο οποίο συνεχίζει να αναπτύσσεται. Βασίζεται σε ψυχολογικές και γλωσσολογικές θεωρίες για τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου και στόχος της δημιουργίας του ήταν να αποτελέσει ένα συνδυασμό λεξικού με γλωσσολογικό θησαυρό ώστε να χρησιμοποιηθεί ως ένα εργαλείο αυτόματης ανάλυσης κειμένου. Πιο συγκεκριμένα, το WordNet ομαδοποιεί τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα σε σύνολα συνωνύμων (synsets) κάθε ένα από τα οποία αντιπροσωπεύει μια διακριτή λεξικολογική έννοια. Παρέχει μικρούς ορισμούς και παραδείγματα χρήσης και ένα αριθμό από σχέσεις μεταξύ συνωνύμων. Επιπλέον, παρέχει έναν αριθμό από έννοιες. Η έννοια μιας λέξης του WordNet αποτελείται από: Ένα αριθμό που σηματοδοτεί τη συχνότητα εμφάνισης του όρου με τη συγκεκριμένη έννοια στα γλωσσολογικά κείμενα που έχουν χρησιμοποιηθεί από το WordNet. Βάση αυτού μπορεί να προκύψει η πιο “δημοφιλής” έννοια για κάθε λέξη (Most Frequent Sense ή First Sense) η οποία χρησιμοποιείται συχνά ως μια γρήγορη εναλλακτική της αποσαφήνισης. Ένα σύνολο

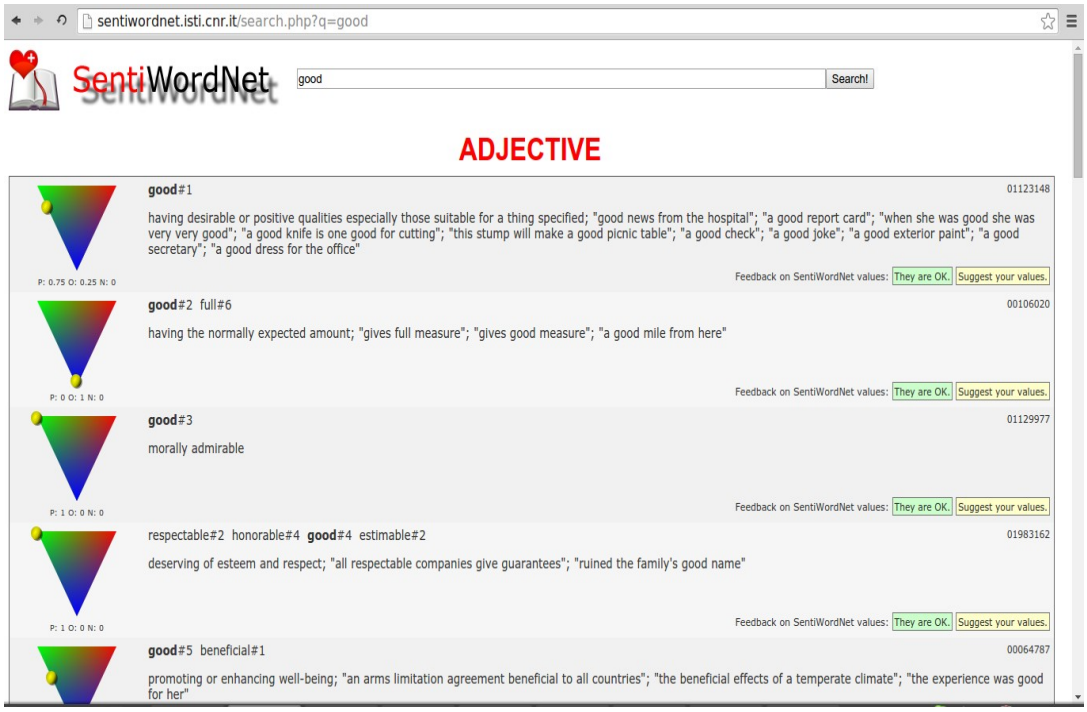
λέξεων που σηματοδοτεί τα συνώνυμα (synsets) του συγκεκριμένης ερμηνείας της λέξης. Ένα σύνολο από φράσεις της καθομιλουμένης που περιέχουν τη λέξη με τη συγκεκριμένη έννοια.



Εικόνα 3.3: Παράδειγμα λειτουργίας WordNet, Πηγή [9]

3.4.2 SentiWordNet

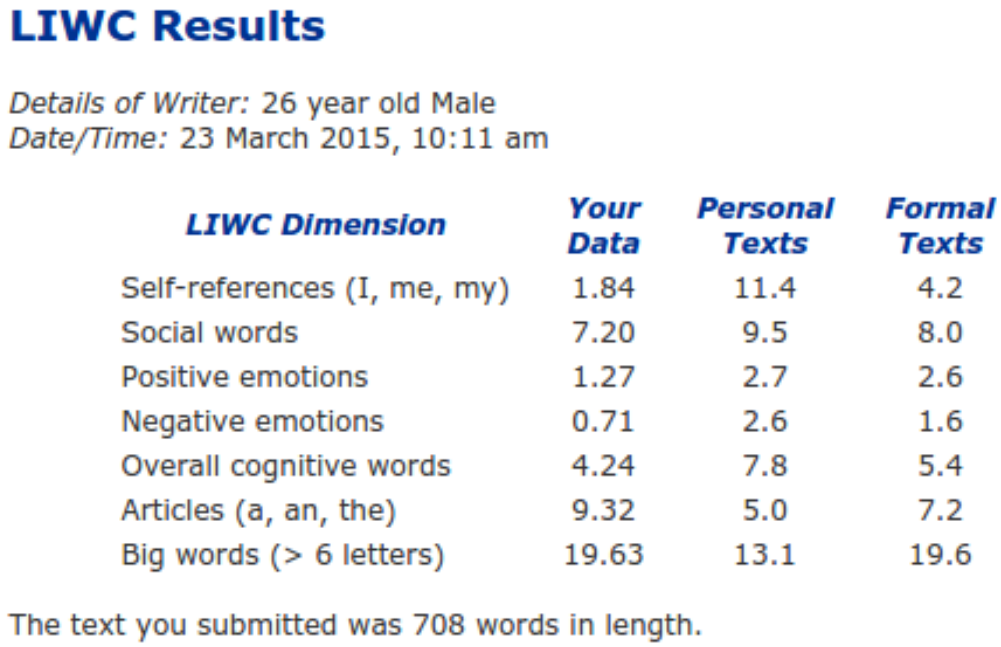
Το SentiWordNet πρώτο-ξεκίνισε από τους [ES06], είναι ένας λεξιλογικός πόρος ελεύθερης πρόσβασης, ο οποίος προορίζεται για την υποστήριξη συστημάτων ταξινόμησης συναισθήματος και εξόρυξης δεδομένων. Αναπτύχθηκε το 2010 από τους Baccianella Stefano, Andrea Esuli και Fabrizio Sebastiani [BES10]. Χρησιμοποιεί ως βασικό δομικό στοιχείο τα σύνολα συνωνύμων (synsets) της λεξιλογικής βάσης δεδομένων WordNet. Το SentiWordNet είναι το αποτέλεσμα της αυτόματης επισήμανσης των συνόλων συνωνύμων του WordNet σύμφωνα με τις έννοιες του θετικού, του αρνητικού και ουδέτερου, που περιγράφουν τους όρους που το απαρτίζουν. Οι τρεις βαθμολογίες, παράγονται από το συνδυασμό των αποτελεσμάτων από ένα σύνολο οκτώ τριμερών ταξινομητών, χαρακτηρίζονται από παρόμοια επίπεδα ακρίβειας, αλλά παρουσιάζουν διαφορετική συμπεριφορά ταξινόμησης. Η χρήση των συνόλων συνωνύμων (synsets) και όχι των ίδιων των όρων ως βασικών μονάδων ανάπτυξης του λεξικού προσφέρει τη δυνατότητα διερεύνησης των ποικίλλων σημασιολογικών αποχρώσεων του ίδιου όρου καθώς και των ιδιοτήτων που λαμβάνουν σε σχέση με την έκφραση γνώμης. Οι τιμές που υποδεικνύουν το συναίσθημα κάθε συνόλου συνωνύμων (synsets) κυμαίνονται στο διάστημα [0.0, 1.0] και το άθροισμα τους θα πρέπει να είναι πάντα ίσο με 1. Παρακάτω ακολουθεί ένα παράδειγμα χρήσης της εφαρμογής του SentiWordNet.



Εικόνα 3.4: Παράδειγμα λειτουργίας SentiWordNet, Πηγή [10]

3.4.3 Linguistic Inquiry and Word Count

Το λεξικό LIWC2007 είναι μια εφαρμογή του LIWC που διατίθεται σαν βάση δεδομένων που περιέχει μεγάλο αριθμό κατηγοριοποιημένων κανονικών εκφράσεων. Αναπτύχθηκε από τους James W. Pennebaker, Roger J. Booth, and Martha E. Francis , 2007 [PBF07]. Και υποστηρίζει 82 γλώσσες. Το λεξικό αυτό αποτελείται από σχεδόν 4.500 λέξεις και ρίζες λέξεων που προέρχονται από ψυχολογικές και διανοητικές καταστάσεις αλλά και σε γραμματικά, δομικά και γλωσσικά στοιχεία όπως το μέρος του λόγου, το μήκος λέξεων, τα επιφωνήματα. Η λειτουργία του λεξικού αυτού είναι η εξέταση κάθε λέξης χωριστά, όταν βρεθεί η λέξη που αντιστοιχεί σε κάποια κατηγορία του λεξικού, τότε αυξάνεται ο μετρητής του εκάστοτε χαρακτηριστικού. Το LIWC2007 έχει κατορθώσει να πετύχει πολύ μεγάλο εύρος από κατηγορίες και πολύ μεγάλο πλήθος λέξεων. Στη συνέχεια το λεξικό κάνει συλλογή λέξεων (Word Collection), όπου γίνετε η δημιουργία και η προσθήκη λημμάτων. Ακολουθεί η φάση βαθμονόμησης των αξιολογητών, με την οποία πραγματοποιείται η βαθμολόγηση και αξιολόγηση κάθε κατηγορίας χρησιμοποιώντας τρεις κριτές. Επιπλέον, η ψυχομετρική αξιολόγηση, όπου αντικαθίστανται οι σπάνιες κατηγορίες με νέες, ανάλογα με την ανάλυση των κειμένων. Και τέλος, οι προσθήκες, που αναδιαμορφώνεται η δομή του λεξικού, επικεντρώνοντας στο γραπτό και προφορικό λόγο. Ακολουθεί παράδειγμα εφαρμογής από το LIWC με παρουσίαση των αποτελεσμάτων.



LIWC Results

Details of Writer: 26 year old Male
Date/Time: 23 March 2015, 10:11 am

LIWC Dimension	Your Data	Personal Texts	Formal Texts
Self-references (I, me, my)	1.84	11.4	4.2
Social words	7.20	9.5	8.0
Positive emotions	1.27	2.7	2.6
Negative emotions	0.71	2.6	1.6
Overall cognitive words	4.24	7.8	5.4
Articles (a, an, the)	9.32	5.0	7.2
Big words (> 6 letters)	19.63	13.1	19.6

The text you submitted was 708 words in length.

Εικόνα 3.5: Παράδειγμα λειτουργίας LIWC,Πηγή [11]

3.4.4 Multi Perspective Question Answering Subjectivity Lexicon (MPQA)

Το λεξικό MPQA Subjectivity Lexicon, δημιουργήθηκε από τους Theresa Wilson, Janyce Wiebe, Paul Hoffmann, 2005, [WWC05]. Έχει σχέση με το συστήματος εξόρυξης γνώμης Opinion Finder. Η λογική λειτουργίας του είναι η επέκταση μιας λίστας από στοιχεία υποκειμενικότητας με χρήση γλωσσολογικών πόρων. Ξεκινώντας τη διαδικασία γίνετε ομαδοποίηση αναλόγως την αξιοπιστία (reliability) τους και μετέπειτα γίνετε χαρακτηρισμός σε θετικές, αρνητικές, ουδέτερες ή σε θετικές και αρνητικές ταυτόχρονα. Χρησιμοποιείτε μια λίστα από 8221 υποκειμενικών στοιχείων. Ύστερα δίνονται πληροφορίες για τον τύπο υποκειμενικότητας, το μήκος του κάθε στοιχείου σε λέξεις , η μορφή που εμφανίζει κάθε λέξη, το μέρος του λόγου που ανήκει, και μορφολογικό χαρακτήρα του και την προϋπάρχουσα πολικότητά του για κάθε στοιχείο.

3.4.5 Bing Liu's Opinion Lexicon

Το Bing Liu's Opinion Lexicon προτάθηκε από τους Minqing Hu and Bing Liu, 2004, [HL04]. Είναι ένα λεξικό που διαθέτει 6789 λέξεις οι οποίες είναι χωρισμένες σε θετικές και αρνητικές. Αφορά την αγγλική γλώσσα. Αποτελείτε από δύο λίστες, μια λίστα των 2006 θετικών και μια άλλη, των 4783 αρνητικών λέξεων που συνήθως εκφράζουν κάποιο συναίσθημα ή κάποια γνώμη. Το λεξικό αυτό διαθέτει πολλές μορφολογικές παραλλαγές και παραφθορές λέξεων και εκφράσεων που συναντάμε συχνά στην καθημερινή ομιλία και στις

συνομιλίες των ανθρώπων στα κοινωνικά δίκτυα, αλλά δεν είναι απαραίτητο ότι θα εντοπίσει με απόλυτη επιτυχία την ακρίβεια στην εξόρυξη γνώμης.

3.4.6 General Inquirer

Το Harvard General Inquirer αναπτύχθηκε από Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, 1966 . Διαθέτει δύο βασικές κατηγορίες , αυτή των 1915 θετικών και αυτή των 2291 αρνητικών λέξεων. Πρόκειται για ένα λεξικό το οποίο επισημειώνει πληροφορίες για κάθε λέξη, όπως για συντακτικό, σημασιολογία, μέρος του λόγου κ.α. Με το λεξικό θέλουμε να κάνουμε ανάλυση κειμένου χρησιμοποιώντας 182 κατηγορίες, που αντιπροσωπεύουν ένα σύνολο από λέξεις και ερμηνείες. Για το λόγω αυτό χρησιμοποιούνται δύο μικρότερα λεξικά για να συμπεριλάβουν όλες τις κατηγορίες. Το πρώτο είναι το HarvardIV-dictionary που έχει τις λέξεις που δείχνουν χαρά, πόνο, αρετή, λέξεις που φανερώνουν υπερβολή ή εγκράτεια, κ.α. και δεύτερο το Lasswell value dictionary, που διακρίνει τα λήμματα σε τέσσερις τομείς σεβασμού (ισχύς, ευθύτητα, υπόληψη, δεσμός) και σε τέσσερις τομείς ευημερίας (πλούτος, ευεξία, φώτιση και ικανότητα). Επιπρόσθετα υπάρχουν και δύο κατηγορίες μορφολογικών και σημασιολογικών δεικτών, όπως άρθρο, αριθμητικό, πρόθεση και αρσενικό, θηλυκό, χώρος και χρόνος.

3.4.7 Affective Norms for English Words (ANEW)

Το ANEW είναι μια πρόταση από τους Margaret M. Bradley and Peter J. Lang, 1999 [BL99]. Συμπληρώνοντας τη δουλειά των Center for Emotion and Attention (CSEA, University of Florida) με το υπάρχον Διεθνές Σύστημα Συναισθηματικής Εικόνας (International Affective Picture System) (IAPS, Lang, Bradley & Cuthbert, 1999) και το Διεθνές Σύστημα Συναισθηματικών Ψηφιοποιημένων Ήχων (International Affective Digitized Sounds) (IADS, Bradley & Lang, 1999), με συλλογές οπτικών και ηχητικών ερεθισμάτων αντίστοιχα, που περιλαμβάνουν αυτές τις συναισθηματικές αξιολογήσεις. Πρόκειται για ένα σύνολο συναισθηματικών αξιολογήσεων για ένα μεγάλο αριθμό λέξεων της αγγλικής γλώσσας, επιθυμεί να σχηματίσει ένα σύστημα λεκτικού υλικού το οποίο θα εξετάζει ως προς την ευχαρίστηση (pleasure), τη διέγερση (arousal) και την κυριαρχία (dominance). Η μεθοδολογία που ακολουθήθηκε εν συντομία είναι η εξής: Ξεκινάει με τη αξιολόγηση σε κλίμακες βαθμολόγησης κάθε συναισθηματικής κατάστασης (pleasure, arousal, dominance). Εκεί αξιολογήθηκαν 1040 λέξεις, επίθετα, ρήματα, ουσιαστικά. Δεύτερο βήμα, η αντιστοίχηση της μέσης τιμής σε κάθε κλίμακα, (pleasure, arousal, dominance).

3.5 Συμπεράσματα

3.5.1 Σύγκριση λεξικών

Όλα τα παραπάνω λεξικά παρέχουν βασική κατηγοριοποίηση πολικότητας. Οι λέξεις και τα λήμματα που περιέχονται στο καθένα από αυτά είναι διαφορετικά μεταξύ τους και για αυτό είναι δύσκολο να τα συγκρίνουμε και να αποφανθούμε για το ποιο είναι το καλύτερο. Μια σύγκριση που θα μπορούσαμε να κάνουμε και να βγάλουμε κάποια συμπεράσματα είναι πόσο συχνά διαφωνούν απόλυτα μεταξύ τους και παρέχουν ακριβώς αντίθετες τιμές για την πολικότητα κάποιας λέξης. Χαρακτηριστικά αποτελέσματα μας δείχνει ο παρακάτω πίνακας με στοιχεία από τα πέντε βασικά λεξικά.

Disagreement levels for the sentiment lexicons reviewed above.

	MPQA	Opinion Lexicon	Inquirer	SentiWordNet	LIWC
MPQA	–	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		–	32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–

Εικόνα 3.5: Σύγκριση μεταξύ λεξικών, Πηγή [12]

Για τις διαφορές μπορούμε να φανταστούμε τουλάχιστον δύο λογικούς λόγους. Ο πρώτος είναι ότι μπορεί να επιλύθηκε η εύρεση πολικότητας υπέρ κάποιας συγκεκριμένης έννοιας. Ο δεύτερος θα μπορούσε να είναι ότι συνδυάζουν κάποιες τιμές που προέρχονται από διατριβές-άρθρα, και έτσι δημιουργούν κάποιες συγκρούσεις ως προς τον τρόπο που δίνονται οι ερμηνείες και για αυτό δημιουργείται και ασάφεια.

3.5.2 Προβλήματα και ελλείψεις των μεθόδων με λεξικά

Σε ότι αφορά τις προσεγγίσεις με λεξικά αποτελούν μια δημοφιλή επιλογή γιατί δεν χρειάζονται εκπαίδευση, και επειδή είναι πιο κατάλληλες για μεγάλο εύρος περιεχομένων. Οι λεξικολογικές προσεγγίσεις χρησιμοποιούν την συναισθηματική κατεύθυνση των λέξεων που φέρουν γνώμη (opinion words) όπως “great, sad, excellent” που βρίσκεται στο δοσμένο κείμενο προκειμένου να υπολογίσουν το συνολικό συναίσθημα. Αντί να χρησιμοποιούν εκπαίδευση δεδομένων οι λεξικολογικές προσεγγίσεις στηρίζονται σε προκατασκευασμένα λεξικά από λέξεις με συγγενική συναισθηματική κατεύθυνση.

Η φύση της μεθόδου αυτής μπορεί να οδηγήσει σε λάθος αποτελέσματα και συμπεράσματα αν δεν προσαρμοστούμε σε περιπτώσεις συντακτικής και σημασιολογικής ανάλυσης. Πιο συγκεκριμένα μπορούμε να οδηγηθούμε σε λάθος συμπεράσματα αν δεν υπολογίσουμε, την

άρνηση, την ειρωνεία , την ένταση των λέξεων(Intensifiers), το θέμα, τη σειρά των λέξεων, τους ιδιωματισμούς γλώσσας. Για παράδειγμα οι παρακάτω προτάσεις πρέπει να εξεταστούν όχι μόνο μέσω κάποιου λεξικού, αλλά και η σημασιολογία και η σύνταξή τους για να βγάλουμε σωστά αποτελέσματα.

π.χ.(Άρνηση)

Bad.

Not bad.

π.χ.(Σειρά λέξεων)

You are right, I don't like ice cream.

You are not right, I do like ice cream.

Επιπροσθέτως, στις προσεγγίσεις με λεξικά παρατηρούμε και κάποια προβλήματα που προκαλούν σκεπτικισμό για τις μεθόδους αυτές.

- Ο αριθμός των λέξεων στα λεξικά είναι πεπερασμένος, αυτό προκαλεί προβλήματα όταν θέλουμε να εξάγουμε συναίσθημα από κάποιο δυναμικό μέσο, όπως το Twitter, όπου οι νέοι όροι, οι συντομογραφίες και οι δύσμορφες λέξεις υπάρχουν παντού στο κείμενο.
- Τα λεξικά συναισθήματος τείνουν να αναθέτουν μια καθορισμένη συναισθηματική κατεύθυνση και να δίνουν βαρύτητα στις λέξεις, ανεξάρτητα από το πώς αυτές οι λέξεις χρησιμοποιούνται μέσα στο κείμενο. Οι λέξεις μπορούν να εκφράσουν διαφορετικό συναίσθημα σε διαφορετικά περιεχόμενα. Για παράδειγμα η λέξη "great" πρέπει να είναι αρνητική σε περιεχόμενα για ένα "problem", και θετική σε περιεχόμενα για ένα "smile".

Για παράδειγμα τα emoticons, οι καθομιλούμενες εκφράσεις, οι συντομογραφίες, κτλ, χρησιμοποιούνται συχνά στα tweets. Αυτές οι εκφράσεις μπορεί να έχουν σημασιολογική/συναισθηματική πολικότητα αλλά δεν υπάρχουν σε κάποιο γενικό λεξικό γνώμης (opinion lexicon). Αυτό οδηγεί σε χαμηλή ανάκληση (recall) που είναι πρόβλημα για μεθόδους βασισμένες σε λεξικά (lexicon-based method), και εξαρτάται εξολοκλήρου από την παρουσία των λέξεων που φέρουν γνώμη (opinion words) που είναι απαραίτητες για να καθορίσουν την συναισθηματική πολικότητα.

Επίσης δεν μπορεί να δοθεί λύση με την προσθήκη αυτών στα λεξικά γνώμης (opinion lexicon) γιατί πρόκειται για ένα πολύ δυναμικό εργαλείο που οι εκφράσεις αυτές αλλάζουν συνέχεια και νέες εμφανίζονται ακολουθώντας την μόδα που επιβάλει το Διαδίκτυο. Επιπλέον, η συναισθηματική πολικότητα μπορεί να είναι άμεσα σχετιζόμενη από τη θεματολογία, αυτό το πρόβλημα είναι πολύ δύσκολο να το λύσει κάποιος μόνος του και να

κάνει τις απαραίτητες προσθήκες στο λεξικό γνώμης (opinion lexicon). Καταλήγουμε στο συμπέρασμα ότι χωρίς ένα ολοκληρωμένο λεξικό, η ανάλυση συναισθήματος δεν μας δίνει ικανοποιητικά αποτελέσματα.

Για τους λόγους αυτούς έχει δοθεί βαρύτητα και σε άλλες μεθόδους όπως η επιβλεπόμενη μηχανική μάθηση που παρουσιάζονται παρακάτω.

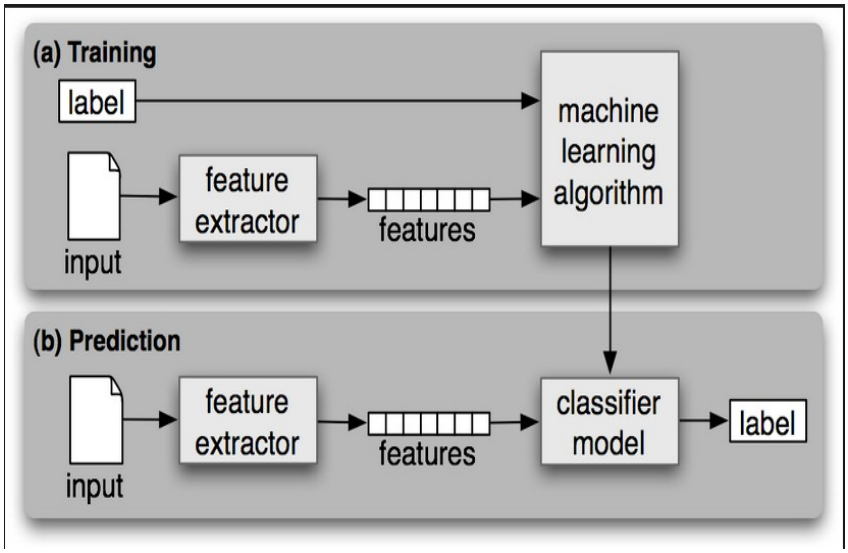
4 Τεχνικές επιβλεπόμενης μηχανικής μάθησης

Η επιβλεπόμενη μηχανική μάθηση (supervised machine learning) ή η μάθηση από παραδείγματα (learning from examples) αποτελεί τη πιο δημοφιλή τεχνική κατηγοριοποίησης (classification) συναισθήματος. Είναι η μάθηση που στηρίζεται στην κατηγοριοποίηση των αντικειμένων εισόδου, δηλαδή έχοντας ένα προκαθορισμένο σύνολο από κλάσεις, σκοπός μας είναι να τοποθετήσουμε τα αντικείμενα που εξετάζουμε σε κάθε μια από τις κλάσεις αυτές (πχ θετικά, αρνητικά, ουδέτερα).

Στην επιβλεπόμενη μάθηση κάθε κείμενο αναπαριστάται με ένα διάνυσμα χαρακτηριστικών έτσι ώστε ο ταξινομητής (classifier) να αναγνωρίσει και να μάθει τις πιο αντιπροσωπευτικές διαφορές ανάμεσα σε κείμενα που ανήκουν σε διαφορετικές κατηγορίες, για το λόγο αυτό πρέπει να χρησιμοποιηθούν σύνολα εκπαίδευσης. Μερικοί από τους πιο γνωστούς αλγορίθμους επιβλεπόμενης μάθησης είναι οι Naive Bayes, Maximum Entropy και Support Vector Machines (SVM).

Οι τεχνικές με χρήση της επιβλεπόμενης μηχανικής μάθησης έχουν πολύ καλά αποτελέσματα καθώς επιτυγχάνουν πολύ καλά ποσοστά ακρίβειας που υπερτερούν σε πολλές περιπτώσεις από τις τεχνικές μη-επιβλεπόμενης μάθησης, αλλά παρουσιάζουν και κάποια μειονεκτήματα. Απαιτούν μεγάλο χρόνο και προσπάθεια για την κατασκευή του συνόλου εκπαίδευσης του ταξινομητή, ώστε να καταφέρουμε να βρούμε τις καλύτερες τιμές ή να διαπιστώσουμε τους απαραίτητους κανόνες. Επίσης, η ακρίβεια εξαρτάται από το σύνολο εκπαίδευσης και για αυτό τα εκπαιδευτικά πρότυπα πρέπει να είναι αντιπροσωπευτικά του εκάστοτε πληθυσμού.

Στο παρακάτω σχήμα παρουσιάζεται η λογική λειτουργίας στη περίπτωση μιας επιβλεπόμενης μηχανικής μάθησης. Αναλυτικότερα η διαδικασία αυτή παρουσιάζεται παρακάτω και θα αναφερθούμε εν συντομία σε μερικούς αλγορίθμους επιβλεπόμενης μηχανικής μάθησης που συναντάμε περισσότερο.



Σχήμα 4.1: Μοντέλο επιβλεπόμενης μηχανικής μάθησης

Αρχικά, στη διάθεσή μας έχουμε το σύνολο των δεδομένων (dataset) για να υλοποιήσουμε τη μηχανική μάθηση και τη ανάλυση συναισθήματος. Αυτό το αρχικό σύνολο μπορεί να διαιρεθεί σε δύο κατηγορίες. Το σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης ή αλλιώς training set είναι το σύνολο που δίνεται σαν είσοδος στον ταξινομητή με σκοπό να το "μάθει". Το σώμα ή αλλιώς corpus είναι αυτό που το έχουμε χαρακτηρίσει σαν θετικό ή αρνητικό, κάτι που φυσικά είναι κάτι το υποκειμενικό. Το σύνολο εκπαίδευσης μπορούμε να πούμε ότι διαδραματίζει ένα πάρα πολύ σημαντικό ρόλο στη διαδικασία επιβλεπόμενης μάθησης διότι πρέπει να καταφέρουμε να βρούμε τις καταλληλότερες τιμές ώστε να διαπιστώσουμε τους κανόνες που χρειαζόμαστε κατά την ταξινόμηση. Επιπλέον, είναι σημαντικό να γνωρίζουμε ότι η διαδικασία επιλογής του συνόλου εκπαίδευσης θα επηρεάσει πολύ την ακρίβεια του αποτελέσματος και για αυτό πρέπει πάντα να είναι όσο το δυνατόν πιο αντιπροσωπευτικά του εκάστοτε πληθυσμού.

Το σύνολο ελέγχου (test set) είναι το σύνολο που θα χρησιμοποιηθεί για τον έλεγχο μετά την εκπαίδευση του ταξινομητή και πριν την πρόβλεψη πάνω στο αντικείμενο που μας ενδιαφέρει.

4.1 Προ-επεξεργασία

Η προ-επεξεργασία είναι το στάδιο κατά το οποίο προσαρμόζουμε το κείμενο (corpus) στην κατάλληλη μορφή πριν την εκπαίδευση του δικτύου. Η διαδικασία αυτή αφορά γενικά οποιαδήποτε κείμενα, αλλά λαμβάνει υπόψη της και τα ιδιαίτερα χαρακτηριστικά κειμένων όπως αυτά του Twitter που είναι 140 χαρακτήρες που μπορούν να περιέχονται αναφορές προς άλλους χρήστες (@username) και υπερσυνδέσμους. Η πιο συνηθισμένες λειτουργίες που εκτελούμε έτσι ώστε να φέρουμε το κείμενο στην κατάλληλη μορφή ώστε να μπορέσουμε να το επεξεργαστούμε με τις μεθόδους επιβλεπόμενης μηχανικής μάθηση είναι οι παρακάτω.

- Οι αναφορές προς άλλους χρήστες (@username) και υπερσύνδεσμοι, συχνά αφαιρούνται [PP10], [BS10] ή αντικαθίστανται με κατάλληλες λέξεις-κλειδιά (placeholders) [GBH09], [MKZ13], [KWM11].
- Επίσης, αντικαθίστανται τα hashtags με κατάλληλες λέξεις-κλειδιά [KWM11], [BS10] ή αφαιρούν τον χαρακτήρα "#" από τα hashtags [OFM13].
- Μια μέθοδος που αφορά τα emoticons, είναι η αφαίρεσή τους [GBH09], [PP10] ή να αντικαθιστούν τα emoticons με λέξεις συναισθήματος από ένα χειροκίνητα κατασκευασμένο λεξικό emoticons μέσω της Wikipedia και τις συντομογραφίες με την πλήρη μορφή τους, [OFM13].
- Επιπλέον, αφαιρούνται τα retweets, [OFM13], (δηλαδή το RT μπροστά από τη λέξη).
- Αφαιρούνται κοινές λέξεις [KWM11].
- Αφαιρούνται συνήθεις λέξεις [OFM13].
- Αφαιρούνται άρθρα (a, an, the stopwording), [KWM11], [PP10].
- Και τέλος, οι λέξεις άρνησης (no, not) συνενώνονται με την προηγούμενη ή επόμενη λέξη [PP10].

Έχοντας αφαιρέσει αυτή την περιττή πληροφορία, μένει το καθαρό κείμενο. Αλλά το μηχάνημα δεν μπορεί να ξεχωρίσει τι αποτελεί μια πρόταση, τι είναι λέξη, τι σημαίνουν τα σημεία στίξης κτλ. Χρειάζεται μια περαιτέρω επεξεργασία για να καθοριστούν όλα αυτά. Η διαδικασία αυτή καλείται tokenization. Είναι η κατάτμηση σε λεκτικές μονάδες (tokenization), δηλαδή η εξαγωγή όρων οι οποίοι αποτελούν λεκτικές μονάδες (tokens) από ένα κείμενο. Σαν "token" ορίζεται ένα συνεχόμενο πλήθος από χαρακτήρες ή νούμερα δηλαδή κάθε λέξη είναι ένα token. Τα διαφορετικά tokens χωρίζονται με κενά (whitespaces) ή με σημεία στίξης [PP10]. Μετά το tokenization εφαρμόζεται μια άλλη τεχνική, η λημματοποίηση που έχει σαν στόχο να αντιστοιχίζει τις διάφορες μορφές των λέξεων στο θέμα τους [JYZ+11], [OFM13]. Το θέμα δεν είναι απαραίτητο να είναι ταυτόσημο με την ρίζα της λέξης.

4.2 Χαρακτηριστικά – Features

Τα χαρακτηριστικά στα οποία βασίζεται ο ταξινομητής για να πάρει την απόφαση του ονομάζονται features. Διαφορετικά μπορούμε να πούμε πως features είναι ιδιότητες που βοηθούν στο να αναγνωριστεί μια οντότητα. Για παράδειγμα τα features της οντότητας "bicycle" είναι "bicycle handlebars", "wheels", "pedal" κτλ. Για να χρησιμοποιηθούν τα features στη πράξη πολλές φορές πρέπει να μετατραπούν στην κατάλληλη μορφή, ώστε να μπορέσουν να λειτουργήσουν ως είσοδος στον εκάστοτε ταξινομητή. Αυτό συμβαίνει γιατί στους ταξινομητές τα features αναπαριστώνται σαν διανύσματα ή σαν δυαδικά μεγέθη για παράδειγμα στο ταξινομητή Naive Bayes τα features που επιλέγονται είναι σε δυαδική μορφή ώστε ο ταξινομητής να μπορέσει να χρησιμοποιήσει σαν είσοδο αυτό το χαρακτηριστικό.

Μπορούμε να χωρίσουμε τα κυριότερα χαρακτηριστικά (features) σε δύο κατηγορίες, σε αυτά που αφορούν όλα τα είδη κειμένου και σε αυτά που αναφέρονται στο Twitter.

Για όλα τα είδη κειμένου εντάσσουμε τα εξής χαρακτηριστικά:

- Το βασικότερο χαρακτηριστικό (feature) όπως αναφέρθηκε και πιο πάνω, είναι η ίδια η λέξη το token που προκύπτει από τη διαδικασία του tokenization κατά τη φάση της προ-επεξεργασίας του κειμένου μας.
- Τα N-γράμματα (N-grams): Μονογράμματα, διγράμματα [PL08], [KWM11], συνδυασμό μονογραμμάτων και διγραμμάτων [GBH09], Λέξεις και χαρακτήρες ν-γραμμάτων (3, 4, 5) [MKZ13]. Έχει βρεθεί ότι η χρήση των SVM ταξινομητών με μοναδικό χαρακτηριστικό τα μονογράμματα αποφέρει το καλύτερα αποτελέσματα, ενώ η χρήση μόνο των διγραμμάτων οδηγεί σε χειρότερα αποτελέσματα εξαιτίας του αραιού χώρου χαρακτηριστικών (feature space) [GBH09]. Αντίθετα οι [PP10] συμπεραίνουν πως τα διγράμματα πετυχαίνουν τη καλύτερη ακρίβεια γιατί "αποτελούν μία καλή ισορροπία ανάμεσα στην κάλυψη του εύρους (αναφορικά με τα μονογράμματα) και στην ικανότητα αναγνώρισης συναισθηματικών μοτίβων έκφρασης (αναφορικά με τα τριγράμματα)". Επιπλέον οι [BS10] βλέπουν ότι η χρήση ν-γραμμάτων και του μέρους του λόγου (POS tags) βελτιώνει την ακρίβεια μόνο στην περίπτωση των μεγάλων κειμένων ενώ στα POS ν-γράμματα η επίλυση συνωνύμων (stemming) και η αφαίρεση κοινών λέξεων (stopwording) δεν οδηγούν σε καλύτερα αποτελέσματα.
- Η χρήση του Part-Of-Speech tagger: Το πρόβλημα της γραμματικής επισημείωσης (part-of-speech tagging) συνίσταται στην κατάταξη κάθε λέξης ενός κειμένου σε μια κατηγορία ανάλογα με το τι μέρος του λόγου είναι. Δεδομένης της φύσης του προβλήματος, οι λύσεις που έχουν προταθεί χρησιμοποιούν ταξινομητές διαφόρων τεχνολογιών. Η ανάγκη χρήσης ταξινομητή προκύπτει από το γεγονός ότι για τις λέξεις ενός κειμένου ως μονάδες δεν μπορεί να αναγνωριστεί μονοσήμαντα το μέρος του λόγου στο οποίο ανήκουν, λόγω πολλών αμφισημιών (π.χ. η λέξη "αποταμιεύσεις", ανάλογα με τα συμφραζόμενα, θα μπορούσε να λειτουργεί ως ρήμα ή ως ουσιαστικό). Οι part-of-speech taggers, κάποιες φορές, προχωρούν ένα βήμα παραπάνω από την ταξινόμηση λέξης στο σωστό μέρος του λόγου: υπό προϋποθέσεις και ανάλογα με τη γλώσσα στην οποία είναι γραμμένο το κείμενο, μπορούν να μαντέψουν και πρόσθετες ιδιότητες της λέξης, όπως γένος, πτώση, αριθμό κ.λπ. Αξίζει να αναφερθεί πως, όπως σε πολλά text information retrieval υποπροβλήματα, οι αποδοτικές υλοποιήσεις διαφέρουν από γλώσσα σε γλώσσα. Χρήση της ιδιότητας της κάθε λέξης (μέρος του λόγου) συναντάμε στις εργασίες [GBH09], [MKZ13], [PP10], [BS10], πλήθος και ποσοστό των κυριότερων POS tags στις [KWM11], και μέτα-χαρακτηριστικά (meta-features) POS tags [BF10]. "Σε αντίθεση με τα χαρακτηριστικά μικροιστολογίων που ήταν τα πιο χρήσιμα, τα POS tags οδηγούν σε

μείωση της ακρίβειας και δεν είναι μάλλον κατάλληλα για χρήση σε κείμενα από μικροιστολόγια" [KWM11].

- Τα σημεία στίξης οι [DTR10], [MKZ13]. Οι [DTR10] παρατηρούν ότι οι λέξεις, τα σημεία στίξης και τα εκφραστικά μοτίβα είναι τα πιο σημαντικά χαρακτηριστικά ενώ τα ν-γράμματα οδηγούν σε οριακή βελτίωση.
- Το πλήθος των θαυμαστικών οι [DTR10], [KS12], και η ύπαρξη αυτών από τον [BF10].
- Τα κεφαλαία γράμματα και λέξεις: [DTR10], [KS12], [BF10].
- Το πλήθος των επαναλαμβανόμενων γραμμάτων, [KS12].
- Η άρνηση ως χαρακτηριστικό. Η δοθείσα πολικότητα αντιστρέφεται από θετική σε αρνητική και αντίστροφα όταν μία άρνηση προηγείται της λέξης, [BF10]. Οι [GBH09] παρατήρησαν ότι προσθέτοντας την άρνηση (negation) ως ξεχωριστό χαρακτηριστικό δεν παρατηρείται βελτίωση των αποτελεσμάτων.

Για τα χαρακτηριστικά των tweets έχουμε:

- Τα Hashtags οι [MKZ13], [BF10], δυαδικά χαρακτηριστικά μικροιστολογίων για την ύπαρξη εξειδικευμένων όρων (hashtags, emoticons), [KWM11].
- Τα Emoticons [KS12], [MKZ13], [BF10].
- Το Μήκος του κάθε tweet [DTR10].
- Καθώς επίσης και retweets, υπερσυνδέσμους, συχνότητα κάθε χαρακτηριστικού κανονικοποιείται διαιρώντας με το πλήθος των όρων του κάθε tweet.

Μια ακόμα παράμετρος που μας απασχολεί σε ότι έχει να κάνει με τα χαρακτηριστικά (features) είναι το κριτήριο με το οποίο θα επιλεγούν τα τελικά features για να χρησιμοποιηθούν από τον ταξινομητή, μια από της δημοφιλέστερες τεχνικές για αυτό το σκοπό είναι το TF-IDF (Term Frequency- Inverse Document Frequency). Το TF-IDF αποτελείται από δυο ποσότητες, την συχνότητα του όρου (TF) και την αντίστροφη συχνότητα όρου (IDF). Η ποσότητα TF είναι απλώς η συχνότητα εμφάνισης του όρου t στο κείμενο d. Χρησιμοποιώντας μόνο την ποσότητα TF, οι καλύτεροι όροι θα ήταν είτε μη τετριμμένες λέξεις (όπως το "ή" , "και") είτε όροι που δεν φέρουν ιδιαίτερη πληροφορία (π.χ. η λέξη "δεδομένα" σε ένα κείμενο που μιλάει για βάσεις δεδομένων). Έτσι, η ποσότητα IDF λειτουργεί σαν ένα βάρος σημαντικότητας του όρου ως προς το κείμενο, σε σχέση με ολόκληρη την συλλογή κειμένων που εξετάζεται. Το μέτρο TF-IDF δίνει μεγάλο βάρος σε έναν όρο που εμφανίζεται συχνά σε ένα κείμενο αλλά σπάνια σε ολόκληρη την συλλογή, καθιστώντας τον χαρακτηριστικό όρο του κειμένου. Τελικά, σαν καλύτερα features επιλέγονται εκείνα που έχουν το μεγαλύτερο TF-IDF.

Οι [BF10] διαπιστώνουν ότι επειδή χρησιμοποιούν μια πιο αφηρημένη αναπαράσταση των δεδομένων και όχι μεμονωμένους όρους του, η προσέγγισή τους εμφανίζει μεγαλύτερη ανοχή στο θόρυβο και μεροληψία (bias) του συνόλου εκπαίδευσης σε σχέση με άλλες μεθόδους ενώ εμφανίζει καλύτερη συμπεριφορά ως προς την ικανότητας γενίκευσης όταν χρησιμοποιούνται σχετικά λίγα εκπαιδευτικά πρότυπα.

4.3 Ταξινομητές

Ο ταξινομητής (classifier) είναι αυτός ο οποίος με την βοήθεια των features θα λάβει τις κατάλληλες αποφάσεις. Γενικά ο ταξινομητής είναι ένα μαθηματικό εργαλείο το οποίο είναι υπεύθυνο στο να ταξινομήσει (να αναθέσει μια ετικέτα σε) μια δεδομένη είσοδο.

4.3.1 Ταξινομητής Naive Bayes

Ο Naive Bayes είναι ένας από τους πιο δημοφιλής ταξινομητές κειμένου. Ο ταξινομητής Naive Bayes είναι ένας απλός πιθανοτικός ταξινομητής που βασίζεται στην εφαρμογή του θεωρήματος Bayes και στόχος του είναι να ταξινομήσει ένα δείγμα X σε μια από τις προκαθορισμένες κατηγορίες. Αυτό που τον κάνει τόσο δημοφιλή είναι η απλότητα, η γραμμική χρονική πολυπλοκότητα και οι καλές του επιδόσεις σε προβλήματα ταξινόμησης. Για να μπορέσει να χρησιμοποιηθεί σε προβλήματα ταξινόμησης πρέπει πρώτα να εκπαιδευτεί, έτσι ώστε να αρχικοποιηθούν οι παράμετροί του. Η εκπαίδευση γίνεται με τα εκπαιδευτικά έγγραφα, δηλαδή έγγραφα τα οποία περιέχουν και την κατηγορία στην οποία ανήκουν. Μετά το τέλος της εκπαίδευσης, ο Naive Bayes είναι σε θέση να υπολογίσει την πιθανή κατηγορία στην οποία ανήκει ένα νέο-άγνωστο έγγραφο. Ο στόχος αυτός είναι ισοδύναμος με το να αναθέσει, ανάλογα με την κατηγορία στην οποία ταξινομείται, μια ετικέτα (label) στο δείγμα X. Ο όρος Naive επαφίεται στο γεγονός ότι η λειτουργία του ταξινομητή στηρίζεται σε μια ισχυρή παραδοχή ανεξαρτησίας. Η ισχυρή παραδοχή που κάνει, είναι ότι η παρουσία (ή μη) ενός feature σε μια κλάση είναι ανεξάρτητο από την παρουσία (ή μη) ενός άλλου. Αν και υπάρχουν περιπτώσεις που αυτό πράγματι συμβαίνει, δεν ισχύει πάντα. Παρότι αυτή η υπόθεση δεν ανταποκρίνεται στην πραγματικότητα, έχει αποδειχτεί ότι λειτουργεί αρκετά αποτελεσματικά. Το πλεονέκτημα της χρήσης αυτού του μοντέλου είναι η απλότητα του, καθώς και ότι απαιτεί μόνο ένα μικρό σχετικά training set. Συγκεκριμένα, ο Naive Bayes υπολογίζει ξεχωριστά τις πιθανότητες το νέο αυτό έγγραφο να ανήκει σε κάθε μια από τις δοσμένες κατηγορίες και στη συνέχεια δίνει ως κατηγορία κατάταξης, αυτή με τη μεγαλύτερη πιθανότητα. Για τον υπολογισμό των πιο πάνω πιθανοτήτων λαμβάνεται υπόψη τόσο η προγενέστερη πιθανότητα της κάθε κατηγορίας, όσο και η κάθε λέξη που περιέχεται στο νέο έγγραφο. Η όλη ιδέα του Naive Bayes είναι να επιλέξει την πιο πιθανή ετικέτα για μια είσοδο (posterior probability), όπου η εκ των προτέρων πιθανότητα (prior probability)

για κάθε ετικέτα είναι γνωστή. Από εκεί και πέρα, ανάλογα με την συνεισφορά του κάθε feature ξεχωριστά, υπολογίζεται το πόσο πιθανό είναι να λάβει την συγκεκριμένη ετικέτα ή άλλη.

4.3.2 Μηχανές Διανυσμάτων Υποστήριξης ή αλλιώς Support Vector Machines

Οι Μηχανές Διανυσμάτων Υποστήριξης ή αλλιώς SVM, πρωτάρχισαν από τον Vapnik και τους συνεργάτες του το 1963. Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVMs) είναι μια τεχνική η οποία ανήκει στην ομάδα των μηχανών εκμάθησης (learning machines) και ως στόχο έχει την επεξεργασία δεδομένων. Είναι μια μέθοδος κατηγοριοποίησης που μπορεί να χρησιμοποιηθεί τόσο για γραμμικά όσο και για μη-γραμμικά δεδομένα. Οι Μηχανές Διανυσμάτων Υποστήριξης είναι ένας δυαδικός, μη-πιθανοτικός ταξινομητής αφού προβλέπει, για κάθε στοιχείο της εισόδου, σε ποιες από τις δυο πιθανές κλάσης πρέπει να τοποθετηθεί. Στο SVM μοντέλο τα δεδομένα τα αναπαριστάμε σαν σημεία στον χώρο, είναι με τέτοιο τρόπο τοποθετημένα ώστε δεδομένα από διακριτές κατηγορίες να είναι διαχωρισμένες με όσο το δυνατόν με μεγαλύτερο κενό. Για να γίνει η πρόβλεψη γα την κατηγορία στην οποία ανήκει, ελέγχεται σε ποιο από τα δύο μέρη (στα οποία χωρίζει το χώρο το κενό) βρίσκεται το σημείο που την αναπαριστά. Η βασική ιδέα ενός SVM είναι να δημιουργήσει ένα υπερεπίπεδο ως την επιφάνεια απόφασης με τέτοιο τρόπο ώστε η απόσταση ανάμεσα στα θετικά και αρνητικά παραδείγματα να είναι το μέγιστο. Η μηχανή επιτυγχάνει αυτή την επιθυμητή ιδιότητα, ακολουθώντας μια αρχή που έχει ως βάση την θεωρία στατιστικής μάθησης.

Αρχικά, κατά την φάση της εκπαίδευσης, το SVM αντιστοιχεί τα αρχικά δεδομένα σε έναν χώρο υψηλών διαστάσεων (χώρο των χαρακτηριστικών) και με βάση αυτή την διάσταση, ψάχνει για γραμμικά διαχωριζόμενα υπερ-επίπεδα.

Όπως καταλαβαίνουμε, υπάρχουν άπειρα υπερ-επίπεδα (στην ουσία πρόκειται για γραμμές) που διαχωρίζουν τις δύο κλάσεις, αλλά όμως δεν είναι όλες βέλτιστες. Σε αυτή την φάση, στόχος είναι η εύρεση του βέλτιστου υπερ-επιπέδου δηλαδή αυτού που ελαχιστοποιεί το σφάλμα κατηγοριοποίησης στα άγνωστα δεδομένα.

Από την στιγμή που θα βρεθεί αυτό το μοναδικό ελάχιστο, για ένα δεδομένο σύνολο δεδομένων, η δοσμένη SVM θα συγκλίνει πάντα ντετερμινιστικά στην ίδια λύση (ανεξάρτητα από τις αρχικές υποθέσεις). Αυτό το υπερ-επίπεδο ονομάζεται maximum marginal hyperplane (MMH) και είναι εκείνο που μπορεί να διαχωρίσει τα δεδομένα στο χώρο των χαρακτηριστικών. Μετά την εκπαίδευση, το SVM μπορεί πλέον να αναθέτει νέα στοιχεία σε κάθε κατηγορία. Στους SVM δεν υπάρχει ο κίνδυνος της υπερκάλυψης (overfitting) αφού η χρήση του MMH οδηγεί σε έναν αλγόριθμο εκπαίδευσης που έχει μετασχηματιστεί (ή απλοποιηθεί) σε ένα πρόβλημα βελτιστοποίησης. Σε περίπτωση που οι δυο κατηγορίες είναι μη-διακριτές, ο SVM ψάχνει ένα επίπεδο που μεγιστοποιεί το περιθώριο και που ταυτόχρονα

ελαχιστοποιεί μια ποσότητα ανάλογη του αριθμού των σφαλμάτων κατηγοριοποίησης. Παρότι η εκπαίδευση μπορεί να είναι αργή, η ακρίβεια είναι υψηλή χάρη στην ικανότητα μοντελοποίησης σύνθετων, μη γραμμικών ορίων απόφασης.

Η λογική μιας μηχανής εκμάθησης είναι να δίνει την τιμή y μιας συνάρτησης (άγνωστη προς εμάς) που αντιστοιχεί σε δοσμένο σημείο διάνυσμα x. Αυτό γίνεται ως εξής: Για δεδομένο σύνολο l σημείων διανυσμάτων x ∈ R και έχοντας τις αντίστοιχες τιμές y ∈ R που παίρνει η άγνωστη συνάρτηση, εκπαιδεύουμε τη μηχανή εκμάθησης να μάθει τη σχέση που συνδέει τα διανύσματα x με τα y. Δηλαδή, η μηχανή μαθαίνει την αντιστοίχηση διανυσμάτων x→ y και έτσι για ένα σημείο x m , διαφορετικό από αυτά του συνόλου l της εκμάθησης, θα μας δώσει την τιμή y m που θα έπαιρνε η άγνωστη συνάρτηση.

Στην περίπτωση ταξινόμησης με τα SVM, το σύνολο των σημείων l αποτελείται από δύο υποσύνολα τα k και n . Έτσι, το αποτέλεσμα της συνάρτησης θα είναι +1 ή -1 (y i =+1 ή y i =-1) ανάλογα σε ποιο υποσύνολο ανήκει το δοθέν σημείο x i. Τα δύο αυτά υποσύνολα ονομάζονται κλάσεις και η τιμή +1 (-1) είναι η "ετικέτα" της κλάσης. Δηλαδή, σε αυτή τη περίπτωση τα SVM μαθαίνουν να κατατάσσουν σωστά τα σημεία x i στις δύο κλάσεις. Τα σημεία x i και οι αντίστοιχες τιμές τους, y i , αποτελούν την πληροφορία εκπαίδευσης (training set). Τα σημεία x i ονομάζονται πρότυπα εκπαίδευσης (training patterns) ενώ οι τιμές y i που αντιστοιχούν σε αυτά, στόχοι εκπαίδευσης (training targets).

4.3.3 Ταξινομητής μέγιστης εντροπίας (Maximum Entropy)

Ο ταξινομητής Maximum Entropy αποτελεί μια γενίκευση του Naive Bayes. Στον ταξινομητή Naive Bayes καθορίζεται μια παράμετρος για κάθε ετικέτα (την εκ των προτέρων πιθανότητα) και μια παράμετρος για κάθε ζευγάρι χαρακτηριστικό (feature) -ετικέτας (την συνεισφορά του κάθε χαρακτηριστικό (feature) προς το likelihood της ετικέτας). Αντίθετα, στον Maximum Entropy ο χρήστης είναι αυτός που καθορίζει ποιούς συνδυασμούς από ετικέτες και ποιά χαρακτηριστικά (features) πρέπει να έχουν τις δικές τους παραμέτρους. Αυτό ισχύει διότι είναι δυνατόν, χρησιμοποιώντας μια παράμετρο, να συσχετιστεί ένα χαρακτηριστικό (feature) με παραπάνω από μια ετικέτες ή να συσχετιστούν πολλά χαρακτηριστικά (features) με μια μοναδική ετικέτα. Σε αντίθεση με πριν, ο Maximum Entropy δεν υποθέτει την ανεξαρτησία των χαρακτηριστικά (features) και συνεπώς δεν χρησιμοποιούνται πιθανότητες για τον καθορισμό των παραμέτρων του μοντέλου. Επειδή δεν γίνονται υποθέσεις ανεξαρτησίας, δεν είναι δυνατόν να υπολογιστούν άμεσα όλες οι εξαρτήσεις για όλους τους συνδυασμούς από χαρακτηριστικά (features). Για αυτόν το λόγο, για να βρεθεί το σύνολο των παραμέτρων που θα μεγιστοποιούν την απόδοση του ταξινομητή, χρησιμοποιούνται επαναληπτικές μέθοδοι βελτιστοποίησης (iterative optimization techniques). Αυτές οι τεχνικές αναζητούν το σύνολο των παραμέτρων που μεγιστοποιεί το συνολικό likelihood του συνόλου εκπαίδευσης. Αρχικά αρχικοποιούνται οι

παράμετροι του μοντέλου με τυχαίες τιμές. Στην συνέχεια, επαναληπτικά, ανανεώνονται οι παράμετροι αυτές για να έρθουν πιο κοντά στις βέλτιστες τιμές. Αν και οι μέθοδοι βελτιστοποίησης εξασφαλίζουν ότι οι παράμετροι θα φτάσουν όντως στις βέλτιστες τιμές, δεν είναι δυνατόν να καθοριστεί το πότε θα φτάσουν σε αυτές, με συνέπεια η διαδικασία της εκπαίδευσης (ειδικά για μεγάλα σύνολα εκπαίδευσης) να διαρκεί αρκετό χρόνο. Κάθε συνδυασμός από ετικέτες και χαρακτηριστικά (features) που λαμβάνει δική του παράμετρο ονομάζεται joint-feature. Joint features είναι ιδιότητα των (labeled) τιμών με ετικέτα, ενώ τα απλά χαρακτηριστικά (features) είναι ιδιότητα των (unlabeled) τιμών χωρίς ετικέτα. Τα joint-features που χρησιμοποιούνται για να κατασκευαστεί ο Maximum Entropy είναι αντίστοιχα με αυτά του Naive Bayes. Δεδομένου των joint-features, η βαθμολογία που δίνεται σε μια ετικέτα για μια δεδομένη είσοδο είναι το γινόμενο των παραμέτρων που συσχετίζονται με τα joint features και εφαρμόζονται στην είσοδο και την ετικέτα. Ο ταξινομητής μέγιστης εντροπίας (MaxEnt) είναι στενά συνδεδεμένος με τον ταξινομητή Naive Bayes, έχει διαφορετικό στοιχείο ότι αντί να επιτρέπει σε κάθε χαρακτηριστικό (feature) να έχει λόγο ανεξάρτητα, το μοντέλο αυτό χρησιμοποιεί μια βελτιστοποίηση με βάση την αναζήτηση για να βρει τα βάρη για τα χαρακτηριστικά που μεγιστοποιούν την πιθανότητα των δεδομένων εκπαίδευσης. Τα χαρακτηριστικά που μπορούμε να ορίσουμε για ένα ταξινομητή Naive Bayes είναι εύκολο να χρησιμοποιηθούν τα ίδια και στον MaxEnt, αλλά στο MaxEnt μοντέλο μπορούμε να χειριστούμε μείγμα δεδομένων από ακέραιες τιμές, πραγματικές, λογικές και χαρακτηριστικά πραγματικών τιμών. Για κάθε λέξη w και class c ∈ C, καθορίζουμε ένα κοινό χαρακτηριστικό f(w, c)= N όπου N είναι το πλήθος των w που εμφανίζονται μέσα στο κείμενο στην κλάση c. (το πλήθος N θα μπορούσε να είναι επίσης λογική τιμή, μια εγγραφή για παρουσία ή απουσία).

Μέσω επαναληπτικής βελτιστοποίησης, εκχωρείτε ένα βάρος σε κάθε κοινό χαρακτηριστικό ούτως ώστε να μεγιστοποιηθεί η λογαριθμική πιθανοφάνεια των δεδομένων εκπαίδευσης.

Η πιθανότητα της κλάσης c δοθέντος του κειμένου d και των βαρών λ είναι:

$$P(c|d,\lambda) \stackrel{def}{=} \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c' \in C} \exp \sum_i \lambda_i f_i(c',d)}$$

4.4 Πρόβλεψη

Όταν τελειώσει η φάση της εκπαίδευσης του ταξινομητή, μπορεί το μηχανικό σύστημα να ξεκινήσει να ταξινομεί την εκάστοτε είσοδο που του δίνουμε. Στην περίπτωση αυτή ο ταξινομητής δέχεται ως είσοδο ένα σύνολο από tweets ή ένα κείμενο ελέγχου (test set) που

είναι διαφορετικό από το σύνολο εκπαίδευσης (training set) και ταξινομεί το καθένα στις προκαθορισμένες κατηγορίες. Σε περιπτώσεις περιορισμένου αριθμού δεδομένων (dataset) χρησιμοποιείται η μέθοδος n-fold cross validation, κατά την οποία τα δεδομένα χωρίζονται σε n ίσα υποσύνολα και στη συνέχεια με βάση αυτά δημιουργούνται n μοντέλα γνώσης, κάθε φορά ένα από τα n υποσύνολα αυτά εξαιρείται από το σύνολο εκπαίδευσης (training set), το οποίο χρησιμοποιείται ως το σύνολο ελέγχου (test set). Μετά ο ταξινομητής θα υπολογίσει την πιθανότητα να λάβει το κείμενο ή τα tweets μια ετικέτα. Η μεγαλύτερη πιθανότητα είναι αυτή που τελικά κρατάμε ως αποτέλεσμα του ταξινομητή.

Για να εξετάσουμε κατά πόσο η ταξινόμηση είναι σωστή έχουν αναπτυχθεί αρκετές μέθοδοι, όπως μετρικές για την ακρίβεια (Accuracy), την εγκυρότητα (Precision), την ανάκληση (Recall), το F-Measure που απαντάνε στο ερώτημα κατά πόσο ο ταξινομητής έχει ταξινομήσει σωστά.

#	Predicted positives	Predicted negatives
Actual positive instances	Number of True Positive instances (TP)	Number of False Negative instances (FN)
Actual negative instances	Number of False Positive instances (FP)	Number of True Negative instances (TN)

- $\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN}$
- $\text{Precision} = \frac{TP}{TP + FP}$
- $\text{Recall} = \frac{TP}{TP + FN}$
- $\text{F1} = \frac{2 \times \text{Pr}ecision \times \text{Re}call}{\text{Pr}ecision + \text{Re}call}$

Πίνακας 4.2: Μετρικές για ταξινόμηση

4.5 Πίνακας μεθόδων

Παρακάτω ακολουθεί ένας πίνακας για μεθόδους επιβλεπόμενης μηχανική μάθησης που συναντήσαμε στη βιβλιογραφία. Οι εργασίες αυτές κεντρίζουν το ενδιαφέρον λόγω των αλγορίθμων και των τεχνικών που χρησιμοποιούν (NB, SVM, MaxEnt), του συνόλου δεδομένων (Twitter, reviews, Facebook) που κάνουν χρήση, αλλά και της απόδοσης που επιτυγχάνουν. Επιλέχθηκαν με βάση το πλήθος εμφανίσεων σε σχετικά άρθρα (citations), που τα καθιστά θεμέλιο για το πεδίο της έρευνας αυτής. Αλλά επίσης, και με το ενδιαφέρον θέμα

τους και το σύγχρονο της κάθε εργασίας, που τα καθιστά την αιχμή της τεχνολογίας στον τομέα της ανάλυσης συναισθήματος.

Συγγραφείς	**Τίτλος**	**Έτος**	**Αλγόριθμος Μηχανικής Μάθησης**	**Σύνολο Δεδομένων (Dataset)**	**Απόδοση**
Pak, Paroubek [PP10]	Twitter as a corpus for sentiment analysis and opinion mining	2010	NB, SVM, CRF	Posts from Twitter	F 0.5 = 0.63
Bermingham, Smeaton [BS10]	Classifying sentiment in microblogs: is brevity an advantage?	2010	Naive Bayes, SVM	Twitter	NB και μονογράμματα ακρίβεια 74.85 %
Go, Bhayani, Huang [GBH09]	Twitter sentiment classification using distant supervision.	2009	NB, Maximum Entropy, SVM	Twitter	SVM με μονογράμματα 82.9 %
Mohammad, Kiritchenko, Zhu [MKZ13]	Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.	2013	SVM	Twitter	-
Jiang, Yu, Zhou, Liu, Zhao [JYZ+11]	Target- dependent twitter sentiment classification	2011	SVM	tweets	Επίδοση 68.2 %
Pang, Lee, Vaithyanathan [PLV02]	Thumbs up? Sentiment Classification using Machine Learning Techniques	2002	Naive Bayes, maximum entropy, SVM	Movie review	support vector machine είχε καλύτερη επίδοση. 81.5%
Tromp, Pechenizkiy [TP11]	SentiCorr: Multilingual sentiment analysis of personal correspondence.	2011	AdaBoost	Twitter, Facebook and Hyves , MS Outlook e-mail	-
Bifet, Frank [BF10.]	Sentiment knowledge discovery in twitter streaming data	2010	Naive Bayes	Twitter data stream, Stanford Twitter Sentiment, Edinburgh Twitter Corpus	82.45 %, 86.26 %

Chen, Ibekwe-SanJuan, SanJuan, Weaver [CIS+06]	Visual analysis of conflicting opinions	2006	Naïve Bayes, SVM	Reviews	77.5%
Gindl, Liegl [GL08]	Evaluation of different sentiment detection methods for polarity classification on web-Based reviews	2008	Naïve Bayes, Maximum Entropy	Reviews	83.8%
Agarwal, Xie, Vovsha, Rambow, Passonneau [AXV+11]	Sentiment analysis of twitter data	2011	SVM	Twitter	75.39 %, overall gain of over 4%
Saif, He, Alani [SHA11]	Semantic smoothing for twitter sentiment analysis	2011	Naive Bayes	Twitter	81.3%
Liu, Li, Guo [LLG12]	Emoticon smoothed language models for twitter sentiment analysis	2012	Γλωσσικό μοντέλο	Twitter	82.5%
Aston, Munson, Liddle, Hartshaw, Livingston, Hu [AML+14]	Sentiment analysis on the social networks using stream algorithms	2014	Massive Online Analysis (MOA) and Weka frameworks	Twitter	F-score 85%, 78%
Pandey, Iyer [PI09]	Sentiment analysis of microblogs	2009	BLR(Gauss), (Laplace) Naive Bayes SVM(Linear) SVM(Poly) SVM(RBF)	Twitter	Naive Bayes και SVM είναι οριακά καλύτερα από τους άλλους

Πίνακας 4.3: Πίνακας για επιβλεπόμενη μηχανική μάθηση

5 Υβριδικά μοντέλα

Τα κοινωνικά δίκτυα αλλά πιο συγκεκριμένα το Twitter που εστιάζεται περισσότερο η ανάλυση συναισθήματος στις μέρες μας, έχει κάποια μοναδικά χαρακτηριστικά όπως αναφέραμε παραπάνω, που το διαφοροποιούν σε σχέση με την ανάλυση συναισθήματος. Παρατηρούμε ότι μέχρι τώρα η ανάλυση συναισθήματος που εφαρμοζόταν κυρίως σε κριτικές ιστοτόπων αφορούσε περισσότερο συναισθήματα που είχαν να κάνουν με προϊόντα ή ταινίες ή εταιρείες κτλ. Οι [KS12] αναφέρουν ότι στο Twitter τα μηνύματα απαιτούν μια διαφορετική και πιο απαιτητική επεξεργασία για την εύρεση του συναισθήματος, γιατί τα tweets δίνουν μια πιο πλούσια και μεγαλύτερης ποικιλίας πηγή πληροφορίας και συναισθήματος από διάφορες πηγές που αφορά διάφορους παραλήπτες. Για αυτό δίνεται μεγάλη έμφαση στο Twitter έτσι ώστε να το αξιοποιήσουμε στο μέγιστο δυνατό καθώς μπορούμε μέσω του κειμενικού σώματος (corpus) να προσεγγίσουμε διαφορετικές διαστάσεις και ποικίλες εφαρμογές.

5.1 Επιβλεπόμενη μηχανική μάθηση μαζί με λεξικολογικές προσεγγίσεις

Οι περισσότερες προσεγγίσεις για την αναγνώριση του συναισθήματος των tweets μπορούν να κατηγοριοποιηθούν σε δύο κατηγορίες, προσεγγίσεις μηχανικής μάθησης, και προσεγγίσεις βασισμένες σε λεξικά. Μερικές από αυτές τις μεθόδους τείνουν να πετύχουν καλύτερη και σταθερού επιπέδου ακρίβεια όταν εφαρμόζονται σε σύνολα δεδομένων (datasets) που τα γνωρίζουμε καλά, όπου τα δεδομένα με ετικέτες είναι διαθέσιμα για εκπαίδευση ή όταν το κείμενο που έχουμε αναλύσει έχει καλυφθεί καλά από το λεξικό συναισθημάτων.

1. **Προσεγγίσεις με λεξικά.**

Οι τεχνικές βασισμένες σε λεξικά (lexicon-based) μπορούμε να πούμε ότι είναι βολικές διότι:

- δεν χρειάζονται εκπαίδευση

- είναι πιο κατάλληλες για μεγάλο εύρος περιεχομένων μιας και στηρίζονται σε προκατασκευασμένα λεξικά από λέξεις με συγγενική συναισθηματική κατεύθυνση.

Αλλά έχουν και σημαντικά μειονεκτήματα:

- Ο αριθμός των λέξεων στα λεξικά είναι πεπερασμένος.
- Τα λεξικά συναισθήματος αναθέτουν μια καθορισμένη συναισθηματική κατεύθυνση και δίνουν βαρύτητα στις λέξεις, ανεξάρτητα από το πώς αυτές οι λέξεις χρησιμοποιούνται μέσα στο κείμενο.
- Δεν γίνεται να προστεθούν λέξεις που φέρουν γνώμη (opinion words) που είναι απαραίτητες για να καθορίσουν την συναισθηματική πολικότητα και να υπάρξει λύση για την πληρότητα των λεξικών γνώμης (opinion lexicon) γιατί πρόκειται για ένα πολύ δυναμικό εργαλείο που οι εκφράσεις αυτές αλλάζουν συνέχεια.

2. Προσεγγίσεις επιβλεπόμενης μηχανικής μάθησης.

Σε ότι αφορά τις επιβλεπόμενες προσεγγίσεις μηχανικής μάθησης έχουν πολύ καλά αποτελέσματα καθώς επιτυγχάνουν πολύ καλά ποσοστά ακρίβειας που υπερτερούν σε πολλές περιπτώσεις από τις τεχνικές μη-επιβλεπόμενης μάθησης. Οι τεχνικές επιβλεπόμενης μάθησης είναι βασισμένες στο να εκπαιδεύσουν ταξινομητές από διάφορα χαρακτηριστικά όπως n-grams, Part-Of-Speech (POS) tags, και συντακτικά χαρακτηριστικά των tweets (π.χ. hashtags, retweets, punctuations, κτλ.). Αυτές οι μέθοδοι μπορούν να πετύχουν 80%-84% σε ακρίβεια (accuracy).

Τα μειονεκτήματα που έχουν αυτές οι μέθοδοι είναι :

- Για να εκπαιδεύσει κάποιος τα δεδομένα είναι κάτι ακριβό , και ιδίως όταν τα δεδομένα είναι συνεχώς μεταβαλλόμενα όπως σε ένα δυναμικό σύστημα σαν το Twitter.
- Επιπλέον οι ταξινομητές που εκπαιδεύονται σε δεδομένα ενός συγκεκριμένου πεδίου μπορεί να παράγουν χαμηλές επιδόσεις σε κάποιο άλλο πεδίο.
- Απαιτούν μεγάλο χρόνο και προσπάθεια για την κατασκευή του συνόλου εκπαίδευσης του ταξινομητή, ώστε να καταφέρουμε να βρούμε τις καλύτερες τιμές ή να διαπιστώσουμε τους απαραίτητους κανόνες.
- Επίσης η ακρίβεια εξαρτάται από το σύνολο εκπαίδευσης και για αυτό τα εκπαιδευτικά πρότυπα πρέπει να είναι αντιπροσωπευτικά του εκάστοτε πληθυσμού.

Έχει αναφερθεί από τους [HBB13] ότι η χρήση μαζί λεξικολογικών προσεγγίσεων και προσεγγίσεων μηχανικής μάθησης συνδυασμένες δίνουν καλύτερα αποτελέσματα. Αυτοί είναι οι λόγοι για τους οποίους όσοι ασχολούνται με την ανάλυση συναισθήματος και ειδικότερα με την ανάλυση συναισθήματος στο Twitter στρέφονται όλο και περισσότερο σε υβριδικές μεθόδους.

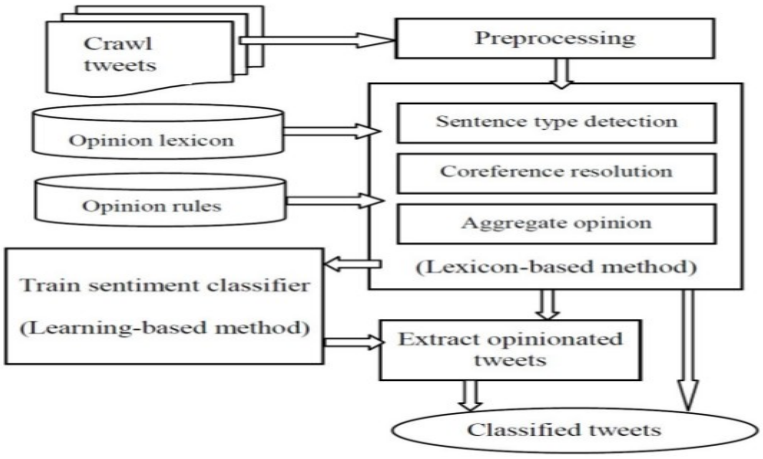
Παρακάτω θα αναφερθούμε σε ορισμένες χαρακτηριστικές περιπτώσεις συνδυασμού προσεγγίσεων με λεξικά και προσεγγίσεων με επιβλεπόμενη μηχανική μάθηση, πιο συγκεκριμένα θα δούμε χρήση λεξικών για εμπλουτισμό των χαρακτηριστικών και για βελτίωση των επιβλεπόμενων μεθόδων. Με αυτό τον τρόπο καλύπτεται το κενό που είχαν προσπάθειες ανάλυσης συναισθήματος με μεμονωμένες μεθοδολογίες σε αποτελεσματικότητα, σε ακρίβεια, σε χρόνο και σε διευκόλυνση συλλογής κατάλληλων χαρακτηριστικών.

5.1.1 Συνδυασμός Lexicon-based και SVM

Στην εργασία "**Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis**", [ZGD+11] παρουσιάζεται μια νέα μέθοδος για ανάλυση συναισθήματος σε επίπεδο περιεχομένου για δεδομένα που εξάγονται από το Twitter. Κάνουν χρήση λεξικών, και έπειτα για βελτίωση αποτελεσμάτων εφαρμόζουν και επιβλεπόμενη μέθοδο. Διαπιστώνουν ότι το Twitter έχει κάποια δικά του χαρακτηριστικά, που μερικά από αυτά είναι επιβλαβή για τις βασισμένες σε λεξικά (lexicon-based) προσεγγίσεις. Για παράδειγμα "I bought iPad yesterday, just lovvee it :-)". Ξεκάθαρα εκφράζει μια θετική γνώμη για το iPad από τη λέξη "lovvee" και από το emoticon ":-)". Αλλά η μέθοδος βασισμένη σε λεξικό (lexicon-based) θα αξιολογούσε το tweet ότι εκφράζει μια ουδέτερη (no/neutral) γνώμη για iPad, μιας και δεν υπάρχει μια γενική λέξη φέρουσα γνώμης (opinion word) στο tweet. Αυτό οδηγεί σε χαμηλή ανάκληση (recall) που είναι πρόβλημα για μεθόδους βασισμένες σε λεξικά (lexicon-based method), και εξαρτάται εξολοκλήρου από την παρουσία των λέξεων που φέρουν γνώμη (opinion words) που χρειάζονται για να καθορίσουν την συναισθηματική πολικότητα. Αυτές είναι πολύ δύσκολο να προστεθούν μιας και το Twitter είναι από τα πλέον δυναμικά συστήματα στο διαδίκτυο. Επίσης, οι πολικότητες μπορούν να είναι σχετιζόμενες από το εκάστοτε θέμα.

Για αυτό προτείνουν αυτή τη μέθοδο που υιοθετεί:

- Πρώτα μια προσέγγιση βασισμένη σε λεξικά (lexicon-based) (Ding et al., 2008) για να πραγματοποιήσει αναγνώριση συναισθήματος σε επίπεδο περιεχομένου. Αυτή η μέθοδος μπορεί να μας δώσει πολύ υψηλή ακρίβεια (precision), αλλά χαμηλή ανάκληση (recall). Για να βελτιωθεί η ανάκληση (recall), επιπρόσθετα tweets που θα πρέπει να είναι γνωμοδοτημένα, αναγνωρίζονται αυτόματα εισάγοντας πληροφορία από τα αποτελέσματα της λεξικολογικής μεθόδου (lexicon-based method).
- Έπειτα, εκπαιδεύεται ένας ταξινομητής (SVM) για να εισάγει πολικότητες στις έννοιες των νέων tweets που αναγνωρίζουμε. Αντί να βάζουμε μόνοι μας τις ταμπέλες, αυτό γίνετε αυτόματα με τη λεξικολογική προσέγγιση (lexicon-based approach).



Σχήμα 5.1: Περιγραφή μεθόδου

Στην αρχή, κατά την προ-επεξεργασία: Διαγράφονται τα διπλότυπα, αφαιρούνται ονόματα χρηστών και οι υπερσυνδέσμοι, αντικαθιστούν συντομογραφίες με την κανονική τους μορφή, αναγνωρίζουν γραμματικά τους επιμέρους όρους των μηνυμάτων (POS tagging).

Ύστερα υπολογίζουν τη συναισθηματική τιμή κάθε όρου με βάση την ομοιότητά του με λέξεις από το λεξικό συναισθημάτων (Ding et al., 2008) και επιλύουν τις απλές αναφορές αντιστοιχίζοντας αντωνυμίες με την πιο κοντινή οντότητα του κειμένου.

Ο αλγόριθμος διαχωρίζει τις προτάσεις σε τρεις κατηγορίες

- δηλωτικές,
- προστακτικές,
- ερωτηματικές

ενώ παράλληλα μπορεί να αναγνωρίσει συγκρίσεις, αρνήσεις και αντιθετικές περιόδους. (Οι πρώτες δύο περιπτώσεις προτάσεων συνήθως εκφράζουν γνώμες, ενώ ο τρίτος τύπος δεν εκφράζει κάποια πληροφορία για γνώμη σε κάποια οντότητα.). Εκπαιδεύουν, ένα δυαδικό ταξινομητή Support Vector Machines (SVM) με πρότυπα που προκύπτουν από την παραπάνω μη-επιβλεπόμενη διαδικασία ο οποίος ταξινομεί τα tweets στις τελικές τους κατηγορίες.

Τα πειράματα που πραγματοποιούν δείχνουν ότι η ανάκληση (recall) και και το F-score βελτιώνετε κατά πολύ, και ξεπερνά πολλές από τις τεχνολογίες αιχμής.

5.1.2 Συνδυασμός χαρακτηριστικών (features) σε SVM με λεξικό MPQA

Στο άρθρο "**Robust sentiment detection on twitter from biased and noisy data**", [BF10] προτείνεται μια προσέγγιση για αυτοματοποιημένη ανίχνευση συναισθήματος στα μηνύματα του Twitter (tweets) που εξερευνά τον τρόπο που γράφονται κάποια tweets και τη μετά-πληροφορία των λέξεων που περιέχεται σε αυτά. Πρόκειται για μια υβριδική μέθοδο αφού χρησιμοποιεί επιβλεπόμενες μεθόδους με ταξινομητή SVM αλλά και λεξικό MPQA για

να καθοριστεί ο βαθμός υποκειμενικότητας και πολικότητας της εκάστοτε λέξης για να εμπλουτίσει τα χαρακτηριστικά του ταξινομητή.

Πολλά συστήματα και προσεγγίσεις που υπάρχουν για να εφαρμόσουν ανάλυση συναισθήματος σε κείμενα χρησιμοποιούν πρωταρχική αναπαράσταση των λέξεων με τα ν-γράμματα (n-grams) ως χαρακτηριστικά (features) για να κτίσουν το μοντέλο της ανάλυσης συναισθήματος και να εφαρμόσουν αυτό το μοντέλο πάνω σε ένα μεγάλο πλήθος κειμένων. Όμως, αυτό που ώθησε την ομάδα αυτή να ασχοληθεί με μια διαφορετική, υβριδική μέθοδο είναι ο περιορισμός της χρήσης αυτής της τεχνικής στο Twitter λόγο του μικρού μεγέθους των tweets. Με μέγιστο μέγεθος 140 χαρακτήρες.

Σε αυτή την εργασία, προτάθηκε ένα μοντέλο δύο φάσεων:

- Στην πρώτη φάση, τα tweets κατηγοριοποιούνται σε υποκειμενικά ή αντικειμενικά
- και στη συνέχεια τα υποκειμενικά διακρίνονται σε θετικά ή αρνητικά tweets.

Για να μειώσουμε τη προσπάθεια για καταχώρηση ετικετών (labeling) κατά τη δημιουργία των ταξινομητών, αντί να χρησιμοποιούν ως noisy labels τα emoticons χρησιμοποιούν τη "γνώμη" τριών εργαλείων ανίχνευσης συναισθήματος:

1. Twendz
2. Twitter Sentiment
3. TweetFeel

Μετά διαγράφουν τα tweets στα οποία δεν συμφωνούν τα αποτελέσματα μεταξύ τους. Διαχωρίζουν τα χαρακτηριστικά τους σε δύο κατηγορίες :

- Τα μετά-χαρακτηριστικά (meta-features) που περιλαμβάνει χαρακτηριστικά όπως POS tags, ο βαθμός υποκειμενικότητας και πολικότητας της εκάστοτε λέξης όπως αυτά προσδιορίζονται στο λεξικό MPQA. Σημειώνουν ότι πρέπει να αντιστρέφεται η πολικότητα από θετική σε αρνητική και αντίστροφα όταν μία άρνηση προηγείται της λέξης.
- Τα χαρακτηριστικά σύνταξης του tweets (tweet-syntax) που περιλαμβάνει τα πιο ειδικά χαρακτηριστικά του Twitter όπως η ύπαρξη retweets, hashtags, υπερσυνδέσμων, θαυμαστικών, ερωτηματικών, emoticons και κεφαλαίων γραμμάτων. Η συχνότητα κάθε χαρακτηριστικού κανονικοποιείται διαιρώντας με το πλήθος των όρων του κάθε tweet.

Όλα μαζί τα χαρακτηριστικά και από τις δύο κατηγορίες είναι 20.

Στα πειράματα που έκαναν, έδειξαν ότι εφόσον τα χαρακτηριστικά είναι ικανά να καταγράψουν μία πιο αφηρημένη αναπαράσταση από τα tweets, είναι πιο αποδοτική λύση από τις προηγούμενες.

Τα καλύτερα αποτελέσματα προκύπτουν χρησιμοποιώντας ως ταξινομητή τον SVM και στις δύο φάσεις επιτυγχάνουν ακρίβεια 81.9% στην αναγνώριση υποκειμενικότητας και 81.3% στην αναγνώριση πολικότητας ενώ ως βάση αναφοράς θεωρούνται τα μονογράμματα με

ακρίβεια 72.4% και 79.1% αντίστοιχα στις δύο φάσεις. Συμπεραίνουν ότι τα μέτα-χαρακτηριστικά είναι πιο σημαντικά στη φάση προσδιορισμού της πολικότητας ενώ τα χαρακτηριστικά σύνταξης κατά την φάση αναγνώρισης της υποκειμενικότητας.

Επειδή χρησιμοποιούν μια πιο αφηρημένη αναπαράσταση των δεδομένων και όχι μεμονωμένους όρους του, η προσέγγισή τους εμφανίζει μεγαλύτερη ανοχή στο θόρυβο και μεροληψία (bias) του συνόλου εκπαίδευσης σε σχέση με άλλες μεθόδους ενώ εμφανίζει καλύτερη συμπεριφορά ως προς την ικανότητας γενίκευσης όταν χρησιμοποιούνται σχετικά λίγα εκπαιδευτικά πρότυπα.

5.1.3 Χρήση MPQA λεξικού για επιλογή χαρακτηριστικών (feature selection) σε επιβλεπόμενη μάθηση

Στο άρθρο "**Twitter sentiment analysis: The good the bad and the omg!**", [KWM11], προσεγγίζουν τον καθορισμό της πολικότητας με χρήση γλωσσολογικών χαρακτηριστικών των tweets. Προτείνουν μια υβριδική μέθοδο όπου γίνετε χρήση επιβλεπόμενων μεθόδων με τους ταξινομητές AdaBoost και Support Vector Machines (SVM) αλλά και χρήση MPQA λεξικού για καθορισμό πολικότητας των χαρακτηριστικών.

Κατά τη φάση της προ-επεξεργασίας:

- Αντικαθιστούν τα ονόματα χρηστών, τους υπερσυνδέσμους και τα hashtags με κατάλληλες λέξεις-κλειδιά
- Διορθώνουν την ορθογραφία των λέξεων από εμφατική επιμήκυνση και χρήση κεφαλαίων γραμμάτων
- Αντικαθιστούν τις συντομογραφίες με την κανονικής τους μορφή
- Αφαιρούν επίσης κοινές λέξεις και άρθρα (stopwording)
- Αναγνωρίζουν γραμματικά την κάθε λέξη (POS tagging).

Τα χαρακτηριστικά (features) που χρησιμοποιούν είναι μονογράμματα, διγράμματα. Διαλέγουν τα πρώτα 1000 μονογράμματα και διγράμματα με βάση το κέρδος πληροφορίας κατά το δείκτη Chi-squared, τη πρότερη πολικότητα των λέξεων κατά το MPQA λεξικό, το πλήθος και ποσοστό των κυριότερων POS tags και δυαδικά χαρακτηριστικά μικρο-ιστολογίων για την ύπαρξη εξειδικευμένων όρων (hashtags, emoticons). Εφαρμόζουν τους ταξινομητές AdaBoost και τον Support Vector Machines (SVM) και καταλήγουν σε ενδιαφέροντα συμπεράσματα, ότι σε αντίθεση με τα χαρακτηριστικά μικρο-ιστολογίων που ήταν τα πιο χρήσιμα, τα POS tags οδηγούν σε μείωση της ακρίβειας και δεν είναι μάλλον κατάλληλα για χρήση σε κείμενα από μικρο-ιστολόγια.

5.2 Υβριδικές μέθοδοι με επιβλεπόμενη μηχανική μάθηση και σημασιολογικές προσεγγίσεις

Ένα πολύ συνηθισμένο φαινόμενο είναι οι προσεγγίσεις ανάλυσης συναισθήματος να είναι πλήρως εξαρτώμενες από τη παρουσία των λέξεων ή των συντακτικών χαρακτηριστικών που αντανακλούν το συναίσθημα, αλλά σε πολλές περιπτώσεις το συναίσθημα κρύβεται στη σημασιολογία. Χαρακτηριστικό παράδειγμα, "πήγαινε διάβασε το βιβλίο". Η πρόταση μπορεί να εκφράζει θετική άποψη όταν αναφέρεται σε κριτική βιβλίου. Η ίδια πρόταση, όμως, μπορεί να εκφράζει εντελώς διαφορετική άποψη όταν χρησιμοποιείται σε κριτική ταινίας. Ανάλογα με τη σημασιολογία δηλαδή μπορεί να αλλάξει η συναισθηματική πολικότητα ενός κειμένου. Επίσης στη σημασιολογία μπορεί να διαφαίνεται και το αντικείμενο στο οποίο αναφερόμαστε. Η γενικότερη αντίληψη της θετικής ή αρνητικής άποψης δεν εξαρτάται άμεσα από το εκάστοτε θέμα συζήτησης. Ωστόσο, το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετείται. Οι σημασιολογικές προσεγγίσεις μπορούν να κατηγοριοποιηθούν σε δύο μεγάλες κατηγορίες, είναι οι προσεγγίσεις σημασιολογίας με συμφραζόμενα (contextual semantic approaches), και οι εννοιολογικές σημασιολογικές προσεγγίσεις (conceptual semantic approaches). Παρακάτω αναφέρονται κάποια χαρακτηριστικά παραδείγματα τέτοιων προσεγγίσεων.

5.2.1 Επέκταση επιβλεπόμενων μεθόδων με χρήση DBpedia, WordNet και SentiWordNet

Στην εργασία "**Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging**", [HBB13] πρότειναν μια επιβλεπόμενη μέθοδο με χρήση επιπλέον χαρακτηριστικών (features) από την DBpedia, το WordNet και το SentiWordNet. Πρόκειται για μια υβριδική μέθοδο με χρήση της σημασιολογίας για να εμπλουτιστούν τα χαρακτηριστικά και να βελτιωθεί η αποδοτικότητα της μεθόδου. Προτάθηκε να χρησιμοποιηθούν επιπλέον χαρακτηριστικά (features) για να βελτιωθεί η εκπαίδευση των ταξινομητών (SVM, Naive Bayes). Αυτά τα χαρακτηριστικά (features) επεκτείνουν το διάνυσμα χαρακτηριστικών (feature vector) του μονο-γραμματικού (unigram) μοντέλου

1. χρησιμοποιώντας τις έννοιες που εξάγονται από DBpedia
2. τα ρήματα και τα επίθετα τα λήφθηκαν από το WordNet
3. τα Senti- features εξάχθηκαν από το SentiWordNet.

Επίσης κατασκευάστηκε ένα λεξικό από emoticons, συντομογραφίες και αργκό λέξεων των tweets που φάνηκε πολύ χρήσιμο προτού το επεκτείνουμε με τα διαφορετικά χαρακτηριστικά (features).

Για χαρακτηριστικά (features) χρησιμοποιήθηκαν :

Λέξεις από το κείμενο (μονογράμματα) (Bag of words, uni-gram), η παρουσία ενός URL ή όχι, αν ήταν retweeted, ο αριθμός των "Not", των χαρούμενων emoticons, και των λυπημένων emoticons, των θαυμαστικών και ερωτηματικών, των λέξεων που ξεκινούν με κεφαλαίο γράμμα, των "@". Το μέρος του λόγου (part of speech) στο όποιο ανήκουν και άλλα συντακτικά και σημασιολογικά χαρακτηριστικά.

- Από τη DBpedia εξάχθηκε το περιεχόμενο κάθε tweet (π.χ. DBpedia concepts for Chapel Hill are (Settlement, PopulatedPlace, Place). Για αυτό αν υποθέσουμε ότι ο κόσμος αναφέρεται θετικά για κάποια κατοικία, θα είναι πολύ πιθανό να αναφέρεται θετικά για το Chapel Hill.)).
- Επιπλέον, σχετιζόμενα επίθετα και ρήματα εξαγόμενα από το WordNet.
- Μερικά επιπλέον senti-features εξαγόμενα από το SentiWordNet όπως ο αριθμός των θετικών, των αρνητικών και των ουδέτερων λέξεων ώστε να υπολογιστεί η θετικότητα, η αρνητικότητα, και η ουδετερότητα των tweets, η πολικότητα τους και η αντικειμενικότητά τους.

Προσθέτοντας τα καινούργια χαρακτηριστικά (features) βελτιώθηκε το f-measure, ακρίβεια (accuracy) 2% με SVM και 4% με Naive Bayes. Πιστεύουν ότι χρησιμοποιώντας κάποιο επιπλέον λεξικό μαζί με το SentiWordNet θα πετύχουν ακόμα καλύτερα αποτελέσματα. Διαπιστώθηκε ότι τα επίθετα είναι τα πιο χρήσιμα χαρακτηριστικά (features) και πρέπει να επικεντρωθούμε στην εξαγωγή ταιριαστών και παρόμοιων επιθέτων. Για τις συντομογραφίες όπως LOL (loud of laughing), μπορεί να είναι καλύτερο να αντικατασταθούν από κάποια έκφραση (επίθετο) που αποδίδει καλύτερα το συναίσθημα του συγγραφέα. Όμως μπορούν να εισαχθούν στο λεξικό αυτά τα επίθετα και να διαχειριστούν καλύτερα τα emoticons.

5.2.2 Σημασιολογικά χαρακτηριστικά μαζί με Naive Bayes

Στο άρθρο "**Semantic Sentiment Analysis of Twitter**", [SHA12.], προτείνουν μια θεωρητική προσέγγιση για να εισάγουν σημασιολογία σαν ένα επιπλέον χαρακτηριστικό στο training set για την ανάλυση συναισθήματος. Για κάθε εξαγόμενη οντότητα π.χ. iPhone από τα tweets, προσθέτουν μια σημασιολογική έννοια όπως "Apple product", σαν ένα επιπλέον χαρακτηριστικό, και μετράνε την συσχέτιση του χαρακτηριστικού αυτού στο τελικό αποτέλεσμα αρνητικό/θετικό συναίσθημα.

Διαπιστώνουν πως υπάρχουν δύο κύριες κατευθύνσεις σε ότι αφορά την ανάλυση συναισθήματος στα μικρο-ιστολόγια.

1. Η κατεύθυνση για ανακάλυψη νέων μεθόδων, όπως η εκτέλεση με γράφους και εφαρμογή κοινωνικών σχέσεων για ανάλυση συναισθήματος.
2. Η κατεύθυνση που επικεντρώνεται στην αναγνώριση νέων συνόλων από χαρακτηριστικά για να προσθέσει στο μοντέλο εκπαίδευσης για την ανάλυση συναισθήματος, όπως τα χαρακτηριστικά από μικρο-ιστολόγια συμπεριλαμβανομένων των hashtags, emoticons, την παρουσία ενισχυτών όπως τα κεφαλαία γράμματα και οι επαναλήψεις.

Ενώ προηγούμενες δουλείες σε χαρακτηριστικά για ανάλυση συναισθήματος απλώς χρησιμοποιούσαν τα χαρακτηριστικά αυξάνοντάς τα, σε αυτή την εργασία επικεντρώνονται στην δεύτερη κατηγορία, ερευνώντας μια θεωρητική προσέγγιση ενός συνόλου από χαρακτηριστικά που εξάγονται από σημασιολογία που εμφανίζεται στα tweets. Τα σημασιολογικά χαρακτηριστικά αποτελούνται από σημασιολογικές έννοιες (π.χ. "person", "company", "city") που αναπαριστούν τις οντότητες (π.χ. "Steve Jobs", "Vodafone", "London") εξαγόμενες από τα tweets. Η λογική πίσω από την επιλογή αυτών των χαρακτηριστικών είναι ότι οι συγκεκριμένες οντότητες και τα πεδία τείνουν να έχουν πιο συχνή σχέση με θετικά ή αρνητικά συναισθήματα. Γνωρίζοντας τις σχέσεις αυτές μπορούν να βελτιώσουν τη ανάλυση συναισθήματος.

Αξιολογούν τρία εργαλεία εξαγωγής περιεχομένου και αναγνώρισης θεματολογίας

1. AlchemyAPI
2. Zemanta
3. OpenCalais

και χρησιμοποίησαν αυτό με την καλύτερη επίδοση σε όρους ποσότητας και ακρίβειας των αναγνωρισμένων θεμάτων.

Πιο συγκεκριμένα μετά την εξαγωγή σημασιολογικών οντοτήτων και εννοιών (Extracting Semantic Entities and Concepts) ακολουθεί η σημασιολογική ενσωμάτωση χαρακτηριστικών (Semantic Feature Incorporation). Σε αυτό το στάδιο, προτείνουν τρεις διαφορετικές μεθόδους για ενσωμάτωση σημασιολογικών χαρακτηριστικών στην εκπαίδευση του ταξινομητή Naive Bayes (NB).

- Σημασιολογική αντικατάσταση: Σε αυτή τη μέθοδο, αντικαθιστούν όλες τις οντότητες των tweets με τις αντίστοιχες σημασιολογικές έννοιες. Αυτό οδηγεί στη μείωση του μεγέθους του λεξιλογίου.
- Σημασιολογική επαύξηση: Σε αυτή τη μέθοδο επαυξάνουν το αρχικό μέγεθος των χαρακτηριστικών με σημασιολογικές έννοιες σαν επιπλέον χαρακτηριστικά για την εκπαίδευση του ταξινομητή.
- Σημασιολογική παρεμβολή: Ένας πιο βασικός τρόπος να ενσωματωθούν σημασιολογικές έννοιες, είναι με παρεμβολή. Παρεμβάλουν το μοντέλο των μονογραμμάτων (unigram) του Naive Bayes, με ένα μοντέλο που παράγει λέξεις που

διαθέτουν σημασιολογικές έννοιες. Μπορούν να παρεμβάλουν αυθέραιτους τύπους χαρακτηριστικών όπως σημασιολογικές έννοιες, συχνότητες μερών του λόγου (POS), συναισθηματικά θέματα κτλ.

Εφαρμόζουν την προσέγγιση τους για να προβλέψουν το συναίσθημα από 3 διαφορετικά Twitter datasets,

1. το general Stanford Twitter Sentiment (STS) dataset,
2. το Obama-McCain Debate (OMD)
3. και το Health Care Reform (HCR).

Τα αποτελέσματά δείχνουν μέση αύξηση της F harmonic, ακρίβεια στην αναγνώριση αρνητικών και θετικών συναισθημάτων κατά 6.5% και 4.8% πάνω από τα βασικά επίπεδα των χαρακτηριστικών μονό-γραμμάτων (unigrams) και του μέρους του λόγου (part-of-speech). Επίσης συνέκριναν σε σχέση με την προσέγγιση του φέροντος συναισθήματος (sentiment-bearing), και βρήκαν ότι τα σημασιολογικά χαρακτηριστικά παράγουν καλύτερη ανάκληση (Recall) και F score όταν ταξινομούν αρνητικό συναίσθημα, και καλύτερη ακρίβεια (Precision) με χαμηλότερα ανάκληση (Recall) και F score όταν ταξινομούν θετικό συναίσθημα.

5.2.3 Lexicon-based προσέγγιση μαζί με συμφραζόμενα και την εννοιολογική σημασιολογία

Λόγω των περιορισμών των λεξικολογικών προσεγγίσεων που έχουν αναφερθεί παραπάνω και των ιδιαίτερων χαρακτηριστικών του Twitter οι [SFH+14], στην εργασία "**SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter**", οδηγήθηκαν να προτείνουν μια νέα υβριδική μέθοδο. Διαπιστώνουν πως το συναίσθημα δεν εξαρτάται μόνο από τις έννοιες και δεν είναι σταθερό αλλά εξαρτάται και από τα συμφραζόμενα. Η μέθοδός τους ονομάζετε SentiCircle και είναι μια θεωρητική λεξικολογική (lexicon-based) προσέγγιση που παίρνει υπόψη τα συμφραζόμενα και την εννοιολογική σημασιολογία και υπολογίζει τη συναισθηματική κατεύθυνση και την ένταση του συναισθήματος στο Twitter.

Για παράδειγμα το "Wind" και το "Humidity" έχει αρνητική έννοια στο SentiCircles καθώς τείνουν να παρουσιάζονται σε αρνητικούς όρους στα tweets. Όμως το σημασιολογικό θέμα τους "Weather Condition" θα έχει και αυτό αρνητικό συναίσθημα. Το tweet "Cycling under a heavy rain.. What a #luck!" είναι πιθανό να πάρει ένα αρνητικό συναίσθημα καθώς περιέχει τη λέξη "rain" όπου είναι αντιστοιχισμένη με το θέμα "Weather Condition"που έχει λάβει αρνητική αξιολόγηση. Επιπλέον, η λέξη heavy σε αυτό το περιεχόμενο είναι επίσης πιθανό να αντιστοιχηθεί με αρνητικό συναίσθημα λόγο της σχέσεις με τη λέξη "rain" και με το "Weather Condition".

Για να αντιμετωπίσουν αυτά τα προβλήματα, χτίζουν μια δυναμική αναπαράσταση περιεχομένου που θα ενσωματώνει δύο τύπους σημασιολογίας:

1. Σημασιολογία συμφραζομένων δηλαδή σημασιολογία που προκύπτει από τη συνύπαρξη λέξεων.
2. Εννοιολογική σημασιολογία, δηλαδή σημασιολογία που προέρχεται από τις οντότητες του DBpedia.

Ορίζουν ένα δείκτη TDOC (Term Degree of Correlation), που είναι μια μετρική εμπνευσμένη από το δείκτη TF-IDF, που υπολογίζει τη σχέση ανάμεσα σε μία λέξη και τους συμφραζόμενους όρους με την ίδια σημασιολογία. Αναπαριστούν το σύστημα σε πολικές συντεταγμένες και υπολογίζουν τη πολικότητα του συναισθήματος και την ισχύ του συναισθήματος χρησιμοποιώντας τριγωνομετρικές ιδιότητες.

- Εξάγουν τη σημασιολογία του περιεχομένου όπως (π.χ., "person", "company", "city") από τις υπάρχουσες οντότητες (π.χ. "Steve Jobs", "Vodafone", "London") που εμφανίζονται στα tweets.
- Εδώ χρησιμοποιούν το AlchemyAPI για να εξάγουν όλες τις ονοματισμένες οντότητες στα tweets μαζί με τα συσχετιζόμενα περιεχόμενα.
- Εισάγουν τις έννοιες μέσα στο SentiCircle χρησιμοποιώντας το Semantic Augmentation method που είναι η μέθοδος που αναφέρθηκε στην προηγούμενη εργασία, όπου προσθέτουν τη σημασιολογική έννοια στο αρχικό tweet προτού εφαρμόσουν τη μέθοδο αναπαράστασης π.χ. "headache" και το θέμα του (περιεχόμενο) "Health Condition" θα εμφανιστεί μαζί με το SentiCircle).
- Κάθε εξαγόμενη έννοια θα αναπαρίσταται με το SentiCircle ώστε να υπολογίζεται το συνολικό συναίσθημα.

Η λογική πίσω από την πρόσθεση αυτών των περιεχομένων είναι ότι οι συγκεκριμένες οντότητες και τα θέματα τείνουν να έχουν πιο συχνή σχέση με τους όρους που φέρουν θετικό ή αρνητικό συναίσθημα. Αυτό μπορεί να βοηθήσει να καθοριστεί το συναίσθημα των σημασιολογικά σχετιζόμενων ή όμοιων οντοτήτων που δεν εκφράζουν κάποιο συναίσθημα.

Τέλος συγκρίνουν τη μέθοδο μας με τρία Twitter datasets,

1. Obama McCain Debate
2. Health Care Reform
3. Standford Sentiment Gold Standard

χρησιμοποιώντας τρία λεξικά συναισθήματος (sentiment lexicons). Τα αποτελέσματα δείχνουν ότι η προσέγγισή αυτή ξεπερνά σημαντικά τις άλλες μεθόδους που βασίζονται μόνο σε λεξικά. Τα αποτελέσματα είναι ανταγωνιστικά με τη μέθοδο SentiStrength, και πολύ διαφορετικά από το ένα dataset στο άλλο. SentiCircle (72.39%) ξεπερνά το SentiStrength (71.7%.) σε ακρίβεια κατά μέσο όρο αλλά αποτυγχάνει στο F-measure.

5.3 Πίνακας υβριδικών μεθόδων

Παρακάτω ακολουθεί ένας συγκεντρωτικός πίνακας για τις μεθόδους που χρησιμοποιούν υβριδικά μοντέλα για την ανάλυση συναισθήματος. Κάνουν χρήση επιβλεπόμενης μηχανικής μάθησης σε συνδυασμό με λεξικολογικές προσεγγίσεις, αλλά και συνδυασμό με σημασιολογικές προσεγγίσεις. Οι μελέτες αυτές έχουν τις περισσότερες αναφορές (citations) και είναι από τις πιο σύγχρονες στη βιβλιογραφία.

Συγγραφείς	**Τίτλος**	**Έτος**	**Αλγόριθμος**	**Σημασιολογία**	**Λεξικό**	**Σύνολο Δεδομένων (Dataset)**	**Απόδοση**
Barbosa, Feng, [BF10]	Robust Sentiment Detection on Twitter from Biased and Noisy Data	2010	SVM	-	MPQA lexicon	Tweets	81.9 %, 81.3 %
Zhang, Ghosh, Dekhil, Hsu, Liu, [ZGD+11]	Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis	2011	SVM	-	lexicon based	Δεδομένα που συλλέγουν από το Twitter	-
Kouloumpis ,Wilson, Moore, [KWM11]	Twitter sentiment analysis: The good the bad and the omg!	2011	AdaBoost, SVM	-	MPQA λεξικό	Endinburgh Twitter Corpus, Stanford Twitter Sentiment Corpus	AdaBoost 75%
Hamdan, Béchet, Bellot, [HBB13]	Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging	2013	SVM, NaiveBayes	DBpedia	WordNet, SentiWordNet	Tweets	βελτιώθηκε το f-measure, ακρίβεια (accuracy) 2% με SVM και 4% με NaiveBayes
Saif, He, Alani, [SHA12.]	Semantic Sentiment Analysis of Twitter	2012	NB	Σημασιολογική αντικατάσταση/ επαύξηση / παρεμβολ	-	general Stanford Twitter Sentiment (STS), Obama-McCain	6.5% και 4.8%

				ή		Debate (OMD), Health Care Reform (HCR)	
Saif, Fernandez, He, Alani, [SFH+14]	SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter	2014	-	DBpedia	Τρία λεξικά	Obama McCain Debate, Health Care Reform, Standford Sentiment Gold Standard	72.39%
Li, Liu, [LL10]	A Clustering-based Approach on Sentiment Analysis	2010	K-means, SVM	-	Wordnet	δεδομένα από κριτικές ταινιών στην αγγλική γλώσσα	78,33%
Hu, Tang, Gao, Liu, [HTG+13]	Unsupervised sentiment analysis with emotional signals	2013	Μη-επιβλεπόμενη μέθοδος	-	λεξικό MPQA	Stanford Twitter Sentiment Corpus, Obama McCain Debate Corpus	74.2% και 70.97%

Πίνακας 5.2: Πίνακας για υβριδικές μεθόδους

5.4 Μη-επιβλεπόμενη μηχανική μάθηση

Σε αντίθεση με τις διαδεδομένες λεξικολογικές μεθόδους (με χρήση λεξικού) οι οποίες παρουσιάζουν χαμηλά ποσοστά ακρίβειας και τις μεθόδους επιβλεπόμενης μηχανικής μάθησης που αν και ακριβείς είναι χρονοβόρες, ακριβές και απαιτούν ανθρώπινη συμμετοχή. Υπάρχει η μη-επιβλεπόμενη μηχανική μάθηση που την συναντάμε και με το όνομα μάθηση από παρατήρηση (Learning from observation). Σε αντιπαράθεση με την επιβλεπόμενη μηχανική μάθηση που πραγματοποιεί κατηγοριοποίηση (classification) έχοντας κάποια εκπαίδευση, αυτή πραγματοποιεί συσταδοποίηση (clustering) στα αντικείμενα εισόδου χωρίς πρότερη εκπαίδευση. Για την ανάλυση συναισθήματος με χρήση μη-επιβλεππόμενης μηχανικής μάθησης, το σύστημα που θέλουμε να εκπαιδεύσουμε τροφοδοτείται με δεδομένα κειμένου τα οποία δεν είναι χαρακτηρισμένα ως προς το συναισθηματικό τους περιεχόμενο. Το σύστημα δηλαδή, δεν έχει πληροφορία για το τι αποτελεί σωστή δράση ή επιθυμητή κατάσταση, και έτσι, δεν μπορεί να εξάγει κανόνες γενίκευσης στους οποίους θα στηρίξει την συναισθηματική ανάλυση. Επομένως, προσπαθεί να ανακαλύψει ανάμεσα στα δεδομένα εκπαίδευσης κάποιες κρυμμένες δομές, ώστε να τα ταξινομήσει σε ομάδες, οι οποίες θα αποτελέσουν τις κλάσεις συναισθήματος. Στην περίπτωση αυτή, το σύστημα τροφοδοτείται

από εισόδους μόνο, και καλείτε να ανακαλύψει πιθανές κρυμμένες δομές ανάμεσα τους, ώστε να τις ταξινομήσει σε ομάδες δεδομένων που παρουσιάζουν κάποια ομοιότητα.

Κατά τη διάρκεια της μεθόδου υπάρχει η ανάγκη για ανάθεση κλάσης συναισθήματος σε κάθε μια από τις συστάδες. Η πιο συνηθισμένη μέθοδος είναι αυτή του υπολογισμού του βάρους των χαρακτηριστικών ενός κειμένου. Σύμφωνα με αυτή, για κάθε όρο υπολογίζεται η ομοιότητα του με λέξεις αναφοράς (π.χ. "good", "bad") με τη βοήθεια κάποιου λεξικού συνωνύμων και ανάλογα με την τιμή αυτής της ομοιότητας υπολογίζεται ένα βάρος για κάθε όρο, με αποτέλεσμα να εντάσσεται το κείμενο συνολικά σε κάποια κλάση συναισθήματος. Μια διαφορετική εκδοχή είπε ο Turney [Tur02], με ένα απλό μη-επιβλεπόμενο αλγόριθμο που ταξινόμησε ως thumbs up or thumbs down βασισμένος σε πληροφορία των φράσεων που είχε επίθετα ή επιρρήματα. Υπολόγισε τη συναισθηματική κατεύθυνση μια φράσης χρησιμοποιώντας τη mutual information της λέξης excellent πλην της λέξης poor.

Αρκετές από τις μη-επιβλεπόμενες μεθόδους επιτυγχάνουν ικανοποιητικά ποσοστά ακρίβειας όταν εφαρμόζονται σε γνωστά θεματικά πεδία όπου το λεξιλόγιο των κειμένων τους καλύπτεται από τα λεξικά συναισθήματος.

5.4.1 Συσταδοποίηση με τον αλγόριθμο K-means με εμπλουτισμό χαρακτηριστικών από Wordnet και σύγκριση με SVM

Στην εργασία "**A Clustering-based Approach on Sentiment Analysis**", [LL10], προτείνουν μια μη-επιβλεπόμενη μέθοδο για ανάλυση συναισθήματος με χρήση συσταδοποίησης (clustering) για ανάλυση σε δεδομένα από κριτικές ταινιών στην αγγλική γλώσσα. Στόχος τους είναι η ταξινόμηση των δεδομένων σε κλάσεις συναισθήματος, θετική ή αρνητική.

Κατά την προ-επεξεργασία κάνουν:

- Εύρεση της ρίζας (stemming)
- Χαρακτηρισμό του μέρους του λόγου (pos tagging)
- Μετατροπή των κειμένων σε διανύσματα με βάση τη συχνότητα των χαρακτηριστικών, και εφαρμογή του SVM προτού κάνουμε τη συσταδοποίηση. Αυτός μας δίνει αποτελέσματα 75.8% ακρίβεια (accuracy) για τη συχνότητα των δεδομένων και 75% για την παρουσία δεδομένων.
- Απομάκρυνση των ετικετών των δεδομένων που χρειαζόντουσαν στο προηγούμενο βήμα αλλά δεν χρειάζονται στον K-means αλγόριθμο.

Για τη συσταδοποίηση χρησιμοποίησαν τον αλγόριθμο K-means με συνάρτηση απόστασης την απόσταση συνημιτόνου.

- **Αλγόριθμος *k*-means συσταδοποίησης (clustering)**

  Πρόκειται για έναν από τους πιο απλούς αλγορίθμους μη-επιβλεπόμενης μηχανικής μάθησης που ασχολούνται με το πρόβλημα της συσταδοποίησης (clustering). Ο

αλγόριθμος αυτός χρησιμοποιεί συσταδοποίηση βασισμένη σε κέντρα (centroids), κατά την οποία μια συστάδα αναπαριστάται από το centroid της, που είναι ο μέσος όρος των σημείων που την αποτελούν. Επομένως, ένα σημείο ανήκει σε μια συστάδα αν η απόσταση του από το κέντρο της είναι η μικρότερη από τις αντίστοιχες αποστάσεις από τα κέντρα των υπολοίπων συστάδων του προβλήματος.

Αρχικά γίνεται μια τυχαία τοποθέτηση των k κέντρων στο χώρο των δεδομένων. Μετά ακολουθεί η ταξινόμηση των δεδομένων στις συστάδες που ορίζονται από τα παραπάνω centroids, ανάλογα με τις αποστάσεις τους από τα centroids, και ο επαναπροσδιορισμός των centroids με χρήση μέσου όρου των ταξινομημένων δεδομένων.

Η ακρίβεια που επιτεύχθηκε ήταν 60,17% και 65,67%. Για να βελτιώσουν τη μέθοδο τους χρησιμοποίησαν τη μέθοδο tf-ifd για την επιλογή των πιο σημαντικών χαρακτηριστικών πετυχαίνοντας ακρίβεια κοντά στα 79% αλλά παρατηρείτε κάποια αστάθεια που είναι χαρακτηριστικό του K-means.

Μετά, προσπαθούν να αντιμετωπίσουν αυτή την αστάθεια με τον εμπλουτισμό του πειράματος. Υπολογίζοντας τιμές-βάρη για τα χαρακτηριστικά ανάλογα με την σχέση τους με τις λέξεις αναφοράς “good”, και “bad”, με τη βοήθεια του Wordnet, και κρατώντας μόνο τα χαρακτηριστικά με υψηλά βάρη.

Έτσι είδαν να βελτιώνεται σημαντικά η ακρίβεια στην περίπτωση των διανυσμάτων συχνότητας χαρακτηριστικών, όπου επιτυγχάνεται ακρίβεια μέχρι και 78,33%. Σημαντικό πλεονέκτημα της μεθόδου αυτής είναι η καλή ταχύτητα της κατηγοριοποίησης σε σχέση με τις μεθόδους επιβλεπόμενης μηχανικής μάθησης, και τα καλύτερα αποτελέσματα σε σχέση με τις συμβολικές τεχνικές (Συμβολικές αναφέρουν εργασίες όπως η εργασία του Turney2002 Thumbs Up or Thumbs Down?). Αυτό φαίνεται και στον παρακάτω πίνακα που συνέταξαν οι συγγραφείς για να υποστηρίξουν τη μέθοδό τους.

	Ακρίβεια (Accuracy)	**Αποδοτικότητα**	**Ανθρώπινη συμμετοχή**
Συμβολικές Τεχνικές	65,83% ([Tur02])	Πολύ γρήγορες	Κυρίως όχι
Επιβλεπόμενες Τεχνικές	77%-82% ([PLV02])	Αργές στην εκπαίδευση αλλά γρήγορες στο έλεγχο	Ναι
Μη επιβλεπόμενες Τεχνικές	77,17%-78,33%	Γρήγορες	Όχι

Πίνακας 5.3: Κριτική μεθόδων

5.4.2 Μη επιβλεπόμενη μεθόδος με χρήση MPQA λεξικού και χρήση μεταπληροφορίας

Στην εργασία "**Unsupervised sentiment analysis with emotional signals**", [HTG+13], πραγματοποιείται μια μη-επιβλεπόμενη ανάλυση συναισθήματος που αφορά τα κοινωνικά δίκτυα. Παρατηρούν ότι ένα ιδιαίτερο χαρακτηριστικό στοιχείο των κοινωνικών δεδομένων είναι ότι συχνά παρέχει πρόσθετες πληροφορίες εκτός του κειμένου. Εισάγουν την έννοια των "σημάτων συναισθήματος" ή οποία ορίζετε ως η πληροφορία που έχει κάποια σχέση με συναισθήματα και συναισθηματική πολικότητα. Χρησιμοποιούν δύο σύνολα δεδομένων (dadasets):

1. Stanford Twitter Sentiment Corpus.
2. Obama McCain Debate Corpus.

για να εξερευνήσουν τη σχέση που έχουν με την κατηγοριοποίηση των δεδομένων τα χαρακτηριστικά όπως τα emoticons και οι κοινές εμφανίσεις αυτών.

Η διαδικασία που ακολουθούν για την εξαγωγή της ανάλυσης συναισθήματος είναι η παρακάτω:

- Χρησιμοποιούν μονογράμματα (unigrams) για να περιγράψουν τις λέξεις του κειμένου
- Χρησιμοποιούν τη πολικότητα από το συναισθηματικό λεξικό MPQA σαν χαρακτηριστικό
- Χρησιμοποιούν ως χαρακτηριστικό την εμφάνιση των λέξεων (term presence)
- Εξάγουν το δείκτη ένδειξης (emotion indication) συναισθήματος : ορίζεται από τα συναισθηματικά σήματα που αντανακλούν έντονα τη συναισθηματική πολικότητα μιας θέσης ή μιας λέξης και μπορεί να συλλεχθούν εύκολα από τα μέσα κοινωνικής δικτύωσης. Όπως είναι τα emoticons, βαθμολογίες προϊόντων, αστέρια εστιατορίων, κτλ. Για καθένα από τα δύο datasets, δύο ομάδες ίδιου μεγέθους επιλέχτηκαν στη μία ομάδα ήταν τα θετικά tweets και στην άλλη ήταν τα τυχαία. Η ίδια διαδικασία και για τα αρνητικά. Δημιούργησαν δύο διανύσματα με τα οποία αναπαράστησαν τις ομάδες αυτές. Τα αποτελέσματα έδειξαν ότι η ένδειξη (emotion indication) συναισθήματος είναι άμεσα συσχετισμένη με την πολικότητα συναισθήματος.
- Εξάγουν το δείκτη συσχέτισης (emotion correlation) συναισθήματος: ορίζεται από τα συναισθηματικά σήματα που αντανακλούν τη συσχέτιση μεταξύ των κειμένων ή λέξεων. Έχει διαπιστωθεί ότι δύο λέξεις που εμφανίζονται συχνά μαζί πρέπει να έχουν την ίδια συναισθηματική πολικότητα. Μπορούμε να διαισθανθούμε ότι είναι απίθανο να συναντήσουμε μπλεγμένες θετικές και αρνητικές απόψεις σε μικρά posts. Έχουμε λοιπόν συναισθηματική συσχέτιση στο post, όπως ένα post-post κοινωνικό δίκτυο, ομοιότητες κειμένου μεταξύ δύο posts, κτλ., και συναισθηματική συσχέτιση με τη λέξη. Όπως συνώνυμα από το WordNet, συν-εμφάνιση πληροφορίας στη Wikipedia, κτλ.

Δημιούργησαν το ESSA, (Emotional Signals for unsupervised Sentiment Analysis) όπου μοντελοποιούν τους δείκτες συναισθήματος και συσχέτισης. Οι δείκτες αυτοί χρησιμοποιήθηκαν στη μέθοδο με παραγοντοποίηση πίνακα για μη- επιβλεπόμενη μάθηση και κανονικοποίηση στη τριπλή παραγοντοποίηση που γίνετε κατά τη διάρκεια της μη-επιβλεπόμενης μεθόδου.

Τα συμπεράσματα που καταλήγουν μετά από πολλά πειράματα αφορούν τη παρούσα εργασία γενικά καθώς και το ρόλο που παίζουν τα σήματα συναισθημάτων και οι δείκτες που χρησιμοποιούνται για την ανάλυση συναισθήματος.

Τα αποτελέσματα είναι πολύ ενθαρρυντικά καθώς, σε σχέση με άλλες μη-επιβλεπόμενες τεχνικές εμφανίζει καλύτερα αποτελέσματα με ποσοστά 74.2% και 70.97% αντίστοιχα στα 2 σύνολα δεδομένων (datasets). Η μέθοδος για χρήση των σημάτων συναισθήματος έδωσε καλύτερα αποτελέσματα και καλύτερη ακρίβεια.

5.4.3 Μη επιβλεπόμενη τεχνική παρόμοια με αυτή των k-Κοντινότερων Γειτόνων (k-Nearest Neighbours – kNN)

Στη παρούσα μελέτη "**Enhanced sentiment learning using twitter hashtags and smileys**", [DTR10], κάνουν χρήση της μη-επιβλεπόμενης μεθόδου που μοιάζει με τη k-Κοντινότερων Γειτόνων (k-Nearest Neighbours – kNN). Με τη μέθοδο αυτή προσπαθούν να ταξινομήσουν αυτόματα το σύνολο δεδομένων των Brendan O'Connor . Για την ταξινόμηση αυτή χρησιμοποιούν σαν δείκτες κατηγοριοποίησης γνωστά και ως noisy labels, 50 Twitter tags και 15 emoticons.

Αξιοποιούν τέσσερις βασικούς τύπους χαρακτηριστικών για την ταξινόμηση συναισθήματος. Τα χαρακτηριστικά που χρησιμοποίησαν είναι:

- Οι λέξεις που αναπαρίστανται από δι-γράμματα μέχρι πεντε-γράμματα
- Το μήκος του κάθε tweet
- Το πλήθος των σημείων στίξης, των θαυμαστικών, των ερωτηματικών, των εισαγωγικών, των κεφαλαίων γραμμάτων, των λέξεων
- Η ύπαρξη ή όχι λέξεων με υψηλή συχνότητα εμφάνισης

Εφαρμόζουν μία τεχνική παρόμοια με αυτή των k-Κοντινότερων Γειτόνων (k-Nearest Neighbours – kNN).

- **Αλγόριθμος *k*-nearest neighbors algorithm (k-NN)**

  Είναι μια μη-παραμετρική μέθοδος για κατηγοριοποίηση (classification) και οπισθοδρόμηση (regression). Και στις δύο περιπτώσεις η είσοδος αποτελείτε από τα k πλησιέστερα σημεία στο χώρο των χαρακτηριστικών. Στην k-NN κατηγοριοποίηση (classification) η έξοδος είναι μια είσοδος σε κάποια κλάση. Ένα αντικείμενο κατηγοριοποιείτε από την πλειοψηφία των ψήφων των k γειτόνων του. Έπειτα το αντικείμενο απλά εισάγετε στην κλάση του πλησιέστερου γείτονα.

Η k-nn σύνδεση συνδυάζει κάθε σημείο του εξωτερικού συνόλου δεδομένων R με τους k-πλησιέστερους γείτονες από το εσωτερικό σύνολο δεδομένων S. Ένα σημείο είναι ένα πολυδιάστατο διάνυσα δεδομένων και η μετρική που χρησιμοποιείται για τον υπολογισμό της απόστασης είναι η Ευκλείδεια απόσταση.

Με τη μέθοδο αυτή πετυχαίνουν βέλτιστη ακρίβεια στο μέσο αρμονικό δείκτη F 1 = 0.86 για τα emoticons και F 1 = 0.8 για τα hashtags στη δυαδική εκδοχή του προβλήματος χρησιμοποιώντας 10-πλη σταυρωτή επικύρωση (10-fold cross-validation). Στο γενικότερο πρόβλημα των τριών κλάσεων, η επίδοση ήταν αισθητά χαμηλότερη (0.64 και 0.31 αντίστοιχα).

Επιπλέον, προτείνουν δύο διαφορετικές μεθόδους για την αυτόματη ανίχνευση της επικάλυψης συναισθήματος και των αλληλεξαρτήσεων ανάμεσα στις λέξεις του κειμένου. Παρατηρούν ότι οι λέξεις, τα σημεία στίξης και τα εκφραστικά μοτίβα είναι τα πιο σημαντικά χαρακτηριστικά ενώ τα ν-γράμματα οδηγούν σε οριακή βελτίωση.

5.5 Πίνακας μεθόδων μη επιβλεπόμενης μάθησης

Ακολουθεί ένας πίνακας που καταγράφονται οι μέθοδοι μη-επιβλεπόμενης μηχανικής μάθησης που συναντάμε στη βιβλιογραφία και έχουν το μεγαλύτερο ενδιαφέρον.

Συγγραφείς	**Τίτλος**	**Έτος**	**Αλγόριθμος**	**Λεξικό**	**Σύνολο Δεδομένων (Dataset)**	**Απόδοση**
Davidov, Tsur, Rappoport, [DTR10]	Enhanced sentiment learning using twitter hashtags and smileys	2010	k-Nearest Neighbours - kNN	-	δημοσιεύσεις του Twitter	F 1 = 0.86 για emoticons, F 1 = 0.8 για hashtags
Ortega, Fonseca, Montoyo, [OFM13]	Ssa-uo: unsupervised twitter sentiment analysis.	2013	Μη-επιβλεπόμενο σύστημα	-	Twitter από SemEval2013	F1= 51.17%.
Kumar, Sebastian, [KS12]	Sentiment analysis on twitter	2012	Χρήση features από επιβλεπόμενη	χρήση λεξικών	Twitter	-
Hatzivassiloglou, McKeown, [HM97]	Predicting the Semantic Orientation of Adjectives	1997	log-linear regression model, non-hierarchical clustering	-	1987 Wall Street Journal corpus	82% των σιαμαίων επιθέτων έχουν παρόμοια πολικότητα

Turney, [Tur02]	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews	2002	-	-	automobile, bank, movie, travel, reviews	65,83%
Clement Levallois, [Lev13]	Umigon: sentiment analysis for tweets based on lexicons and heuristics	2013	Ευριστικές	Χρήση Λεξικού	Twitter	-
O'Connor, Balasubramanyan, Routledge, Smith, [OBR+10]	From tweets to polls: Linking text sentiment to public opinion time series.	2010	Σύνδεση δημοσκοπήσεων με την ανάλυση συναισθήματος σε tweets	MPQA, Opinion Finder lexicon, (Wilson, Wiebe, and Hoffman 2005)	TwitterAPI, 1 δις tweets	80%.
Thelwall, Buckley, Paltoglou, Cai, Kappas, [TBP+10]	Sentiment strength detection in short informal text.	2010	-	SentiStrength	MySpace	-

Πίνακας 5.4: Πίνακας για μη-επιβλεπόμενη μηχανική μάθηση

6 Εφαρμογές

Παραπάνω είδαμε τις μεθόδους με τις οποίες προσεγγίζουμε την ανάλυση συναισθήματος. Από την αρχή υπήρχε η ανάγκη δημιουργίας πραγματικών εφαρμογών που θα υλοποιούσαν την ανάλυση συναισθήματος στην πράξη. Τη σημερινή εποχή υπάρχουν πάρα πολλές εφαρμογές που μπορεί κάποιος να της βρει ελεύθερα στο ίντερνετ ή να της αγοράσει. Και πάρα πολύς κόσμος χρησιμοποιεί εφαρμογές ανάλυσης συναισθήματος που συνήθως συνοδεύονται με άλλες εφαρμογές επεξεργασίας φυσικής γλώσσας NLP. Κάθε μια χρησιμοποιεί τις δικές της μεθόδους, αλγορίθμους και χαρακτηριστικά για να πετύχει το καλύτερο δυνατό αποτέλεσμα. Συνήθως η διαδικασία που ακολουθούν είναι κάτι που δεν φανερώνεται με λεπτομέρειες, αποτελεί τη "μυστική συνταγή" του κατασκευαστή. Οι εφαρμογές της ανάλυσης συναισθήματος που συνήθως συναντάμε αφορούν κυρίως δύο τομείς. Τις ελεύθερες εφαρμογές και τις επιχειρηματικές εφαρμογές.

6.1 Ελεύθερες εφαρμογές

6.1.1 Sentiment viz

Το Tweet Sentiment Visualization, είναι μια εφαρμογή που πραγματοποιεί ανάλυση συναισθήματος στο Twitter και οπτικοποιεί τα αποτελέσματά της. Στο API που διατίθεται πληκτρολογούμε ένα keyword στο Input field, μετά πρόσφατα tweets τα οποία περιέχουν το keyword εξάγονται από το Twitter και οπτικοποιούνται στο πίνακα του σχήματος. Μπορούμε να επιλέξουμε να εμφανιστούν αποτελέσματα που αφορούν διάφορες κατηγορίες όπως:

- **Συναισθήματα (Sentiment)**. Σε άξονες: ευχαρίστηση (pleasure) και διέγερση (arousal) στον οριζόντιο και κάθετο άξονα. Και με επιπλέον κατηγορίες pleasant, happy, elated, excited, alert, active, tense, nervous, stressed, upset, unpleasant, sad, unhappy, depressed, bored, subdued, calm, relaxed, serene, contrented. Αν επιλέξουμε κάποιο σημείο μας λέει όλες τις λεπτομέρειες για την ανάλυση συναισθήματος.

- **Θέμα (Topics)**. Όπου προσδιορίζει τα tweets που συζητάμε ένα κοινό θέμα ή το θέμα. Κάθε θέμα εμφανίζεται σαν μια ορθογώνια ομάδα των tweets, με τις λέξεις-κλειδιά στην κορυφή για να συνοψίσουμε το θέμα, καθώς και μια σειρά στο κάτω μέρος για να προσδιορίσει τον αριθμό των tweets του συμπλέγματος.
- **Heatmap**. Απεικονίζει τον αριθμό των tweets μέσα σε διαφορετικές περιοχές συναίσθηματος. Τονίζει "καυτό" κόκκινες περιοχές με πολλά tweets, και "κρύο" μπλε περιοχές με λίγες μόνο tweets.
- **Tag Cloud**. Οπτικοποιεί τις πιο συχνά εμφανιζόμενους όρους σε τέσσερις συναισθηματικές περιοχές: αναστάτωση στην επάνω αριστερή γωνία, ευτυχισμένος στην επάνω δεξιά, χαλαρή στην κάτω δεξιά, και δυστυχισμένος στο κάτω-αριστερά.
- **Timeline**. Το χρονοδιάγραμμα οπτικοποιεί πότε στάλθηκαν tweets. Τα ευχάριστα tweets φαίνονται στο πράσινο πάνω από το οριζόντιο άξονα, και τα δυσάρεστα tweets σε μπλε κάτω από τον άξονα.
- **Map**. Ο χάρτης δείχνει από πού στάλθηκαν τα tweets.
- **Affinity**. Το γράφημα απεικονίζει τη συγγένεια με συχνά εμφανιζόμενα tweets, άνθρωπους, hashtags, διευθύνσεις URL.
- **Tweets**. Δείχνει την ημερομηνία, τον συγγραφέα, και το σώμα του κάθε τιτιβίσματος, καθώς και τη συνολική απόλαυση (pleasure) ή διέγερση (arousal).

Κάθε tweet ζωγραφίζεται σαν κύκλος.

Κάθε κύκλος, έχει χρώμα, φωτεινότητα, μέγεθος και διαφάνεια που φανερώνουν διάφορες λεπτομέρειες για το κάθε tweet.

- **Χρώμα (Colour)**, το ευχάριστο tweet είναι πράσινο, και το δυσάρεστο tweet είναι μπλε.
- **Φωτεινότητα (Brightness)**, ενεργό tweet είναι πιο φωτεινό, και υποτονικό tweet είναι πιο σκούρο.
- **Μέγεθος (Size)**, ένα μέτρο του δείχνει το πόσο σίγουροι είμαστε για την εκτίμηση του συναισθήματος στο tweet: μεγαλύτερα tweets αντιπροσωπεύουν εκτιμήσεις μεγαλύτερη αυτοπεποίθηση.
- **Διαφάνεια (Transparency)**, ένα δεύτερο μέτρο του πόσο σίγουροι είμαστε για την εκτίμηση του συναισθήματος στο tweet: πιο αδιαφανής (δηλαδή λιγότερο διαφανές) tweet αντιπροσωπεύουν εκτιμήσεις με μεγαλύτερη αυτοπεποίθηση.

Το συναίσθημα το εκτιμούν ως εξής:

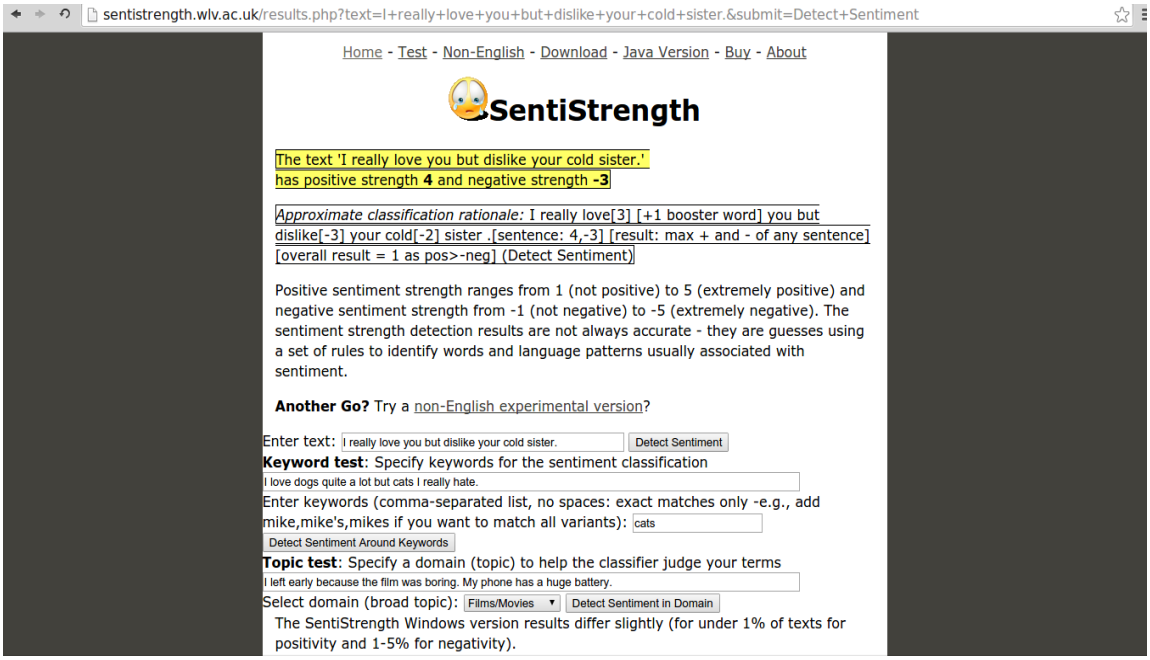
Χρησιμοποιούν ένα λεξικό συναισθήματος για να εκτιμηθεί το συναίσθημα. Αναζητούν κάθε Tweet στο λεξικό, τότε συνδυάζουν τις λέξεις, την ευχαρίστηση και διέγερση στις αξιολογήσεις για το τελικό συναίσθημα του tweet.

Υπολογιστικές μέθοδοι για την εκτίμηση συναισθήματος περιλαμβάνουν αλγόριθμους μηχανικής μάθησης όπως naive Bayes, support vector machines, και maximum entropy προσεγγίσεις, ή συνδυασμούς κοινής λογικής σκέψης και συναισθηματικής οντολογίας.

Εικόνα 6.1: Στιγμιότυπο από Tweet Sentiment Visualization

6.1.2 SentiStrength

Το SentiStrength, εκτιμά την ένταση των θετικών και αρνητικών συναισθημάτων σε κάποιο μικρό κείμενο, ακόμα και για μη επίσημη καθομιλουμένη γλώσσα. Εκτός από την πολικότητα κάθε κειμένου (θετικό/αρνητικό) υπολογίζουν και την αντίστοιχη ισχύ του συναισθήματος με εύρος τιμών 1 έως 5. Έχει πολύ καλή ακρίβεια για μικρά κείμενα που εξάγονται από τα κοινωνικά δίκτυα στην αγγλική γλώσσα. Εξαιρούνται όμως τα κείμενα που περιέχουν πολιτική. Άλλες γλώσσες που μπορεί να εξυπηρετήσει είναι τα: Φινλανδικά, Γερμανικά, Ολλανδικά, Ισπανικά. Ρωσικά, Πορτογαλικά, Γαλλικά, Αραβικά, Πολωνικά, Περσικά, Σουηδικά, Ελληνικά, Ουαλίας, Ιταλικά, Τουρκικά. Ξεκίνησε με την εργασία "Sentiment strength detection in short informal text", [TBP+10], και βελτιώθηκε στη συνέχεια με την "Sentiment strength detection for the social web", [TBP12]. Στην πρώτη εργασία χρησιμοποίησαν ένα σύνολο από 2600 σχόλια και κατασκεύασαν μία λίστα με 298 θετικούς και 465 αρνητικούς όρους ταξινομημένους ως προς την πολικότητά τους μαζί με την αντίστοιχη ισχύ τους. Συμπεριέλαβαν στην λίστα και τα emoticons, τους όρους άρνησης, τις λέξεις που αυξάνουν ή μειώνουν τη ισχύ του συναισθήματος των συμφραζόμενων όρων (booster words). Στην δεύτερη εργασία αυξάνουν τους όρους από 693 σε 2310, εισάγουν μία λίστα με ιδιώματα καθώς και την έννοια της ενίσχυσης της πολικότητας λόγω εμφατικής επιμήκυνσης.



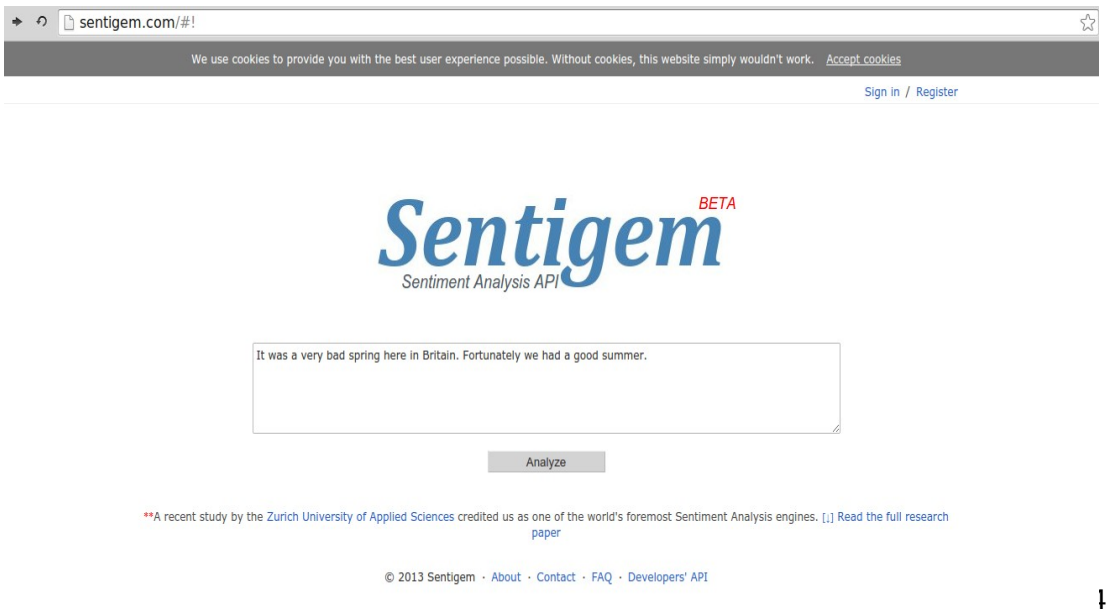
Εικόνα 6.2: Στιγμιότυπο από SentiStrength

6.1.3 Sentigem

Το Sentigem , είναι μια πλατφόρμα που υλοποιεί ανάλυση συναισθήματος σε κείμενο στα Αγγλικά. Είναι ένα API εύκολο στη χρήση του που βασίζεται σε τμήματα του κειμένου. Υπολογίζει το συνολικό συναίσθημα του κειμένου καθώς και το συναίσθημα των επιμέρους φράσεων Positive sentiment/ Negative sentiment/ Neutral sentiment.

Διαπιστώνουν κάποιους κύριους τομείς που εφαρμόζετε η ανάλυση συναισθήματος σήμερα:

- Η πρόβλεψη για την κατεύθυνση (ανοδική- πτωτική) των τιμών στις αγορές.
- Διαχείριση φήμης της εταιρείας.
- Μάρκετινγκ του προϊόντος και μελλοντικός σχεδιασμός.
- Διενέργεια δημοσκοπήσεων.



Εικόνα 6.3: Στιγμιότυπο από Sentigem

6.1.4 Πίνακας ελεύθερων εφαρμογών

Παρακάτω ακολουθεί ένας συγκεντρωτικός πίνακας στον οποίο μπορούμε να δούμε κάποιες ελεύθερες εφαρμογές πάνω στο πεδίο της αυτόματης ανάλυσης συναισθήματος.

Όνομα Εφαρμογής	**Ιστοσελίδα**	**Ίδρυμα**	**Πεδίο εφαρμογής**	**Demo API**	**Κατηγορίες Ανάλυσης Συναισθήματος**	**Άλλα αποτελέσματα**
Sentigem	http://sentigem.com/#!	Zurich University of Applied Sciences	Ανάλυση συναισθήματος σε κείμενο	http://sentigem.com/#!	Θετικό, Αρνητικό, Ουδέτερο	-
sentiment viz	http://www.csc.ncsu.edu/faculty/healey/tweet_viz/	-	Ανάλυση συναισθήματος στο Twitter, οπτικοποιεί τα αποτελέσματά της	http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/	pleasant, happy, elated, excited, alert, active, tense, nervous, stressed, upset, unpleasant, sad, unhappy, depressed, bored, subdued, calm, relaxed, serene,	Συναισθήματα (Sentiment), Θέμα (Topics), Heatmap, Tag Cloud, Timeline, Map, Affinity, Tweets

					contrented	
SentiStrength	http://sentistrength.wlv.ac.uk/	University of Wolverhampton	Μικρό κείμενο, ακόμα και για μη επίσημη καθομιλουμένη γλώσσα	http://sentistrength.wlv.ac.uk/	ένταση των θετικών και αρνητικών συναισθημάτων, με εύρος τιμών 1 έως 5	-
Tweet Annotator	http://www.tweenator.com/index.php?page_id=2	-	annotate tweets message with their sentiment labels	http://www.tweenator.com/index.php?page_id=2	Θετικό, Αρνητικό, Ουδέτερο	-
Sentilo	http://wit.istc.cnr.it/stlab-tools/sentilo/	Semantic Technology Laboratory (STLab)	sentence-based sentiment analysis	http://wit.istc.cnr.it/stlab-tools/sentilo/ui/sentence.html?	Sentiment analysis, sentilometers	visualize a semantic graph representation of a sentence enriched with opinion-related information, e.g. opinion holder, topics, sentiment scores, etc.
Umigon	http://www.umigon.com/	Rotterdam School of Management	Sentiment analysis for tweets, and more	http://www.umigon.com/	-	Positive, Negative, Promoted

Πίνακας 6.4: Ελεύθερες εφαρμογές

6.2 Εμπορικές εφαρμογές

Ανεξαρτήτως από τα αυτόματα συστήματα, ο Bing Liu λέει "η αποδεχόμενη ακρίβεια ακόμα και η μέτρησή της είναι αρκετά δύσκολη επειδή η ανάλυση συναισθήματος είναι ένα πολυδιάστατο πρόβλημα με πολλά υπό-προβλήματα".

Οι κυρίαρχοι πάροχοι όπως Sysomos και Radian6 εκτιμούν ότι με την αυτοματοποιημένη ανάλυση συναισθήματος που πραγματοποιούν μπορούν να πετύχουν ακρίβεια γύρω στο 80%

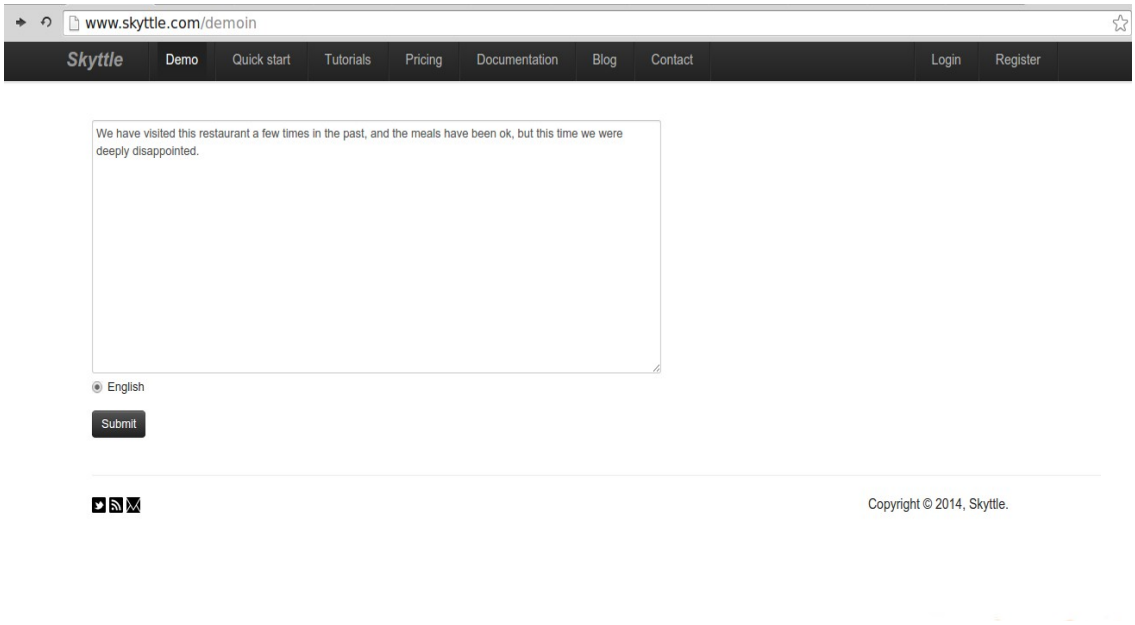
αναφέρουν στο άρθρο [13]. Αλλά ακόμα και συστήματα που δεν μπορούν να φτάσουν σε αυτά τα επίπεδα ακρίβειας δεν σημαίνει ότι είναι άχρηστα, γιατί μερικές φορές δεν είναι απαραίτητο να φτάσεις το 80% για να θεωρηθεί χρήσιμο σύστημα. Άλλες εμπορικές εφαρμογές λειτουργούν σε ένα εύρος από εφαρμογές όπως το συναισθηματικό τόνο (emotional tone checker) από τη Lymbix, η οποία στοχεύει λειτουργίες εταιρικής επικοινωνίας, και ένα παρόμοιο σύστημα από τη Adaptive Semantics που αυτοματοποιεί τα μετριοπαθή σχόλια για το χρήστη. Μερικά παραδείγματα εφαρμογών είναι τα εξής:

6.2.1 Skyttle

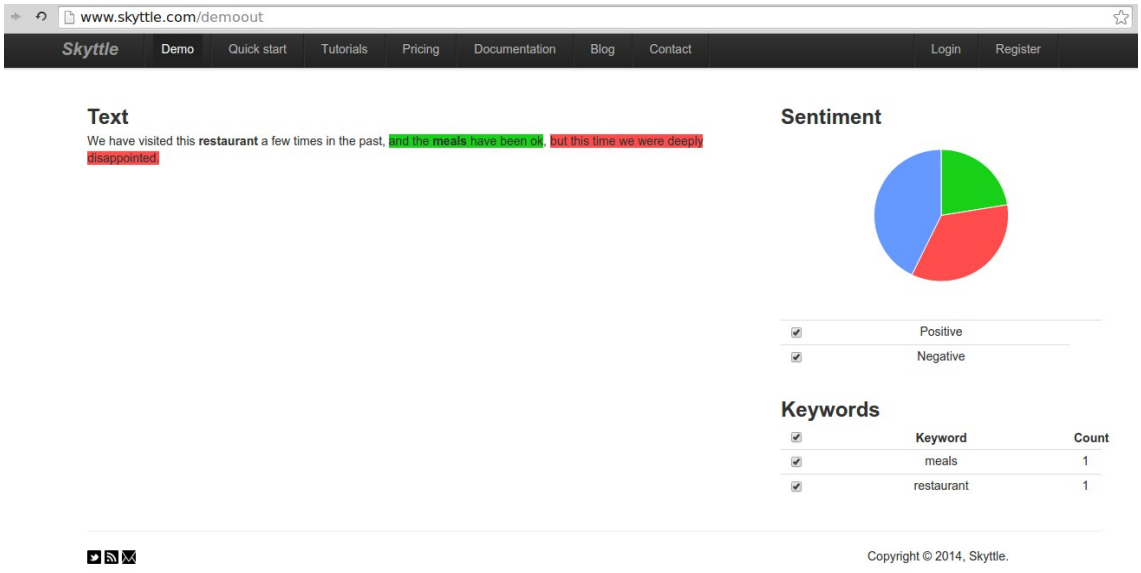
Το Skyttle, επιστρέφει κατηγορίες θετικών, αρνητικών και ουδέτερων για κομμάτια κειμένου. Υποστηρίζει τέσσερις γλώσσες English, French, German, Russian. Είναι ένα API που επιστρέφει την ανάλυση κειμένου σε επίπεδο φράσης. Η ανάλυση, περιλαμβάνει την απόδοση συναισθηματικής πολικότητας (possitive, negative, neutral) στις φράσεις του κειμένου, τον υπολογισμό ποσοστών για τις κατηγορίες συναισθήματος που εμφανίζονται και στον εντοπισμό των λέξεων κλειδιών (keywords) του κειμένου.

Εκτός από την ανάλυση συναισθήματος μπορεί να κάνει επίσης τις παρακάτω λειτουργίες:

- Εξαγωγή λέξεων κλειδιών (Keyword extraction with Skyttle)
- Σχολιασμό των κειμένων για το συναίσθημα και τις λέξεις-κλειδιά (Annotating texts for sentiment and keywords)
- Εύρεση συναισθήματος που σχετίζεται με οντότητες και λέξεις κλειδιά (Finding sentiment associated with entities and keywords).



Εικόνα 6.5: Στιγμιότυπο από Skyttle

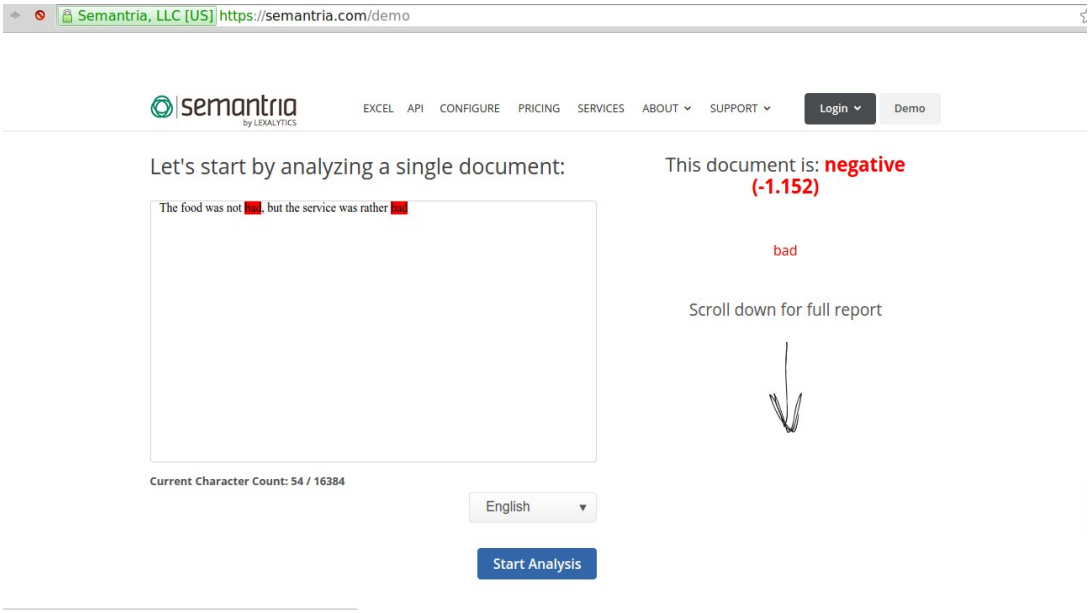


Εικόνα 6.6: Στιγμιότυπο από εκτέλεση Skyttle

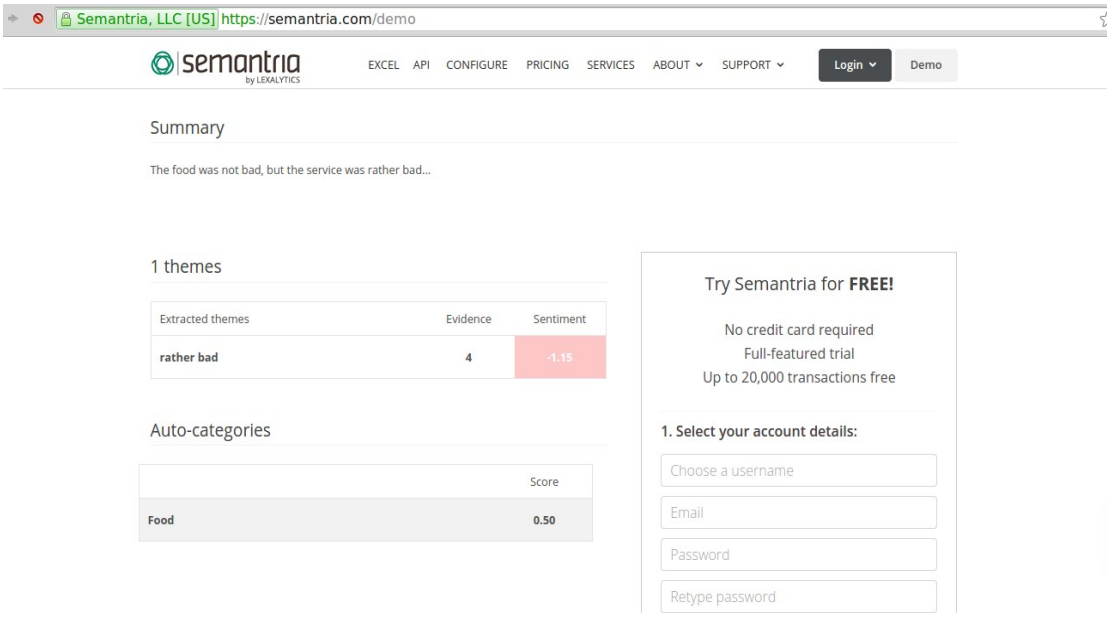
6.2.2 Semantria

Το Semantria, χρησιμοποιεί εργαλεία κειμένου για να κάνει ανάλυση συναισθήματος στα tweets, το Facebook, τις κριτικές (reviews), έρευνες (surveys), σχόλια, ή τα περιεχόμενα επιχειρήσεων. Επιτρέπει την ανάλυση οποιουδήποτε κειμένου μέχρι 16384 χαρακτήρες. Είναι πλήρως εξοπλισμένο, για να υποστηρίζει παραπάνω από δέκα γλώσσες Αγγλικά, Γαλλικά, Πορτογαλικά, Ισπανικά, Γερμανικά, Κινέζικα (Mandarin), Ιταλικά, Κορεατικά, Ιαπωνικά, Ολλανδικά (beta). Το Semantria είναι χτισμένο με τεχνολογίες αιχμής και επιχειρεί στο πεδίο ανάλυσης κειμένων. Από την ανάλυση κειμένου προκύπτει το συνολικό του συναίσθημα που μπορεί να είναι αρνητικό, θετικό, ή ουδέτερο, μαζί με την αντίστοιχη βαθμολογία του συναισθήματος. Εξάγονται οι οντότητες που εμφανίζονται σε αυτό , οι κατηγορίες στις οποίες ανήκουν οι οντότητες, τα θέματα στα οποία αναφέρεται το κείμενο και η περίληψη του. Για την ανάλυση συναισθήματος χρησιμοποιούν συντακτική ανάλυση του κειμένου και μετά εντοπίζουν φράσεις με συναισθηματικό περιεχόμενο. Στο τέλος η βαθμολογία προκύπτει από το συνδυασμό των βαθμολογιών των επιμέρους φράσεων.

Αυτό που καθορίζει πραγματικά το Semantria και το κάνει να ξεχωρίζει από τις άλλες επιχειρήσεις και cloud NLP μηχανές είναι η παραμετροποίηση. Η κατηγοριοποίηση και η εξαγωγή οντοτήτων μπορεί εύκολα να εκπαιδευτεί ώστε να ταιριάζει με τα ειδικά λεξιλόγια κάθε επιχείρησης. Κάθε πτυχή της ανάλυσης συναισθήματος μπορεί να παραμετροποιηθεί μέχρι τα αποτελέσματα να ταιριάζουν με τις ανάγκες του πελάτη.



Εικόνα 6.7: Στιγμιότυπο από Semantria



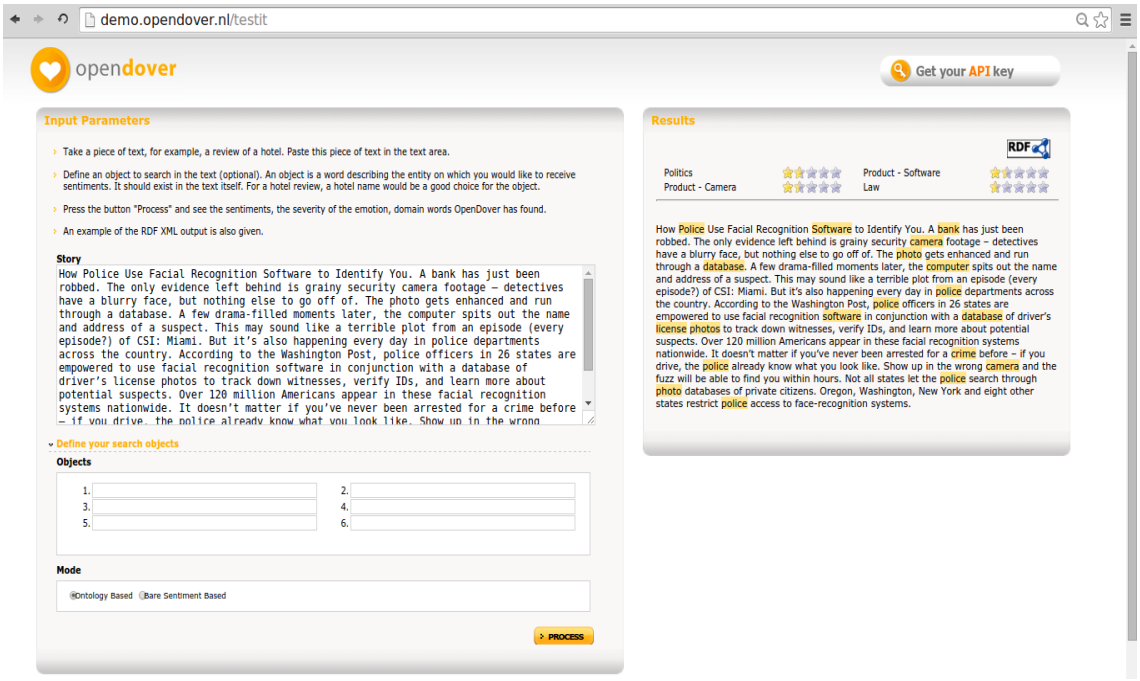
Εικόνα 6.8: Στιγμιότυπο από εκτέλεση Semantria

6.2.3 OpenDover

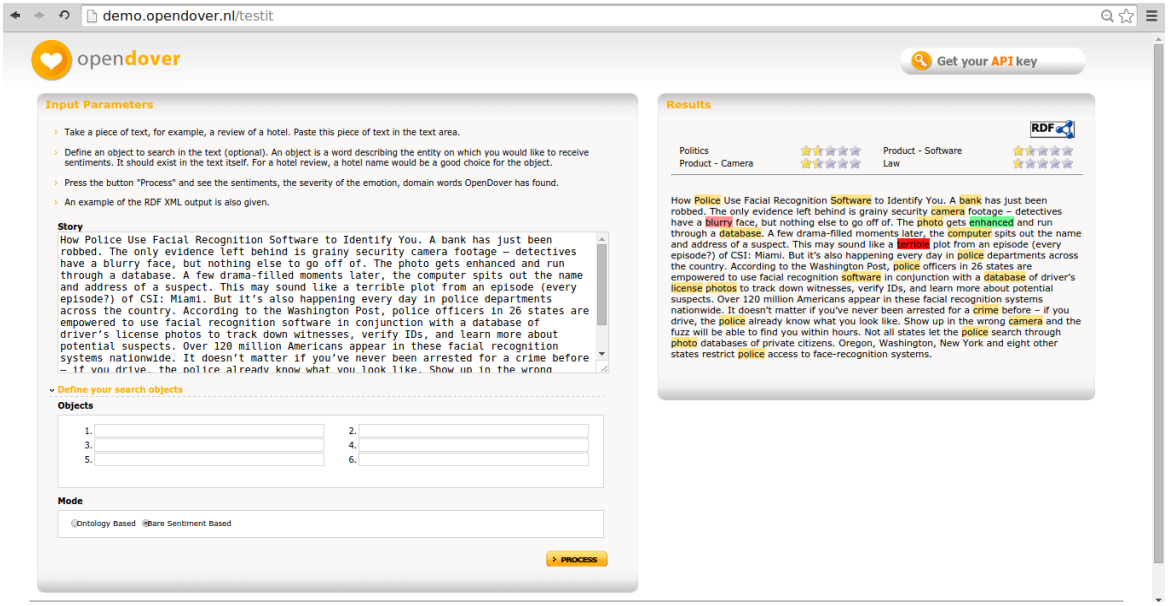
Το OpenDover , είναι μια εφαρμογή που σου επιτρέπει να εξάγεις χαρακτηριστικά συναισθήματος από blogs, content management systems, websites ή άλλες ποικίλες εφαρμογές. Το OpenDover χρησιμοποιεί σημασιολογική τεχνολογία για το συναισθηματικό καθορισμό των κειμένων. Διαπιστώνουν την ανάγκη για ανάλυση συναισθήματος τη σημερινή εποχή μιας και υπάρχει όλο και αυξανόμενη χρήση του διαδικτύου. Με το OpenDover API μπορείς να εισάγεις το κείμενο και να επιλέξεις ανάμεσα σε δύο κατηγορίες ανάλυσης:

1. Βασισμένη στις οντότητες (Ontology based) και
2. Βασισμένη στο συναίσθημα (Bare Sentiment Based).

Επιπλέον μπορείς να δηλώσεις χειροκίνητα μέχρι 6 αντικείμενα (objects) που θέλεις να δώσεις έμφαση στην ανάλυση.



Εικόνα 6.9: Στιγμιότυπο από OpenDover



Εικόνα 6.10: Εκτέλεση OpenDover

Ο Seth Grimes ιδρυτής της εταιρίας Alta Plana [14] και διοργανωτής του συνεδρίου για την ανάλυση συναισθήματος (Sentiment Analysis Symposium) κάθε χρόνο [15], δείχνει σε μια παρουσίαση ένα ενδιαφέρον σχήμα με τις πιο δημοφιλείς εφαρμογές που υπάρχουν στον τομέα της ανάλυσης συναισθήματος το 2014 [16].

Εικόνα 6.11: Εταιρίες ανάλυσης συναισθήματος

Ακολουθεί ένας πίνακας με τις κυριότερες εφαρμογές συναισθηματικής ανάλυσης που υπάρχουν και αφορούν περισσότερο πραγματικές εφαρμογές και επιχειρήσεις.

6.2.4 Πινάκας επιχειρηματικών εφαρμογών

Όνομα Εφαρμογής	**Ιστοσελίδα**	**Πεδίο εφαρμογής**	**Γλώσσες**	**Demo API**	**Κατηγορίες Ανάλυσης Συναισθήματος**	**Άλλες λειτουργίες**
Textalytics	https://www.meaningcloud.com/	Κοινωνικά δίκτυα, forums, blogs αλλά και sites ειδήσεων	Αγγλικά, Ισπανικά	https://textalytics.com/api-text-analysis-demo-en	Θετικό, Αρνητικό, Ουδέτερο	Εύρεση οντοτήτων, χαρακτηριστικές πληροφορίες όπως χρονικές εκφράσεις
Bittext	http://www.bitext.com/	Ανάλυση κείμενου	Αγγλικά, Ισπανικά, Γαλλικά, Πορτογαλικά, Ιταλικά, Γερμανικά, Ολλανδικά, Καταλανικά, Ρώσικα (beta) Βασικά(beta)	http://www.bitext.com/text-analysis-api/	Εξάγει συναίσθημα	Αναγνώριση οντοτήτων, εννοιών και κατηγοριών κειμένου

Repustate	https://www.repustate.com/	Συναισθηματική ανάλυση για κοινωνικά δίκτυα	Αγγλικά, Αραβικά, Κινεζικά, Γερμανικά, Γαλλικά, Ισπανικά, Ιταλικά, Ρωσικά, Πολωνικά	https://www.repustate.com/api-demo/	Τιμές μεταξύ του -1 και του 1, καθορίζοντας με αυτόν το τρόπο το αρνητικό και το θετικό συναίσθημα	-
Skyttle	http://www.skyttle.com/	Ανάλυση κειμένου σε επίπεδο φράσης	Αγγλικά, Γαλλικά, Γερμανικά, Ρωσικά	http://www.skyttle.com/demoin	Θετικό, Αρνητικό, Ουδέτερο	υπολογισμό ποσοστών για τις κατηγορίες συναισθήματος που εμφανίζονται, Εντοπισμό των keywords του κειμένου, Σχολιασμό των κειμένων για το συναίσθημα και τις λέξεις-κλειδιά, Εύρεση συναισθήματος που σχετίζεται με οντότητες και λέξεις κλειδιά
Semantria	https://semantria.com/	Ανάλυση συναισθήματος στα tweets, το Facebook, τις κριτικές (reviews), έρευνες (surveys), σχόλια, ή τα περιεχόμενα επιχειρήσεων	Αγγλικά, Γαλλικά, Πορτογαλικά, Ισπανικά, Γερμανικά, Κινέζικα (Mandarin), Ιταλικά, Κορεατικά, Ιαπωνικά, Ολλανδικά (beta)	https://semantria.com/demo	αρνητικό, θετικό, ή ουδέτερο και αντίστοιχη βαθμολογία του συναισθήματος	Εξάγονται οι οντότητες που εμφανίζονται σε αυτό, οι κατηγορίες στις οποίες ανήκουν οι οντότητες, τα θέματα στα οποία αναφέρεται το κείμενο και η περίληψη του
Lymbix	http://www.lymbix.com/	Κείμενο μέχρι 20,000 λέξεων	Αγγλικά	http://www.lymbix.com/live-demo	Αποφαίνονται για την ποικιλία των συναισθημάτων που παρουσιάζονται σε κάποιο κείμενο	Article Affection/Friendliness Amusement/Excitement Contentment/Gratitude

						Enjoyment/Elatio Anger/Loathing Fear/Uneasiness Humiliation/Shame Sadness/Grief Dominant Emotion Intense Sentence Article Sentiment Coverage Clarity
OpenDover	http://www.opendover.nl/	blogs, content management systems, websites ή άλλες ποικίλες εφαρμογές	Αγγλικά	http://demo.opendover.nl/testit	Εξάγει χαρακτηριστικά συναισθήματος	6 αντικείμενα (objects) που θέλεις να δώσεις έμφαση στην ανάλυση.
uClassify	https://www.uclassify.com/	Κείμενο	Κάνει αναγνώριση κειμένου	https://www.uclassify.com/browse/uclassify/sentiment	θετικό ή αρνητικό	Γλώσσα κειμένου, Θεματολογία, Ανάλυση φύλου, Ανάλυση διάθεσης
Sentiment 140	http://www.sentiment140.com/	Εξερευνά το Twitter για συναισθήματα που αφορούν προϊόντα	Αγγλικά, Ισπανικά, αυτόματη αναγνώριση γλώσσας	http://help.sentiment140.com/api	Πολικότητα: 0: αρνητικό 2: ουδέτερο 4: θετικό	-
Miopia	http://miopia.grupolys.org/	Συναισθηματική ανάλυση κοινωνικών δικτύων	Αγγλικά, Ισπανικά	http://miopia.grupolys.org/demo/	Η επιβλεπόμενη μέθοδος σε 5 κατηγορίες: P+, P, NONE, N, N-, η rule-based μέθοδος 3 κατηγορίες: positive, none, negative	-
Twinword	https://www.twinword.com/	Κείμενο	Αγγλικά	https://www.twinwor	Θετικό ή αρνητικό	Βαθμολογία, λέξεις κλειδιά, Οπτικοποιημένα

				d.com/sentiment-analysis.php		αποτελέσματα (Semantic Vizualization)
TheySay	http://www.theysay.io/	text analytics tool	Αγγλικά	http://apidemo.theysay.io/#sentimentRelationsTab	Θετικό, Αρνητικό, Ουδέτερο συναίσθημα	Απόψεις, συναισθήματα, θεματολογία και οπτικής γωνίας
Weather Sentiment Prediction	-	Ανάλυση καιρού μέσω Tweets	Αγγλικά	http://www.sproutloop.com/prediction_demo/	-	Οπτικοποίηση αποτελεσμάτων, εμφάνιση των μηνυμάτων
AlchemyAPI	http://www.alchemyapi.com/	Κείμενο	Αναγνώριση γλώσσας ανάμεσα σε 97 γλώσσες	http://www.alchemyapi.com/products/demo/	Θετικό, Αρνητικό, Μεικτό	Εξόρυξη Οντότητας, Εξόρυξη λέξης-κλειδιού, Χαρακτηρισμός έννοιας, Εξόρυξη σχέσεων, Κατηγοριοποίηση ταξονομίας, Εύρεση συγγραφέα, Εντοπισμός γλώσσας, Εύρεση εισόδου, Υποστήριξη διασυνδεδεμένων δεδομένων, Διαχείριση της μάρκας και υποστήριξης πελατών, Ευφυΐα πρόβλεψης, Περιεχόμενο και μηνύματα
Angoss	http://www.angoss.com/predictive-analytics-software/applications/text-analy	Text Analytics	Αγγλικά	http://www.angoss.com/request-demo/	Θετικό, Αρνητικό, Ουδέτερο	Εξαγωγή περιεχομένου και οντοτήτων , θεματολογίας κατηγοριοποίηση θέματος, δυνατότητα

	tics/					περίληψης κειμένου
Clarabridge	http://www.clarabridge.com/	Ανάλυση κειμένου από Twitter, Facebook, Yelp!, product forums, κτλ και από πηγές όπως σημειώσεις από τηλεφωνικό κέντρο, αποθήκες, CRM, BI, emails	Αγγλικά	http://www.clarabridge.com/demo/	Παρέχει ανάλυση συναισθήματος	Ανάλυση κειμένου (text mining), (NLP), Συσταδοποίηση και κατηγοριοποίηση κειμένου, SaaS (Λογισμικό ως υπηρεσία)
General Sentiment	http://www.generalsentiment.com/	Blogs, Forums, Twitter, Facebook και συλλογές σχολίων	Αγγλικά	http://www.generalsentiment.com/request-a-demo/	-	-
MonkeyLearn	http://www.monkeylearn.com/	Κείμενο	Αγγλικά	-	Συναισθηματική ανάλυση	Κατηγοριοποίηση κειμένου ανά τομέα (industry/domain)
NetOwl	https://www.netowl.com/	Παραδοσιακές πηγές όπως (news, reports, web pages, email) και από κοινωνικά δίκτυα όπως (Twitter, Facebook, chats, blogs)	Αγγλικά και άλλες ξένες γλώσσες	-	Πραγματοποιεί ανάλυση συναισθήματος	Εξαγωγή οντοτήτων, διασυνδέσεων (links), γεωγραφικές πληροφορίες για το κείμενο, μετάφραση ονομάτων και ταίριασμα για εξαγωγή πληροφοριών
Sysomos	http://sysomos.com/	Ιστοσελίδες κοινωνικών δικτύων όπως blogs, forums, Twitter	Αγγλικά	http://sysomos.com/products/map	Συναισθηματική ανάλυση από διαδικτυακές συνομιλίες καταναλωτών	Καταγραφή κοινωνικών δικύων και brand managers and customer support groups
Radian6	http://www.exacttarget.com/products/s	Κείμενο, κοινωνικά δίκτυα	Αγγλικά	https://socialstudio.ra	Συναισθηματική ανάλυση	Μάρκετινγκ Email Mobile, Social Media Διαφημίσεις,

	ocial-media-marketing/radian6			dian6.com/login		Διαχείριση ταξιδίων, Προγνωστική ευφυΐα
Lexalytics	http://www.lexalytics.com/	Ανάλυση συναισθήματος στα tweets, το Facebook, τις κριτικές (reviews), έρευνες (surveys), σχόλια, ή τα περιεχόμενα επιχειρήσεων	Αγγλικά, Γαλλικά, Πορτογαλικά, Ισπανικά, Γερμανικά, Κινέζικα, Ιταλικά, Κορεατικά, Ιαπωνικά	https://semantria.com/demo	Αρνητικό, θετικό, ή ουδέτερο και αντίστοιχη βαθμολογία του συναισθήματος	Εξάγονται οι οντότητες που εμφανίζονται σε αυτό, οι κατηγορίες στις οποίες ανήκουν οι οντότητες, τα θέματα στα οποία αναφέρεται το κείμενο και η περίληψη του
Brandwatch	https://www.brandwatch.com/?utm_expid=26799417-2.t5gli-Q8Sn-YW4EpmOeYCA.0&utm_referrer=https%3A%2F%2Fwww.brandwatch.com%2Fdemo%2F	Αναλύουν πλούσια δεδομένα από τα κοινωνικά δίκτυα για να βελτιώσουν το μαρκετινγκ επιχειρήσεων	Αγγλικά	https://www.brandwatch.com/demo/	Χρήση συναισθηματικής ανάλυσης	Κατηγοριοποίηση, οπτικοποίηση αποτελεσμάτων, αναλυτικά εργαλεία για περισσότερες πληροφορίες
IBM	http://www.ibm.com/big-data/us/en/big-data-and-analytics/index.html	Ανάλυση κοινωνικών δικτύων	Αγγλικά	-	Ανάλυση συναισθήματος	Watson Analytics Predictive analytics Streaming analytics Business Inteligence Advanced case management
Simply Mesasured	http://simplymeasured.com/#i.2v9lodiebdgwzf	Ολοκληρωμένες υπηρεσίες ανάλυσης σε κοινωνικά δίκτυα	Αγγλικά	http://get.simplymeasured.com/trial#i.2v9lodiebdgwzf	Συναισθηματική ανάλυση	Twitter Follower Analysis, Instagramm user analysis, Facebook insights analysis, Facebook content analysis, Google+ page analysis, LinkedIn company analysis, Vine analysis,Twitter customer servece analysis,

						Social traffic analysis, Traffice source analysis
SAP	http://scn.sap.com/welcome	-	Αγγλικά	-	Χρήση εργαλείων συναισθηματικής ανάλυσης	-
SAS	http://www.sas.com/en_us/software/analytics/sentiment-analysis.html	Κοινωνικά δίκτυα, διαδίκτυο.	Αγγλικά	http://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-sentiment-analysis-104357.pdf	Δυναμική συναισθηματική ανάλυση	Εξόρυξη δεδομένων, Στατιστικές αναλύσεις, Προγνώσεις. Ανάλυση κειμένων. Προσομοιώσεις
TEMIS	http://www.temis.com/home	-	Αγγλικά, και έξι ακόμα γλώσσες	-	-	-
Attensity	http://www.attensity.com/	Κείμενα, κοινωνικά δίκτυα	Αγγλικά	-	Αναλυτική συναισθηματική ανάλυση	Κατανόηση συμπεριφοράς πελάτη, Σημασιολογική κατηγοριοποίηση, Προγνωστική ανάλυση, κτλ
Crimson Hexagon	http://www.crimsonhexagon.com/	Κοινωνικά δίκτυα	Αγγλικά	http://www.crimsonhexagon.com/#schedule-a-demo	Θετικό, αρνητικό, ουδέτερο, συναισθηματική ανάλυση.	Αναγνώριση θεματολογίας, Ανάλυση κειμένου, Καταγραφή κοινωνικών δικτύων,

Πίνακας 6.12: Εφαρμογές που αφορούν επιχειρήσεις

6.3 Παραδείγματα εφαρμογών

Οι εφαρμογές, οι τεχνικές και τα εργαλεία που περιγράφτηκαν πιο πάνω έχουν εφαρμογή στον πραγματικό κόσμο. Πλέον η ανάλυση συναισθήματος είναι όλο και πιο αναγκαία και η χρήση της είναι πολύ διαδεδομένη. Αυτό γίνεται εμφανές μέσα από μερικά παραδείγματα εφαρμογών ανάλυσης συναισθήματος (case studies) που παρουσιάζονται παρακάτω.

6.3.1 Μέτρηση απήχησης γεγονότων

Όταν το ESPN και η TFL (Transport For London) ένωσαν τις δυνάμεις τους για την κάλυψη του World Cup 2014, όλοι περίμεναν ότι θα ήταν μια εξαιρετική συνεργασία. Θα έδειχναν τα αποτελέσματα των αγώνων του World Cup στους πίνακες του μετρό του Λονδίνου. Η ESPN εκπλήρωσε την υποχρέωσή της να ενημερώνει το κοινό και τους οπαδούς συνέχεια και παντού. Αλλά κάπως έπρεπε να μετρηθεί η επιτυχία του σχεδίου αυτού, σε αυτό βοήθησε η Brandwatch. Η οποία χρησιμοποίησε δημογραφικά στοιχεία. Μέτρησε σε ποια μέρη στο London Underground η εκστρατεία είχε τα καλύτερα αποτελέσματα, και το περισσότερο κοινό. Και πόσες θετικές αναφορές είχε η ενέργειά τους αυτή, μετρήθηκε πάνω από 60% θετικές αναφορές. Η ESPN χρησιμοποίησε το Brandwatch για να αξιοποιήσει δημογραφικά στοιχεία αλλά και ανάλυση συναισθήματος για να εξάγει συμπεράσματα για παραπάνω από 2 εκατομμύρια χρήστες του Twitter αναφέρουν εδώ [17].

6.3.2 Επικοινωνία με δημότες

Το 2013 στη Τουλούζη, ο δήμος ήθελε να βρει νέους τρόπους για να συλλάβει τις ανησυχίες των πολιτών της. Προσπαθούν να προσφέρουν στους κατοίκους ένα ευρύ φάσμα υπηρεσιών, που περιλαμβάνουν προγράμματα οικονομικής ανάπτυξης, του περιβάλλοντος, της υγείας, της εκπαίδευσης, της αναψυχής και άλλα. Σε μία πόλη που χαρακτηρίζεται από την άνθηση της αεροδιαστημικής και βιομηχανίες υψηλής τεχνολογίας, η κυβέρνηση της πόλης εργάζεται για να εξυπηρετήσει κατάλληλα τους κατοίκους της, μετατρέποντας τις παραδοσιακές μεθόδους της κυβέρνησης σε πλατφόρμες ηλεκτρονικής διακυβέρνησης και υλοποίηση ενός δικτυακού τόπου, καθώς και εφαρμογές για κινητά και για κοινωνικά δίκτυα.

Για το λόγο αυτό η κυβέρνηση χρησιμοποιεί ανάλυση κοινωνικών δικτύων και δεδομένων για να αναλύσει τις ανάγκες των πολιτών που αναρτούνται στα κοινωνικά δίκτυα, λαμβάνοντας υπόψη παράγοντες, όπως το πλαίσιο, το περιεχόμενο και το συναίσθημα. Έτσι, γνωρίζοντας καλά τις ανησυχίες και τις προσδοκίες των κατοίκων, η κυβέρνηση της πόλης της Τουλούζης μπορεί να ενισχύσει τις δημόσιες σχέσεις της, τον αστικό σχεδιασμό και την αναπτυξιακή πολιτική.

Για να επιτευχθεί αυτό χρειάστηκε να χρησιμοποιηθούν υπηρεσίες από την IBM [18] όπως Διαχείριση της σχέσης με τον Πελάτη (Customer Relationship Management), Ευφυές μάρκετινγκ (Smarter Marketing), (Smarter Planet), Κοινωνικές επιχειρήσεις και εξυπηρέτηση πελατών (Social Business for Customer Service).

6.3.3 Εξυπηρέτηση προβλημάτων σε εταιρείες

Ο Rami Nuseir έγραψε [19] για μια μεγάλη αεροπορική εταιρεία ξεκίνησε να καταγράφει και να παρακολουθεί τα tweets που αφορούν τις πτήσεις τους για να δουν πώς αισθάνονταν οι

πελάτες για τις καθυστερήσεις, τις αναβαθμίσεις, τα νέα αεροπλάνα, και πολλά άλλα. Επιπλέον άρχισαν να χρησιμοποιούν την πλατφόρμα υποστήριξης των πελατών τους (ZenDesk) και την επίλυσή τους σε πραγματικό χρόνο. Μια φορά ένας πελάτης έγραψε αρνητικά σχόλια για χαμένες αποσκευές πριν επιβιβαστεί στην πτήση του. Η αεροπορική εταιρεία χρησιμοποιώντας ανάλυση συναισθήματος κατάφερε και κατέγραψε αμέσως το tweet αυτό, και το διαβίβασε στον υπεύθυνο έτσι ώστε να προσφέρει στο πελάτη μια δωρεάν αναβάθμιση στην πρώτη τάξη στο ταξίδι της επιστροφής. Επίσης παρακολουθούσαν την αποσκευή, και του έδωσαν πληροφορίες σχετικά με το πού ήταν, και πού θα του την παρέδιδαν τη στιγμή που θα έβγαινα από το αεροπλάνο. Ο πελάτης με αυτό τον τρόπο έμεινε κατευχαριστημένος και για πάντα θα λέει πόσο καλά τον εξυπηρέτησε αυτή η εταιρεία.

6.3.4 Αγορά Μπύρας

Σε μια ανάλυση συναισθήματος της βιομηχανίας μπύρας χρησιμοποιώντας Simply Measured, είναι ξεκάθαρο ότι η Bud Light είναι η νικήτρια από όλες στα Facebook Likes. Αλλά όμως, με μια πιο προσεκτική ματιά μπορεί να δεί κανείς ότι και η Coors Light έχει τον ίδιο αριθμό από σχόλια όπως και η Bud Light. Για την ακρίβεια αν συγκρίνει κανείς τα σχόλια αν like για τις δύο αυτές εταιρίες, φαίνεται σαν η Coors Light να υπερέχει από την Bud Light. Η Susan υποστηρίζει [20], ότι το "Like" είναι μια πιο μικρή ένδειξη – δέσμευση από το σχόλιο που είναι κάτι πιο σοβαρό. Αν πάρουμε το μέρος της εταιρείας είναι προφανές ότι θα μας ενδιέφερε να μάθουμε με ποιο τρόπο η εταιρεία της Coors Light κατορθώνει να κάνει αυτούς που την συμπαθούν να σχολιάζουν περισσότερο αν like. Αυτός ο τρόπος ανταγωνιστικής ανάλυσης είναι που βοηθά τις εταιρίες να βελτιώσουν τις κοινωνικές τακτικές τους και τα κόλπα τους.

6.3.5 Διόρθωση βλαβών

Όταν υπηρεσία της DIRECTV για εξυπηρέτηση πελατών και επιδιορθώσεις βλαβών πήγαινε από το κακό στο χειρότερο, το κόστος της αντιμετώπισης μεμονωμένων πελατών γινόταν πάρα πολύ υψηλό. Κάθε εταιρία που προσφέρει παραδοσιακά πολύ μεγάλη υποστήριξη πελατών μέσω τηλεφωνικού κέντρου μπορεί να χρησιμοποιήσει το παράδειγμα από την επικοινωνιακή στρατηγική της DIRECTV [20].

Η DIRECTV έχει αντιληφθεί ότι οι άνθρωποι συνηθίζουν να λένε το πρόβλημά τους πρώτα στα κοινωνικά δίκτυα πριν από κάνουν κλήση σε υπηρεσία υποστήριξης πελατών. Συγκρίνοντας το κόστος του τηλεφωνικού κέντρου dial-in με το κόστος των προληπτικών tweets, σίγουρα τα tweets είναι πολύ πιο οικονομικά. Χρησιμοποιώντας την παρακολούθηση των κοινωνικών δικτύων επιτρέπεται να επιλύονται θέματα εξυπηρέτησης πριν γίνει κρίσιμη η κατάσταση. Ο Miller, Διευθυντής, Στρατηγικής για την Κοινωνική Υποστήριξη Media, λέει

"Αν μπορούμε να διαπιστώσετε προβλήματα στην εκπομπή στα Κοινωνική δίκτυα γρήγορα, μπορούμε να ανιχνεύσουμε την ταχύτητα, να κρατήσουμε τους πελάτες ενήμερους, και να βοηθηθούν οι πελάτες τόσο εντός όσο και εκτός σύνδεσης. Τα κοινωνικά δίκτυα αποτελούν έτσι το σύστημα έγκαιρης προειδοποίησης".

6.3.6 Το Twitter στη Wall Street

Το μέχρι πρότινος μπλοκαρισμένο Twitter στη Wall Street, έκανε το μεγάλο ντεμπούτο του στα γραφεία συνδιαλλαγών μέσω τερματικών του Bloomberg αναφέρεται στο άρθρο [21].

Η Bloomberg LP ανακοίνωσε την ενσωμάτωση των tweets στην υπηρεσία δεδομένων της, η οποία χρησιμοποιείται ευρέως στον χρηματοοικονομικό κλάδο. Το νέο χαρακτηριστικό επιτρέπει στους εμπόρους και άλλους επαγγελματίες την παρακολούθηση των κοινωνικών μέσων μαζικής ενημέρωσης και τις σημαντικές ειδήσεις για τις εταιρείες που παρακολουθούν.

Αυτή η άφιξη του Twitter έρχεται πλαγίως γιατί οι μεγάλες τράπεζες της Wall Street είχαν απαγορεύσει σε μεγάλο βαθμό τη χρήση του Twitter και άλλων κοινωνικών μέσων ενημέρωσης κατά την εργασία, αναφέροντας τους κανονισμούς που διέπουν την επικοινωνία. Αν και ορισμένες επιχειρήσεις που επέτρεπαν σε ορισμένους υπαλλήλους να δουλεύουν πάνω στα μέσα κοινωνικής δικτύωσης, η χρήση ελέγχονταν πολύ.

Τώρα, οι τραπεζικοί υπάλληλοι μπορούν να έχουν μια ευρύτερη και πιο οργανωμένη εικόνα για το τι λέγεται στον κόσμο του Twitter. Κάποιοι στην Wall Street χρησιμοποιούν ήδη τα κινητά τους τηλέφωνα για να παρακολουθούν την ιστοσελίδα για πληροφορίες του πώς θα μπορούσαν να κινηθούν τα αποθέματα.

Η νέα υπηρεσία του Bloomberg δείχνει tweets ταξινομημένα ανά εταιρεία και θέμα, επιτρέποντας στους χρήστες να ψάξουν ανάλογα με τη λέξη κλειδί και να ρυθμίσουν ειδοποιήσεις για μια συγκεκριμένη εταιρεία όταν είναι να πάρει ασυνήθιστη προσοχή.

"Είχαμε πάρει αιτήματα από πελάτες που έβλεπαν ειδήσεις και ήθελαν να γνωρίζουν το τι γράφεται στο Twitter," δήλωσε ο Μπράιαν Ρούνεϊ, διευθυντής ειδήσεων στο Bloomberg, ο οποίος είπε ότι οι αξιωματούχοι από τράπεζες της Wall Street είχαν εκδηλώσει ενδιαφέρον στο να επιτρέψουν στους υπαλλήλους να βλέπουν τα tweets.

Όμως θέλει μεγάλη προσοχή γιατί στο Twitter μπορούν να κυκλοφορήσουν πολλές ανακρίβειες για αυτό το Bloomberg θα δείχνει tweets από τις εταιρείες, από διευθύνοντες σύμβουλους και άλλες ειδήσεις από ιθύνοντες, επιπλέον και από ορισμένους οικονομολόγους και οικονομικούς bloggers. Ο κ. Ρούνεϊ αναφέρθηκε στους οικονομολόγος Nouriel Roubini και Paul Kedrosky ως παραδείγματα.

Επίσης η Thomson Reuters, [22] είναι μια από τις μεγάλες πολυεθνικές στο τομέα των μέσων μαζικής ενημέρωσης και πληροφόρησης, η εταιρεία ιδρύθηκε στο Τορόντο και εδρεύει στη Νέα Υόρκη. Στο άρθρο αυτό [23], μας δείχνει το πώς μπορεί να συνεχίζει βελτιώνει τη θέση

της. Η εταιρία αυτή εντάσσει τώρα την ανάλυση συναισθήματος από το Twitter για την εφαρμογή Eikon που κάνει ανάλυση αγοράς και είναι μια πλατφόρμα συναλλαγών.

Αυτό το μοντέλο είχε εφαρμοστεί πρώτα από την αντίπαλο εταιρία Bloomberg, εδώ όμως η εφαρμογή της Thomson Reuters πηγαίνει ένα βήμα παραπέρα, δημιουργώντας απεικονίσεις (visualizations) και τα διαγράμματα που βασίζονται σε αυτό το είδος των δεδομένων. Κοιτάζοντας τα γραφικά, οι έμποροι (traders) και οι άλλοι χρήστες του Eikon θα είναι σε θέση να εξετάσουν περαιτέρω τα στοιχεία για την παρακολούθηση συγκεκριμένων Tweets, ανθρώπους και εταιρείες στο Twitter.

Προς το παρόν, η ανάλυση συναισθήματος χρησιμοποιεί μόνο το Twitter, αλλά η Thomson Reuters εργάζεται για την προσθήκη περισσότερων πηγών περιεχομένου, συμπεριλαμβανομένων των blogs. Η Thomson Reuters πιστεύει ότι είναι η πρώτη κύρια πλατφόρμα που αφορά τα χρηματοοικονομικά και παρέχει συναισθηματική ανάλυση από το twitter με αυτόν τον τρόπο σε ευρεία κλίμακα. Το Eikon έχει 120.000 άτομα που χρησιμοποιούν την υπηρεσία στην επιφάνεια εργασίας, και ότι αυξάνεται εκθετικά κάθε εβδομάδα.

Με την Επιτροπή Κεφαλαιαγοράς επισήμως να αναγνωρίζει ότι οι επιχειρήσεις μπορούν να επικοινωνούν τα νέα, νόμιμα μέσω του Twitter. Το Twitter και η ανάλυση συναισθήματος σε αυτό έχει γίνει όλο και περισσότερο κάτι που ενδιαφέρει όλους τους επιχειρηματίες και τους επενδυτές. Η Thomson Reuters αναγνωρίζει ότι “το 50% των επιχειρήσεων χρησιμοποιούν μηχανήματα για αναγνώριση ειδήσεων (news feeds)".

“Τα οικονομικά που συνδέονται με τη συμπεριφορά (Behavioral Finance) είναι ένας τομέας αυξανόμενου ενδιαφέροντος στις χρηματοπιστωτικές αγορές. Ωστόσο, ήταν δύσκολο για τους ανθρώπους να παρακολουθούν λόγω του μεγάλου όγκου και τις λεπτομέρειες των δεδομένων και την ανάγκη να το ερμηνεύσει και να εντοπίσουν τις τάσεις αμέσως” είπε ο Philip Brittan, προϊστάμενος τεχνολογίας.

6.3.7 Το Twitter στη μουσική βιομηχανία

Το άρθρο των NewYorkTimes [24], αναφέρεται η επίδραση του Twitter στη μουσική, και πώς εταιρίες όπως η 300 [25], προσπαθούν να εκμεταλλευτούν την πληροφορία που υπάρχει μέσα στο Twitter και να βγάλουν χρήσιμα συμπεράσματα χρησιμοποιώντας και τεχνικές ανάλυσης συναισθήματος. Ο κ. Cohen αναφέρει "Υπήρξε μια εποχή, όχι πριν πολύ καιρό, όταν πουλούσαμε μουσική αλλά κανείς δεν ήξερε ποιοι είναι αυτοί που αγόραζαν. Έχω περάσει το μεγαλύτερο μέρος της ζωής μου, μη γνωρίζοντας ποιος είναι ο πελάτης. Δεν είναι ντροπή;". Για να καλυφθεί αυτή η ανάγκη αλλά και για να βελτιωθεί η γνώση γύρω από την αγορά της μουσικής η εταιρία 300 έχει πρόσβαση στη βάση δεδομένων του Twitter ακόμα και δεδομένων που δεν διατίθενται ελεύθερα όπως η τοποθεσία του αποστολέα. Αλλά και για το Twitter αυτή η συνεργασία είναι πολύ επωφελής γιατί ενισχύει τη θέση της με τους μουσικούς και ελκύει περισσότερους χρήστες.

6.3.8 Εξερεύνηση προκαταλήψεων συγγραφέα

Η νικήτρια ομάδα του Viafora Big Data Hackathon χρησιμοποίησε την ανάλυση συναισθήματος που παρέχει το σύστημα Semantria για να φτιάξει μια ανοικτή εφαρμογή (Chrome plugin) που θα εντόπιζε επιπλέον στοιχεία για τον συγγραφέα κάποιου άρθρου. Όταν τρέχει η εφαρμογή, ελέγχεται το όνομα του συγγραφέα και μετά εξάγει και αναλύει άλλες δουλειές του συγκεκριμένου χρησιμοποιώντας τεχνικές εξαγωγής γνώσης και ανάλυσης συναισθήματος. Για παράδειγμα αν διαβάζεις ένα άρθρο στη Huffington Post, του John Doe που κάνει κριτική για ένα προϊόν της Apple και κάνεις χρήση της εφαρμογής μπορείς να δεις ότι σε όλα τα άρθρα του John Doe τα iPhone, iPad, και Macbook χαρακτηρίζονται αρνητικά. Έτσι, θα ξέρεις ότι ο συγκεκριμένος έχεις την τάση να μισεί αυτή την μάρκα [26].

6.3.9 Έλεγχος της ζήτησης

Παράδειγμα επιτυχούς εφαρμογής ανάλυσης συναισθήματος αναφέρεται εδώ [27], και περιλαμβάνει μια ταχέως αναπτυσσόμενη ελληνική μάρκα γιαουρτιού. Η εταιρεία χρησιμοποιούσε στοιχεία που δείχνουν ότι η βανίλια ήταν η πιο δημοφιλής γεύση. Αλλά η γεύση που δημιουργούσε το μεγαλύτερο ενδιαφέρον στα κοινωνικά δίκτυα ήταν ο ανανάς. Μετά από έρευνα, προέκυψε ότι στους μεταπωλητές γρήγορα εξαντλούνταν τα αποθέματα ανανά, και έτσι οι πελάτες αγόραζαν τη δεύτερη καλύτερη επιλογή. Χωρίς αυτή την πληροφόρηση του συναισθήματος, η εταιρία θα μπορούσε να χάσει αυτό το σημαντικό πλεονέκτημα. Αντ' αυτού, ήταν σε θέση να ενισχύσει την αξία της εταιρίας, επενδύοντας και βοηθώντας έτσι τους μεταπωλητές να αποθηκεύουν κρατώντας τις σωστές ποσότητες αποθεμάτων των προϊόντων.

Με χρήση παραδοσιακών τρόπων ανάλυσης και μετρήσεων θα απαιτούσε την ανταλλαγή δεδομένων μεταξύ των μεταπωλητών και τον αρχικό προμηθευτή, το οποίο δεν συμβαίνει πάντοτε στην πράξη με αποδοτικά και γρήγορα. Η ανάλυση συναισθήματος ήταν σε θέση να εντοπίσει το πρόβλημα με πιο άμεσο τρόπο.

6.3.10 Γραφειοκρατική οργάνωση επιχειρήσεων

Η Coates Hire είναι μια από τις μεγαλύτερες εταιρίες ενοικίασης εξοπλισμού στην Αυστραλία. Δραστηριοποιείται σε όλες τις περιοχές της χώρας καθώς και στην Ινδονησία. Έχει 2600 υπαλλήλους και πάνω από 230 θυγατρικές και υπηρεσίες, περισσότερους από 20000 πελάτες σε εταιρίες όπως εξορύξεις πετρελαιοειδών και φυσικού αερίου, κατασκευές, βιομηχανική συντήρηση, κυβερνητικά συμβόλαια και εκδηλώσεις. Η επιχείρηση αυτή το 2014 χρειαζόταν να έχει ένα εύκολο τρόπο πρόσβασης στα δεδομένα της. Οι μάνατζερ της Coates Hire έβρισκαν μεγάλη δυσκολία να εντοπίσουν και να βάλουν σε σειρά προτεραιότητας τις υποχρεώσεις που θα συμβάλλουν περισσότερο στην παραγωγικότητα των υποκαταστημάτων. Μια νέα λύση ταμπλό με εργαλεία όπως IBM Analytics, Business Analytics, Business Intelligence, Performance Management, Predictive Analytics, δίνει στους διευθυντές υποκαταστημάτων άμεση εικόνα για ζητήματα που σχετίζονται με περιουσιακά στοιχεία, βοηθώντας τους να αξιολογούν τις σημαντικές προτεραιότητες και να αναλάβουν γρήγορα δράση για την επίλυσή τους. Έτσι έχουν ταχύτερη διορατικότητα που οδηγεί στην καλύτερη λήψη αποφάσεων τονίζοντας τις δυνητικές δημοσιονομικές επιπτώσεις [28].

6.3.11 Ασφαλής επικοινωνία και εξυπηρέτηση κοινού

Από το 1892 λειτουργούσε η Regina Police Service και πλέον είχε φτάσει να απασχολεί 530 εργαζομένους. Το 2014 η αστυνομία της Regina, Canada αναγκάστηκε να κλείσει την σελίδα που διατηρούσε στο Facebook μετά από μην ηλεκτρονική επίθεση, έπρεπε όμως να βρεθεί ένας τρόπος για αποτελεσματική επικοινωνία με τους πολίτες και με μεγαλύτερη ασφάλεια. Χρησιμοποιήθηκε μια εφαρμογή της IBM, Integritie Social Media SMC4® solution, built on IBM® Enterprise Content Management software για να καταγράφει αυτόματα, να ερμηνεύει, να αποκωδικοποιεί και να κάνει διαλογή των μηνυμάτων από κοινωνικά δίκτυα. Με αυτό τον τρόπο η αστυνομία της Regina έσωσε την υπόληψη της από πιθανές μελλοντικές επιθέσεις στα κοινωνικά δίκτυα. Και μπόρεσε να κάνει τη δουλεία της πιο εύκολα στο πλευρό των πολιτών. Εξοικονόμησε πολύ χρόνο από τους εργαζόμενους, και βελτίωσε τον χρόνο ανταπόκρισης με 24/7 παρακολούθηση και εργαλεία αυτόματης απάντησης. Έτσι, μείωσε δραστικά το κόστος με τις αυτόματες εφαρμογές μέσω των κοινωνικών δικτύων και της πολιτικής ελέγχου του οργανισμού [29].

7 Ανακεφαλαίωση και Συμπεράσματα

7.1 Ανακεφαλαίωση

Η παρούσα εργασία ασχολείται με την Ανάλυση Συναισθήματος, τις τεχνικές της, τους αλγόριθμους που χρησιμοποιούνται σε αυτή και τα εργαλεία της.

Αρχικά διαπιστώσαμε τη σημασία και το ενδιαφέρον της ανάλυσης συναισθήματος τη σημερινή εποχή με την όλο και αυξανόμενη χρήση των νέων τεχνολογιών στην επικοινωνία και συγκεκριμένα των μικρο-ιστολογίων. Η φύση αυτών των μέσων, προκαλεί διάφορα προβλήματα και εγείρει σκεπτικισμό σε σχέση με την σωστή εφαρμογή της ανάλυσης συναισθήματος. Έχουμε εμφάνιση πολυγλωσσικού περιεχόμενου, θόρυβο, ιδιαίτερο λεξιλόγιο κ.α.

Έπειτα, αναγνωρίσαμε και διακρίναμε το ζήτημα της ανάλυσης συναισθήματος σε κάποιες κατηγορίες. Πρώτα, σε σχέση με τις χρήσεις που μπορεί να εμφανιστεί. Μετά, σε σχέση με τον τρόπο που εφαρμόζεται˙ δηλαδή σε σχέση με το πώς προσεγγίζουμε το κείμενο (κείμενο, πρόταση, λέξη, οντότητα). Τρίτον, σε σχέση με την τεχνική που ακολουθούμε για να λάβουμε αποτελέσματα (βασισμένες σε λεξικά, επιβλεπόμενες, υβριδικές, μη επιβλεπόμενες).

Ιδιαίτερο ενδιαφέρον και σημασία έχουν οι τεχνικές και οι μέθοδοι που ακολουθούνται για την υλοποίηση της ανάλυσης συναισθήματος. Έτσι στη συνέχεια παρουσιάσαμε αναλυτικά τις τεχνικές με λεξικά, τις τεχνικές με επιβλεπόμενη μάθηση και τις υβριδικές τεχνικές (συνδυασμός λεξικολογικών, επιβλεπόμενων, μη επιβλεπόμενων τεχνικών). Αναφέραμε επίσης τον τρόπο προσέγγισης και λειτουργίας των τεχνικών αυτών, τα πλεονεκτήματα και τα μειονεκτήματά τους, καθώς και τις συγκρίσεις μεταξύ τους.

Ακολούθως, αναφέρθηκαν κάποιες χαρακτηριστικές εφαρμογές που αφορούν τόσο ελεύθερες προσεγγίσεις όσο και εμπορικές προσεγγίσεις, δείχνοντας μερικά παραδείγματα και κατασκευάζοντας πίνακες που δείχνουν με χαρακτηριστικό τρόπο τις εφαρμογές που κυριαρχούν σήμερα.

Τέλος, παρουσιάστηκαν κάποια παραδείγματα εφαρμογών (case studies). Τα παραδείγματα αυτά αντλήθηκαν από σύγχρονες πηγές και θέλουν να καταδείξουν τη χρήση, αλλά και την αναγκαιότητα της συναισθηματικής ανάλυσης σήμερα.

Μελετώντας τα παραπάνω, μπορούμε να καταλήξουμε σε κάποια χρήσιμα συμπεράσματα και σκέψεις για το μέλλον της ανάλυσης συναισθήματος.

7.2 Συμπεράσματα

Η ανάλυση συναισθήματος είναι ένα πάρα πολύ σημαντικό εργαλείο για την ανίχνευση απόψεων. Αυτό γίνεται εμφανές από την πληθώρα των χρήσεων, των τεχνικών, των μεθόδων και των εφαρμογών που υπάρχουν τη σημερινή εποχή.

Υπάρχουν προβλήματα που δεν μας επιτρέπουν να κάνουμε σωστά ανάλυση συναισθήματος. Αυτά είναι προβλήματα που απορρέουν από τη φύση των κειμένων όπως, το μήκος κειμένου, το πολυγλωσσικό περιεχόμενο, το θόρυβο κτλ. Επιπλέον, το συναίσθημα/άποψη μπορεί πολλές φορές να εκφρασθεί με πιο λεπτό/έμμεσο τρόπο χωρίς τη χρήση συναισθηματικά φορτισμένων λέξεων. Επιπρόσθετα, ο προσδιορισμός του κατόχου - εκφραστή της άποψης (opinion holder) που διατυπώνεται στο κείμενο μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Ένα άλλο ζήτημα είναι ότι το συναίσθημα και η υποκειμενικότητα ενός κειμένου εξαρτώνται από το σημασιολογικό πλαίσιο στο οποίο τοποθετούνται. Σημαντικό επίσης πρόβλημα είναι ότι μπορεί να προκύψει διαφορετικό συναίσθημα από τη διαφορετική σειρά των λέξεων και των φράσεων στο κείμενο. Τέλος, υπάρχουν δυσκολίες που προέρχονται από την γενικότερη περιοχή της Επεξεργασίας Φυσικής Γλώσσας, όπως, η αμφισημία, ο χειρισμός της άρνησης, η ειρωνεία και ο σαρκασμός.

Υπάρχουν όμως και προβλήματα που προέρχονται αποκλειστικά από την τεχνική που εφαρμόζουμε για την υλοποίηση της ανάλυσης, τα οποία διακρίνονται παρακάτω.

Προσεγγίσεις με λεξικά :

- Το πλήθος των λέξεων είναι πεπερασμένο, σε αντίθεση με τις ανάγκες για ανάλυση σε ένα δυναμικό μέσο.
- Λέξεις που μπορούν να έχουν διαφορετικό συναίσθημα σε διαφορετικά περιεχόμενα.
- Η συναισθηματική πολικότητα μπορεί να είναι άμεσα σχετιζόμενη με τη θεματολογία.

Προσεγγίσεις με επιβλεπόμενη μηχανική μάθηση :

- Το κόστος.
- Η αδυναμία να εκπαιδευτούν οι ταξινομητές σε περισσότερα από ένα πεδία.
- Ο πολύ μεγάλος χρόνος και κόπος που απαιτείται για την κατασκευή του συνόλου εκπαίδευσης του ταξινομητή.

- Η δυσκολία στην εύρεση αντιπροσωπευτικού δείγματος για την κατάλληλη εκπαίδευση του ταξινομητή.

Παρά τα παραπάνω προβλήματα έχουν βρεθεί τρόποι να ξεπεραστούν αποδοτικά με τις τεχνικές να ανταγωνίζονται μεταξύ τους, παρατηρούμε ότι κάθε μια να έχει τα δικά της πλεονεκτήματα σε σχέση με την άλλη.

Οι επιβλεπόμενες προσεγγίσεις μηχανικής μάθησης έχουν πολύ καλά αποτελέσματα καθώς επιτυγχάνουν πολύ καλά ποσοστά ακρίβειας που υπερτερούν σε πολλές περιπτώσεις από τις τεχνικές μη-επιβλεπόμενης μάθησης. Αυτές οι μέθοδοι μπορούν να πετύχουν 80%-84% ακρίβεια (accuracy). Πράγματι από τα στοιχεία που συγκεντρώσαμε επαληθεύεται ότι για επιβλεπόμενη μηχανική μάθηση η ακρίβεια είναι περίπου 79,97%.

Οι τεχνικές βασισμένες σε λεξικά (lexicon-based) έχουν επίσης κάποια πολύ σημαντικά πλεονεκτήματα, γιατί δεν χρειάζονται εκπαίδευση, και είναι πιο κατάλληλες για μεγάλο εύρος περιεχομένων, μιας και στηρίζονται σε προκατασκευασμένα λεξικά από λέξεις με συγγενική συναισθηματική κατεύθυνση.

Οι υβριδικές προσεγγίσεις εκμεταλλεύονται τα πλεονεκτήματα και προσπαθούν να αποσοβήσουν τα μειονεκτήματα των μεθόδων που εμπεριέχουν, διαπιστώσαμε οτι πετυχαίνουν πολύ καλά αποτελέσματα σε ποσοστό περίπου 77,18%.

Στη μη-επιβλεπόμενη μηχανική μάθηση, σε αντίθεση με τις λεξικολογικές μεθόδους και τις μεθόδους επιβλεπόμενης μηχανικής μάθησης, το σύστημα τροφοδοτείται μόνο από εισόδους, και καλείται να κάνει την ανάλυση συναισθήματος χωρίς ανάγκη από σύνολα ήδη χαρακτηρισμένα. Έτσι, επιτυγχάνουν και αυτές ικανοποιητικά αποτελέσματα, με καλά ποσοστά ακρίβειας όταν εφαρμόζονται σε γνωστά θεματικά πεδία, όπου το λεξιλόγιο των κειμένων τους καλύπτεται από τα λεξικά συναισθήματος. Τα αποτελέσματα για τη μη επιβλεπόμενη μάθηση είδαμε ότι μπορούν να φτάσουν την ακρίβεια σε ποσοστό περίπου 65,66%.

Το συμπέρασμα που βγαίνει μελετώντας και πολλά παραδείγματα εφαρμογών είναι ότι δεν υπάρχει μια μοναδική κατάλληλη συνταγή για ανάλυση συναισθήματος, αλλά εξαρτάται από την εκάστοτε περίπτωση και τα διάφορα δεδομένα που καλούμαστε να αναλύσουμε. Τα καλύτερα αποτελέσματα συνήθως έρχονται με τον κατάλληλο συνδυασμό των τεχνικών και των μεθόδων.

7.2.1 Ανοιχτά ζητήματα

Άνθρωποι ή μηχανές

Η απάντηση στο ερώτημα που θέτει η ανάλυση συναισθήματος είναι κάτι που δυσκολεύει τις μηχανές, αλλά εξίσου και τους ανθρώπους. Το να αποφανθεί ο άνθρωπος για το θετικό, αρνητικό ή ουδέτερο κρύβει δυσκολίες και παγίδες. Οι άνθρωποι συχνά δεν μπορούν να αποφασίσουν το συναίσθημα ενός κειμένου επειδή, όπως και οι μηχανές, δεν έχουν την αντίστοιχη γνώση για να το κάνουν. Για παράδειγμα δύο άνθρωποι με διαφορετικό γνωστικό

υπόβαθρο, διαφορετικές εμπειρίες και διαφορετικά σημεία αναφοράς μπορεί να κάνουν μια διαφορετική ανάλυση για το ίδιο θέμα

Που υστερεί η ανάλυση συναισθήματος από ανθρώπους

Όπως είπαμε οι άνθρωποι μπορεί να βρεθούν σε δύσκολη θέση για να αποφανθούν για κάποιο συναίσθημα. Αλλά το πρόβλημα μπορεί να είναι πολύ μεγαλύτερο όταν πρέπει να αποφανθούν για τεράστιο όγκο δεδομένων σε πολύ μικρό χρονικό διάστημα. Ο άνθρωπος είναι ικανός για συνεπείς και ακριβείς εκτιμήσεις κειμένων πάνω σε περιορισμένο αριθμό κατηγοριών και σε γνωστό πεδίο, αλλά όχι για πολύ μεγάλο όγκο δεδομένων.

Που μπορούν να βοηθήσουν οι μηχανές

Οι μηχανές επειδή μπορούν να μειώσουν τόσο το χρόνο όσο και να διαχειριστούν καλύτερα τον όγκο της εργασίας που έχουν να κάνουν, αποτελούν το πιο χρήσιμο εργαλείο της ανάλυσης συναισθήματος. Απαιτούν όμως ακρίβεια στα αποτελέσματα. Το μυστικό της αποτελεσματικότητας της αυτόματης ανάλυσης συναισθήματος είναι να καταλάβουμε τις επικίνδυνες περιοχές αυτής, οι οποίες είναι, η εξάρτηση περιεχομένου και η εξάρτηση χρόνου.

Εξάρτηση περιεχομένου, σημαίνει ότι ο ταξινομητής που είναι σχεδιασμένος να ταξινομεί ένα θέμα δεν μπορεί να ταξινομήσει επίσης καλά κάποιο άλλο. Για παράδειγμα όταν ταξινομεί κριτικές για ποτήρια δεν θα τα πάει καλά σε ένα πολιτικό ντιμπέιτ. Η χρονική εξάρτηση αναφέρεται στο πότε ένας ταξινομητής γίνεται μη αποδοτικός μετά από πέρασμα κάποιας χρονικής περιόδου. Το λεξιλόγιο του θέματος μπορεί να έχει αλλάξει τόσο που ο ταξινομητής δεν μπορεί πια να “καταλάβει” τα δεδομένα όπως παλιά.

Πως να αντιμετωπίσουμε τα προβλήματα των συστημάτων ανάλυσης

Για να είναι ακριβής ένας ταξινομητής συναισθήματος πρέπει να είναι συνεπής στην εξάρτηση θεματολογίας, και στην εξάρτηση χρόνου όντας συγκεκριμένος και σύγχρονος. Δεν πρέπει να στηριζόμαστε σε ένα και μόνο ταξινομητή που κάνει για όλα, αλλά πρέπει να χρησιμοποιούμε τον καταλληλότερο κάθε φορά και τον πιο ανανεωμένο.

Υπάρχουν δύο τρόποι που μπορούν τα συστήματα ανάλυσης συναισθήματος να χτιστούν και να συντηρηθούν. Ο ένας είναι βασισμένος στη γνώση/γλωσσικές πηγές και ο άλλος στη μηχανική μάθηση. Όσον αφορά στη μηχανική μάθηση που είναι η πιο συνηθισμένη μέθοδος πρέπει να είμαστε σίγουροι ότι οι ταξινομητές είναι σωστά εκπαιδευμένοι. Η δυσκολία έγκειται στο κάθε πότε πρέπει να ενημερώνουμε τους ταξινομητές. Αυτό δεν είναι μια εύκολη απόφαση, διότι χρειάζεται πολύς χρόνος και χρήματα για αυτή την συντήρηση και τις αλλαγές στον ταξινομητή, και δεν υπάρχει κανένας κανόνας που μπορεί κάποιος να ακολουθήσει για να αντιμετωπίσει αυτό το πρόβλημα.

Η αυτοματοποιημένη ανάλυση συναισθήματος είναι ένα εργαλείο και σαν όλα τα εργαλεία έχει κάποια όρια και κάποιες βέλτιστες μεθόδους και απαιτεί συντήρηση για να εξασφαλίσει το καλύτερο αποτέλεσμα. Φυσικά δεν μπορεί να αντικαταστήσει την ανθρώπινη ανάλυση συναισθήματος, και δεν γίνεται να πούμε ότι η αυτόματη ανάλυση συναισθήματος μπορεί να σταθεί από μόνη της και ότι ο ανθρώπινος παράγοντας δεν είναι απαραίτητος. Η αυτοματοποιημένη ανάλυση συναισθήματος πρέπει να θεωρείται σαν ένα πολύ καλό εργαλείο που λειτουργεί συμπληρωματικά στην ανθρώπινη ανάλυση, διότι και οι δύο έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους.

7.2.2 Ανάλυση συναισθήματος και επιπλέον εφαρμογές

Η ανάλυση συναισθήματος είναι μόνο ένα τμήμα της πληροφορίας που μπορεί να αξιοποιηθεί από τα κοινωνικά δίκτυα. Στο ίντερνετ, και ειδικότερα στα μέσα κοινωνικής δικτύωσης, κρύβεται ένας πολύ μεγάλος πλούτος πληροφορίας που ο άνθρωπος καλείται να εξάγει. Για το λόγο αυτό παρατηρούμε ότι οι μεγάλες εταιρείες του χώρου που απευθύνονται σε πραγματικές εφαρμογές και επιχειρήσεις έχουν στο οπλοστάσιό τους πολλές ακόμα λειτουργίες που δρουν μαζί ή συμπληρωματικά με την ανάλυση συναισθήματος. Τέτοιες είναι για παράδειγμα οι Προγνωστικές Αναλύσεις (Predictive analytics), Εξόρυξη Οντότητας (Entity Extraction), Εξόρυξη λέξης-κλειδιού (Keyword Extraction), Χαρακτηρισμός έννοιας (Concept Tagging), Εξόρυξη σχέσεων (Relation Extraction), Κατηγοριοποίηση ταξονομίας (Taxonomy Classification), Εύρεση συγγραφέα (Author Extraction), Εντοπισμός γλώσσας (Language Detection), Εξαγωγή κειμένου (Text Extraction), Εύρεση εισόδου (Feed Detection), Υποστήριξη διασυνδεδεμένων δεδομένων (Linked Data Support), Παρακολούθηση κοινωνικών δικτύων και υποστήριξη δυνατοτήτων για επικοινωνία μεταξύ επαγγελματιών (Social media monitoring and engagement capabilities to communication professionals), Διαχείριση της μάρκας και υποστήριξης πελατών (Brand managers and customer support groups), Μάρκετινγκ ηλεκτρονικού ταχυδρομείου (Email Marketing), Μάρκετινγκ κινητών (Mobile Marketing), Μάρκετινγκ κοινωνικών δικτύων (Social Media Marketing), Διαφημίσεις (Ads), Προσωποποίηση του διαδικτύου (Web Personalization), Διαχείριση ταξιδίων (Journey Management), Ευφυΐα πρόβλεψης (Predictive Intelligence), Περιεχόμενο και μηνύματα (Content and Messaging) και πολλά άλλα. Σίγουρα όμως, μπορούμε να ισχυριστούμε ότι η ανάλυση συναισθήματος σημαίνει καλύτερο και πιο συντομευμένο μάρκετινγκ, πιο γρήγορος εντοπισμός ευκαιριών και απειλών, διαχείριση φήμης και ο τελικός και γενικότερος στόχος το κέρδος.

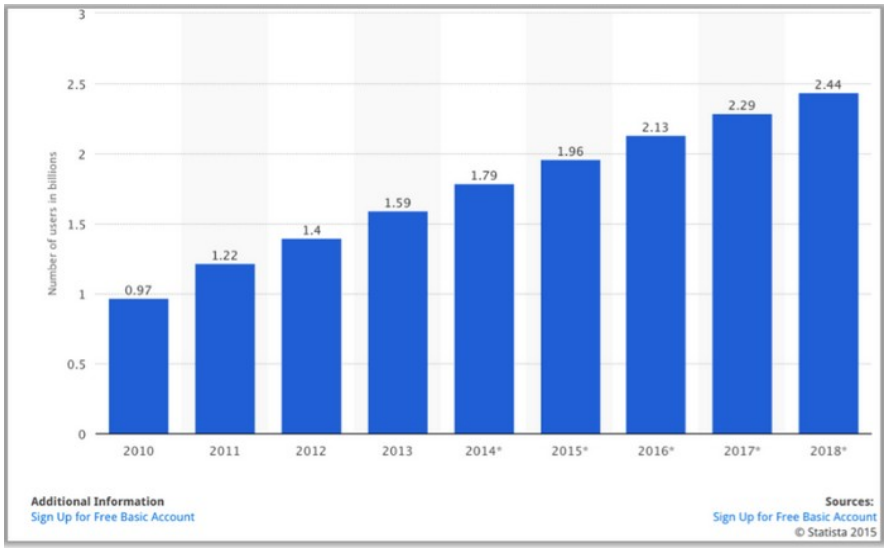
7.2.3 Το μέλλον της ανάλυσης συναισθήματος

Η σημερινή εποχή είναι η πιο κατάλληλη στιγμή για να εξεταστούν οι τεχνολογίες ανάλυσης συναισθήματος. Σύμφωνα με αυτή την παρουσίαση [30], από την NetBase, CMO, Lisa Joy

Rosner, ο μέσος καταναλωτής αναφέρει συγκεκριμένες μάρκες πάνω από 90 φορές την εβδομάδα σε συνομιλίες με τους φίλους, την οικογένεια και τους συναδέλφους. Επιπλέον, το 53% των ανθρώπων στο Twitter προτείνουν εταιρίες ή/και προϊόντα στα tweets τους, με το 48% από αυτούς να μιλούν σχετικά με την πρόθεσή τους να αγοράσουν το προϊόν.

Αυτό σημαίνει ότι το Twitter και τα άλλα μέσα κοινωνικής δικτύωσης είναι ένα τέλειο βοήθημα στους παραδοσιακούς τρόπους για μάρκετινγκ και μάνατζμεντ. Επιπλέον η χρήση του έχει εξαπλωθεί περισσότερο, με τους χρήστες του διαδικτύου άνω των 50 ετών να έχουν σχεδόν διπλασιαστεί στο 42% το προηγούμενο έτος. Οι επιχειρήσεις πλέον μπορούν να πάρουν αμερόληπτες, πιο ειλικρινείς σκέψεις και απόψεις, και οι καταναλωτές να πάνε σε αυτές με πιο φυσικό τρόπο και δωρεάν.

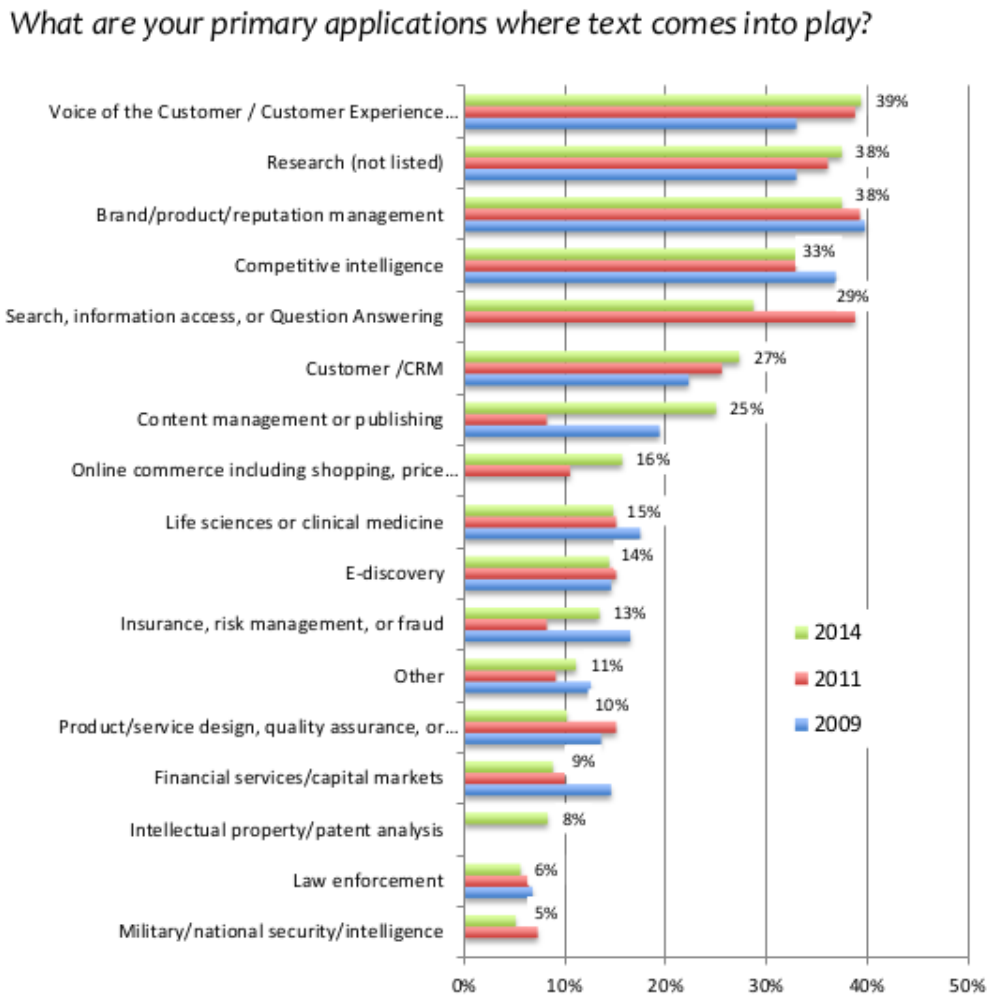
Στη σημερινή εποχή παρατηρείται όλο και περισσότερη χρήση των κοινωνικών δικτύων. Στη παρακάτω μελέτη [31] βλέπουμε πως θα διαμορφωθεί ο αριθμός των χρηστών (σε δισεκατομμύρια) των κοινωνικών δικτύων σε όλο τον κόσμο, από το 2010 έως 2018.



Σχήμα 7.1: Στατιστική για αριθμό χρηστών των κοινωνικών δικτύων 2010-2018

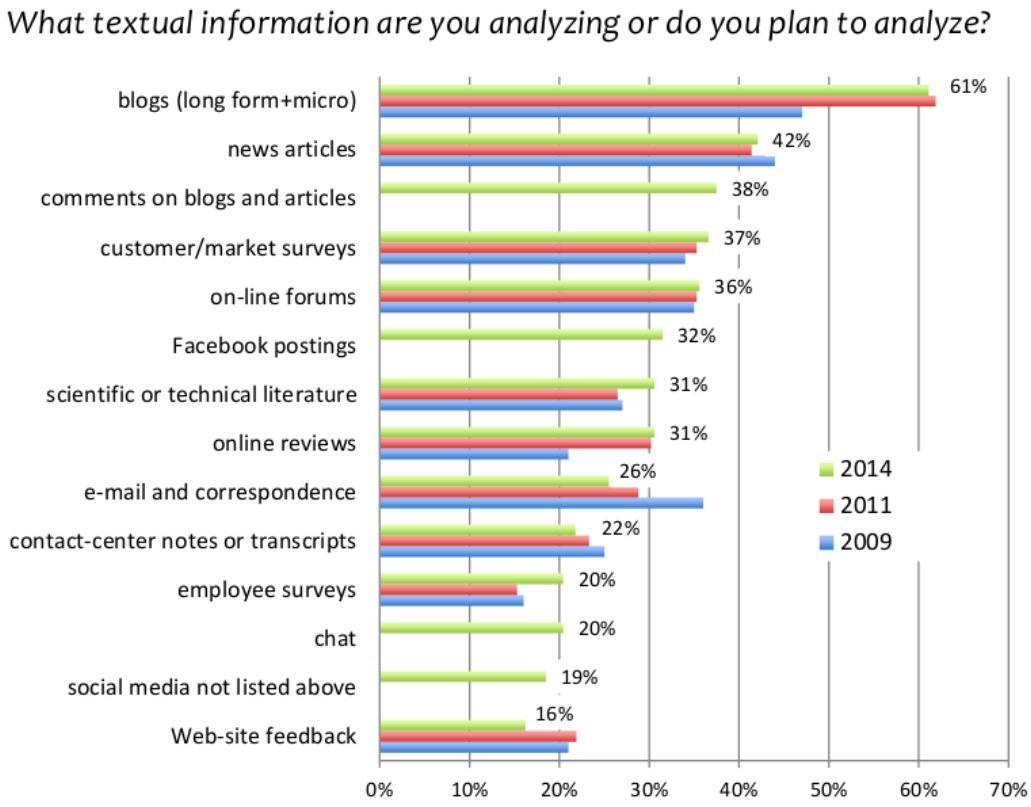
Ο Seth Grimes ιδρυτής της εταιρίας Alta Plana [14] και διοργανωτής του συνεδρίου (Sentiment Analysis Symposium) [15] κάθε χρόνο, αναφέρει κάποιους πολύ χαρακτηριστικούς πίνακες που μας δίνουν μια καλή εικόνα για τη ανάλυση συναισθήματος και το μέλλον αυτής [16].

Στο επόμενο σχήμα βλέπουμε να περιγράφεται με παραστατικό τρόπο, ποιες είναι οι εφαρμογές και οι τομείς που εφαρμόστηκαν τεχνικές ανάλυσης συναισθήματος κατά το πέρασμα των ετών από το 2009, 2011, 2014. Παρατηρούμε κάποια πεδία να έχουν ανοδική πορεία σε σχέση με το χρόνο, όπως οι απόψεις των πελατών (Voice of the customer), ενώ άλλα να παραμένουν στάσιμα ή και να χάνουν την ανοδική τους κατεύθυνση, όπως η έρευνα και τα ερωτηματολόγια. Επιπλέον, βλέπουμε σιγά σιγά να εισάγονται στο πεδίο της ανάλυσης συναισθήματος νέες κατηγορίες θεμάτων, όπως ζητήματα εθνικής ασφάλειας.



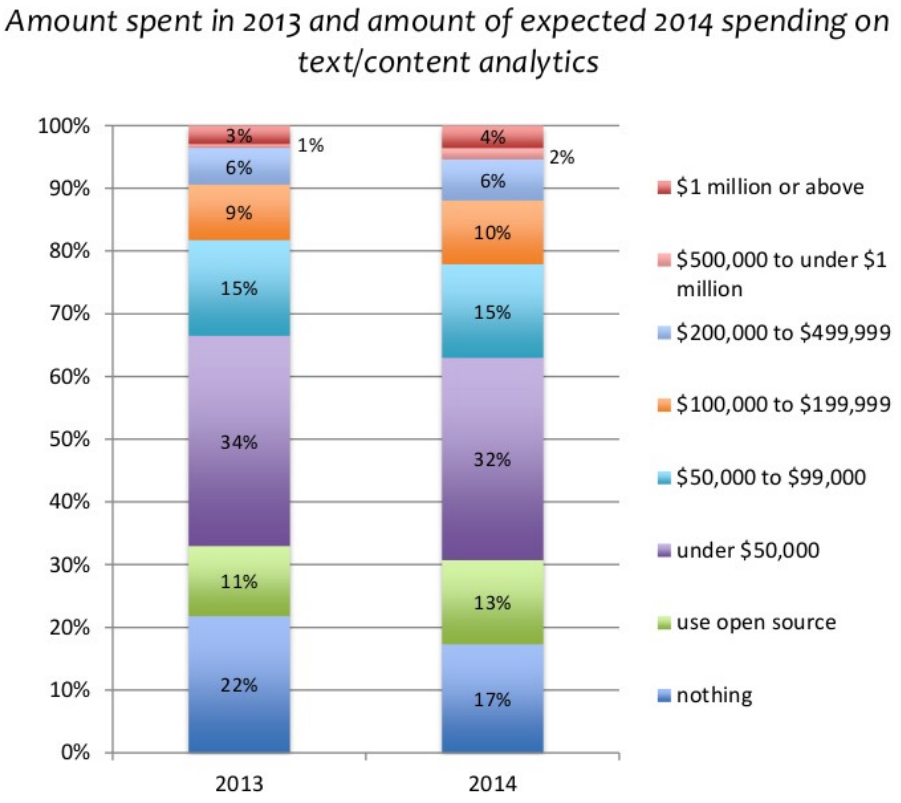
Σχήμα 7.2: Τομείς εφαρμογής ανάλυσης συναισθήματος

Στο επόμενο σχήμα βλέπουμε το είδος των μέσων (κοινωνικών δικτύων, κ.α.) που θέλουμε να αναλύσουμε συναισθηματικά και πως αυτό διαμορφώνεται με το πέρασμα των ετών. Βλέπουμε πως το σημαντικό μέσο που μας απασχολεί είναι τα blogs και τα άρθρα ειδήσεων. Ωστόσο, τα μέσα με μεγαλύτερη δυναμική είναι το Twitter, και το Facebook.



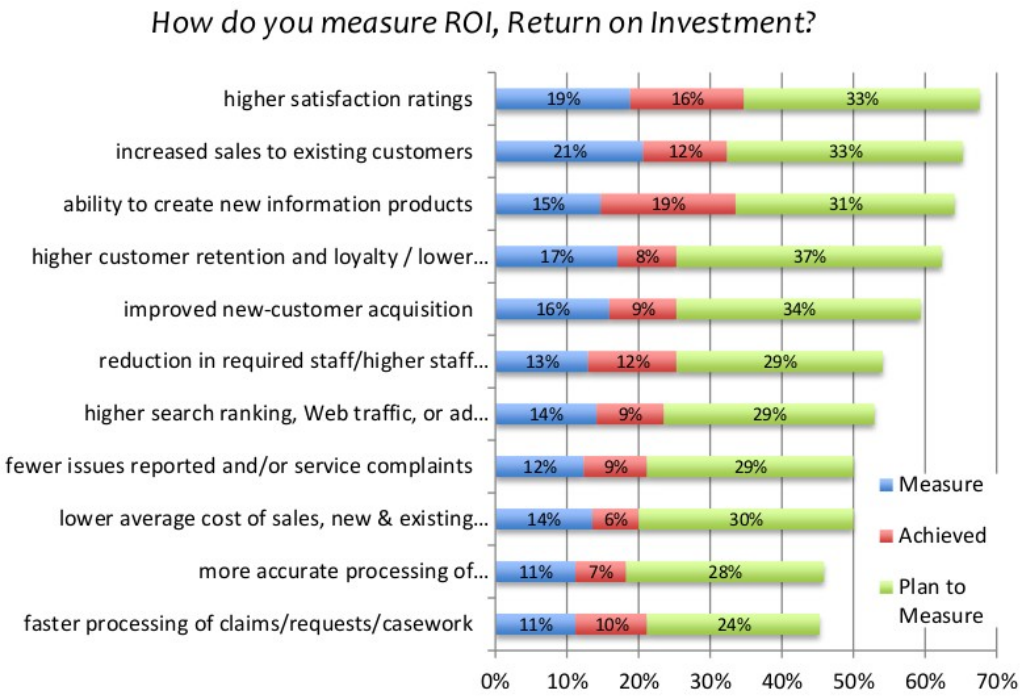
Σχήμα 7.3: Ποια κοινωνικά δίκτυα αναλύουμε

Σημαντική ένδειξη για τη σημασία και τη δυναμική της ανάλυσης συναισθήματος έχει να δούμε την πορεία των επενδύσεων που γίνονται σε αυτό τον τομέα. Το παρακάτω σχήμα μας δείχνει τα ποσά που είναι διατεθειμένες να επενδύσουν οι επιχειρήσεις στην ανάλυση συναισθήματος. Τα μεγέθη των επενδύσεων φτάνουν το ύψος των εκατομμυρίων δολαρίων.



Σχήμα 7.4: Χρηματικά ποσά που ξοδεύονται στην ανάλυση συναισθήματος

Μετρικές έχουν εμφανιστεί και έρευνες γίνονται για να δειχθεί κατά πόσο η ανάλυση συναισθήματος είναι άξια επενδύσεων για τις επιχειρήσεις. Τα ποσοστά ικανοποίησης είναι ανοδικά σε όλα τα πεδία εφαρμογής κατά τη διάρκεια των ετών, πράγμα που φανερώνει τη μεγάλη χρησιμότητα και τα οφέλη της χρήσης της ανάλυσης συναισθήματος.



Σχήμα 7.5: Μέτρο ικανοποίησης από την ανάλυση συναισθήματος

7.2.4 Επόμενα βήματα

Σίγουρα υπάρχουν περισσότερες κατηγορίες που μπορούμε να εντάξουμε την πολικότητα όπως, θυμωμένος, χαρούμενος, λυπημένος, απηυδισμένος, ικανοποιημένος κτλ. Αυτές μπορούν να προσφέρουν περισσότερη πληροφορία από το απλό θετικό, αρνητικό, ουδέτερο, σύστημα βαθμονόμησης. Επιπλέον, οι προχωρημένες λύσεις στον τομέα της αυτόματης ανάλυσης συναισθήματος προχωρούν ένα βήμα παραπέρα προσπαθώντας να εντοπίσουν συναίσθημα, όχι μόνο στο κείμενο, αλλά και σε φωτογραφίες και βίντεο. Στο μεθοδολογικό κομμάτι, μερικά συστήματα συνδέουν το συναίσθημα με εγγραφές, όπως τις πωλήσεις, τις έρευνες, τις πληρωμές κτλ, συμπεριλαμβανομένου του συσχετισμού θέσης που μας οδηγεί προς έναν κόσμο ολοκληρωμένων αναλύσεων.

8 Βιβλιογραφία

[AML+14] Aston, N., Munson, T., Liddle, J., Hartshaw, G., Livingston, D., & Hu, W. (2014). Sentiment analysis on the social networks using stream algorithms. Journal of Data Analysis and Information Processing, 2014.

[ATB+12] Asiaee T, A., Tepper, M., Banerjee, A., Sapiro, G. If you are happy and you know it... tweet. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1602{1606. ACM. 2012

[AXV+11] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media (pp. 30-38). Association for Computational Linguistics.

[BAM+12] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V.. Mining sentiments from tweets. Proceedings of the WASSA 12. 2012

[BBP+12] Matteo Baldoni, Cristina Baroglio, Viviana Patti and Paolo Rena. From Tags to Emotions: Ontology-driven Sentiment Analysis in the Social Semantic Web. pp. 41-54. 2012

[BES10] Baccianella S., Essuli, A. and Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of LREC 10. 2010

[BF10.] Bifet, Albert, and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data."Discovery Science. Springer Berlin Heidelberg, 2010.

[BF10] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44. Association for Computational Linguistics, 2010.

[BL99] Bradley, M.M., & Lang, P.J. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida. 1999

[BRS10] Ramnath Balasubramanyan , Bryan R. Routledge , Noah A. Smith, From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series (2010)

[BS10] Adam Bermingham and Alan F Smeaton. Classifying sentiment in

microblogs: is brevity an advantage? In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1833–1836. ACM. 2010

[CE13] Chalothorn, T., Ellman, J. Tjp: Using twitter to analyze the polarity of contexts.In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA. 2013

[CIS+06] Chen, C., Ibekwe-SanJuan, F., SanJuan, E., & Weaver, C. (2006, October). Visual analysis of conflicting opinions. In Visual Analytics Science And Technology, 2006 IEEE Symposium On (pp. 59-66). IEEE.

[CRV+13] Martinez-Camara, E., Montejo-Raez, A., Martin-Valdivia, M., Urena-Lopez, L Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. 2013

[DH13] Deitrick, W., Hu, W. Mutually enhancing community detection and sentiment analysis on twitter networks. Journal of Data Analysis and Information Processing 1. 2013

[DTR10] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 241–249. Association for Computational Linguistics. 2010.

[ES06] Esuli, A. and Sebastiani, F. SentiWordNet: A high-coverage lexical resource for opinion mining. Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR). 2006

[Fel98] Christiane Fellbaum. WordNet. Wiley Online Library. 1998.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, pages 1–12. 2009

[GL08] Gindl, Stefan, and Johannes Liegl. "Evaluation of different sentiment detection methods for polarity classification on web-based reviews."Proceedings of the 18th European conference on artificial intelligence. 2008.

[HBB13] Hussam Hamdan, Frederic Béchet, Patrice Bellot. Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. Joint Conference on Lexical and Computational Semantics, 2013

[HL04] Hu, M. and Liu, B. Mining and Summarizing Customer Reviews. ACM SIGKDD-2. 2004

[HM97] Hatzivassiloglou and McKeown: Predicting the Semantic Orientation of Adjectives : 1987 Wall Street Journal corpus. 1997

[HTG+13] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international

conference on World Wide Web, pages 607–618. International World Wide Web Conferences Steering Committee, 2013

[HTT+13] Hu, X., Tang, L.,Tang, J., Liu, H. Exploiting social relations for sentiment analysis in microblogging. In: WSDM, pp 537-546. 2013

[HW00] Hatzivassiloglou and Wiebe: Effects of Adjective Orientation and Gradability on Sentence Subjectivity : 1987 Wall Street Journal corpus. 2000

[JYZ+11] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 151–160. Association for Computational Linguistics, 2011

[KH06] Soo-Min Kim and Eduard Hovy, Identifying and analyzing judgment opinions, Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, June, 2006

[KS12] Akshi Kumar and Teeja Mary Sebastian. Sentiment analysis on twitter. IJCSI International Journal of Computer Science Issues, 9(3):372–378. 2012

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! ICWSM, 11:538–541, 2011

[Lev13] Clement Levallois, Umigon: Sentiment Analysis for Tweets Based on Lexicons and Heuristics, Proceedings of the International Workshop on Semantic Evaluation, SemEval ’13, June 2013, Atlanta, Georgia, May 1, 2013

[Liu12] Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[LL10] Li, G., Liu, F. .A Clustering-based Approach on Sentiment Analysis. 978-1-4244-6793-8/10 ©2010 IEEE. 2010

[LLG12] Liu, K.L., Li, W.J., Guo, M. Emoticon smoothed language models for twitter sentiment analysis. In: AAAI. 2012

[MGL09] Prem Melville, Wojciech Gryc, Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 1275-1284. 2009

[MKZ13] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. CoRR, abs/1308.6242, 2013.

[MMP13] Bravo-Marquez, F., Mendoza, M., Poblete, B.: Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the Second International Workshop on Issues of Sentiment

Discovery and Opinion Mining. ACM 2013

[OBR+10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. 2010.

[OFM13] Reynier Ortega, Adrian Fonseca, and Andrés Montoyo. Ssa-uo: unsupervised twitter sentiment analysis. In Second Joint Conference on Lexical and Computational Semantics (* SEM), volume 2, pages 501–507, 2013.

[PB91] Douglas B. Paul, Janet M. Baker. The design for the wall street journal-based CSR corpus. Published in: Proceeding HLT '91 Proceedings of the workshop on Speech and Natural Language, Pages 357-362, Association for Computational Linguistics Stroudsburg, PA, USA©1992

[PBF07] Pennebaker, J.W., Booth, R.J., & Francis, M.E. Linguistic Inquiry and WordCount:LIWC2007. Austin, TX. 2007

[PI09] PANDEY, Vipul; IYER, C. Sentiment analysis of microblogs. CS 229: Machine learning final projects, 2009.

[PL08] Bo Pang, Lillian Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval, Vol. 2, No 1-2, 2008

[PLV02] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In EMNLP, pages 79–86. 2002

[POL10] Petrovic S., Osborne M., Lavrenko V. Edinburgh Twitter corpus.In Workshop on Social Media, NAACL. 2010

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREC, 2010.

[Rem13] D-Remus, R Asvunioeipzig.:Sentiment analysis in twitter using data-driven machine learning techniques. 2013

[Seb02] Fabrizio Sebastiani, Machine learning in automated text categorization, ACM Computin Surveys (CSUR), ACM Press, 2002

[SFH+13] Saif, H., Fernandez, M., He, Y. and Alani, H. Evaluation Datasets for Twitter Sentiment Analysis. In Proceedings of ESSEM. 2013

[SFH+14] Hassan Saif, Miriam Fernandez, Yulan He and Harith Alani. SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter. In The Semantic Web: Trends and Challenges, pages 83–98. Springer, 2014.

[SHA11] Saif, H., He, Y., Alani, H.: Semantic Smoothing for Twitter Sentiment Analysis. In: Proceeding of the 10th International Semantic Web Conference (ISWC). 2011

[SHA12.] Hassan Saif, Yulan He and Harith Alani. Semantic Sentiment Analysis of Twitter, ISWC'12 Proceedings of the 11th international conference on The Semantic Web - Volume Part I, Pages 508-524 Springer – VerlagBerlin , Heidelberg. 2012

[SHA12] Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment

analysis. In: Proceedings, 2nd Workshop on Making Sense of Microposts (#MSM2012) in conjunction with WWW 2012. Layon, France. 2012

[SHF+14] Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani. Adapting Sentiment Lexicons using Contextual Semantics for Sentiment Analysis of Twitter. In:Workshop 5: SemanticSentimentAnalysis2014: Semantic Web and Sentiment Analysis, 25-19 May 2014, Crete, Greece, Springer International Publishing, pp. 54–63.

[SKC09] D. Shamma, L. Kennedy, and E. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In Proceedings of WSM, 2009.

[SSU+11] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland (2011), pp. 53-63

[TAV06] M. Taboada, C. Anthony, K. Voll, Methods for Creating Semantic Orientation Databases, Proceeding of LREC-06, the 5th International Conference on Language Resources and Evaluation, pages 427–432, 2006

[TBP+10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12):2544–2558, 2010.

[TBP12] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1):163–173, 2012.

[TP11] TROMP, Erik; PECHENIZKIY, Mykola. Senticorr: Multilingual sentiment analysis of personal correspondence. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, 2011. p. 1247-1250.

[Tur02] Peter D. Turney., Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.

[WWC05] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(2- 3):165–210, 2005.

[WZJ+13] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, Zhana Bao. A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. Trends and Applications in Knowledge Discovery and Data Mining Lecture Notes in Computer Science Volume 7867, 2013, pp 201-203

[YH03] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion

questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[ZGD+11] Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011, 89, 2011.

Ηλεκτρονικές Πηγές

[1] https://about.twitter.com/company

[2] http://www.niemanlab.org/2013/01/feelings-nothing-more-than-feelings-the-measured-rise-of-sentiment-analysis-in-journalism/

[3] http://www.bloomberg.com/bw/stories/2011-03-01/sentiment-analysis-gives-companies-insight-into-consumer-opinionbusinessweek-business-news-stock-market-and-financial-advice

[4] https://www-03.ibm.com/press/us/en/pressrelease/40120.wss

[5] http://www.theysay.io/stock-market-sentiment/

[6] https://en.wikipedia.org/wiki/Barack_Obama_on_social_media

[7] http://www.businessinsider.com/twitter-facebook-monitoring-2012-11

[8] http://www.nytimes.com/2006/10/04/us/04monitor.html

[9] http://wordnetweb.princeton.edu/perl/webwn

[10] http://sentiwordnet.isti.cnr.it/search.php?q=good

[11] http://www.liwc.net/tryonline.php

[12] http://sentiment.christopherpotts.net/lexicons.html#resources

[13] http://www.informationweek.com/software/information-management/expert-analysis-is-sentiment-analysis-an-80--solution/d/d-id/1087919?

[14] http://altaplana.com/

[15] http://sentimentsymposium.com/

[16] http://www.slideshare.net/SethGrimes/text-analytics-2014-user-perspectives-on-solutions-and-providers

[17] https://www.brandwatch.com/case-study-espn/

[18] http://www-03.ibm.com/software/businesscasestudies/fr/fr/corp?synkey=Z349591K13544V96

[19] http://www.smartdatacollective.com/raminuseir/191466/strange-uses-sentiment-analysis

[20] http://pivotcon.com/3-case-studies-on-the-power-of-social-metrics/

[21] http://dealbook.nytimes.com/2013/04/04/twitter-arrives-on-wall-street-via-bloomberg/?_r=0

[22] http://thomsonreuters.com/en.html

[23] http://techcrunch.com/2014/02/03/twitter-raises-its-enterprise-cred-with-thomson-reuters-sentiment-analysis-deal/

[24] http://mobile.nytimes.com/2014/02/03/business/media/twitter-and-300-team

-up-to-find-musical-talent.html?_r=0
[25] https://twitter.com/300
[26] http://www.smartdatacollective.com/raminuseir/191466/strange-uses-sentiment-analysis
[27] http://timoelliott.com/blog/2012/11/scrooge-didnt-believe-in-sentiment-analysis-either.html
[28] http://www-03.ibm.com/software/businesscasestudies/ca/en/corp?synkey=B059516T66601C14
[29] http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=S050167W39816F15
[30] https://www.slideshare.net/login?from=download&from_source=%2Ftimoelliott%2Fsavedfiles%3Fs_title%3Dlisa-joy-rosner-the-new-market-research%26user_login%3DIABC2010
[31] http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
[32] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[33] http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
[34] http://help.sentiment140.com/
[35] https://bitbucket.org/speriosu/updown
[36] http://sentistrength.wlv.ac.uk/documentation/
[37] http://www.sananalytics.com/lab
[38] http://crl.ucsd.edu/corpora/

9 *Παράρτημα: Σύνολα δεδομένων*

Σύνολα δεδομένων (datasets)

Όπως αναφέραμε και παραπάνω στη παράγραφο για τα σύνολα εκπαίδευσης (training set) αυτά προκύπτουν από το αρχικό σύνολο δεδομένων (dataset) μαζί με το σύνολο ελέγχου (test set). Το σύνολο δεδομένων πρέπει να αποτελείται από κείμενα που ανήκουν στον ίδιο τύπο και στο ίδιο θεματικό περιεχόμενο με αυτά που θέλουμε να ταξινομεί το σύστημα μετά την εκπαίδευση του (π.χ. reviews, tweets, facebook posts κτλ.).

Η χρήση αυτών των συνόλων είναι πάρα πολύ σημαντική για κάποιον που θέλει να εφαρμόσει Μηχανική Μάθηση διότι έτσι περιορίζεται κατά πολύ ο χρόνος που απαιτείται για τη δημιουργία ενός νέου συνόλου δεδομένων για την ανάλυση συναισθήματος και μειώνεται ο κίνδυνος για αστοχία του συνόλου, καθώς τα σύνολα που υπάρχουν είναι πολύ προσεκτικά κατασκευασμένα και μπορούμε να τα βρούμε για κάθε τομέα και θεματολογία σε αφθονία. Επιπλέον, εξασφαλίζεται η ποιότητα του συνόλου εκπαίδευσης καθώς με το πέρασμα του χρόνου βελτιώνονται συνεχώς και διορθώνονται τυχών αστοχίες. Οι [SFH+13], έκαναν πολύ καλή δουλεία στη αξιολόγηση συνόλων δεδομένων (datasets) για την ανάλυση συναισθήματος στο Twitter. Παρακάτω θα αναφέρουμε μερικά από τα κυριότερα σύνολα δεδομένων που συναντήσαμε στη βιβλιογραφία:

- **Pang & Lee dataset:** είναι μια συλλογή από 1.000 αρνητικές και 1.000 θετικές κριτικές ταινιών, η οποία δημιουργήθηκε από τους Pang & Lee [32].

- **MPQA opinion corpus:** Συλλογή από άρθρα ειδήσεων που προέρχονται από ποικιλία ειδησεογραφικών πηγών και έχουν σχολιαστεί από ως προς την άποψη, τις πεποιθήσεις, το συναίσθημα και τις εικασίες που εκφράζουν [33].

- **Wall Street Journal corpus:** [PB91] Αποτελείται από περίπου 25 εκατομμύρια λέξεις που έχουν εξαχθεί από κείμενα της WSJ: Wall Street Journal. Έχει χρησιμοποιηθεί η υποκατηγορία του 1987 Wall Street Journal corpus από τους [HM97], [HW00].

Μερικά από τα πιο σημαντικά σύνολα δεδομένων που αφορούν το Twitter είναι τα παρακάτω:

1. **Stanford Twitter sentiment corpus**

Το Stanford Twitter sentiment corpus [34], το εισήγαγε οι Go, Bhayani και Huang [GBH09]. Αποτελείται από δύο διαφορετικά σύνολα, Σύνολο εκπαίδευσης (training set) και σύνολο τεστ (test set). Το (training set) περιέχει 1.6 εκατομμύρια tweets που είναι αυτόματα χαρακτηρισμένα σαν θετικά ή αρνητικά βασισμένα στα emotions. Για παράδειγμα, ένα tweet χαρακτηρίζεται ως θετικό αν περιέχει :), :-), :), :D, ή =) και ως αρνητικό αν περιέχει :(, :-(, ή : (. Αν και οι απόδοση συναισθήματος χρησιμοποιώντας τα emoticons είναι γρήγορη, η ακρίβεια (accuracy) είναι αμφίβολη επειδή τα emoticons μπορεί να μην αναπαριστούν ορθά το συναίσθημα. Το σύνολο εκπαίδευσης (STS-Test), έχει χαρακτηριστεί χειροκίνητα και περιέχει 177 αρνητικά, 182 θετικά και 139 ουδέτερα tweets. Τα tweets συλλέχθηκαν με χρήση API αναζήτησης του Twitter με όρους αναζήτησης ονόματα προϊόντων, ανθρώπων και εταιρειών. Παρά το μικρό του μέγεθος, το σύνολο STS- Test έχει χρησιμοποιηθεί σε πολλές εφαρμογές για αξιολόγηση ταξινόμησης σε δύο κλάσεις ή αξιολόγηση ταξινόμησης ως προς την υποκειμενικότητα. Για παράδειγμα [GBH09], [SHA11], [SHA12], [SSU+11], και [BAM+12] το χρησιμοποίησαν για να αξιολογήσουν τα μοντέλα για την ταξινόμηση της πολικότητας (positive vs. negative). Επιπλέον ο, [MMP13] χρησιμοποίησε το dataset για αξιολόγηση υποκειμενικότητας (neutral vs polar). Επιπλέον το έχουμε συναντήσει να το χρησιμοποιούν οι [KWM11], [SHA12.], [SFH+14], [HTG+13].

2. **Endinburgh Twitter Corpus**

Το Endinburgh Twitter Corpus δημιουργήθηκε από τους Petrovic, Osborne, Lavrenko, [POL10]. Το κείμενό τους περιέχει 97 εκατομμύρια tweets, και χρειάζεται τουλάχιστον 14 GB χωρητικότητα στο δίσκο όταν δεν είναι συμπιεσμένο. Το συναντήσαμε να το χρησιμοποιούν οι [KWM11].

3. **Health Care Reform (HCR)**

Το Health Care Reform (HCR) dataset δημιουργήθηκε συλλέγοντας tweets που περιέχουν το hashtag "#hcr" (health care reform) το March 2010 από τους [SSU+11]. Ένα υποσύνολο αυτού του σώματος κειμένου(corpus) χαρακτηρίστηκε χειροκίνητα με 5 κατηγορίες (positive, negative, neutral, irrelevant, unsure(other)) και χωρίστηκε σε training (839 tweets), development (838 tweets) and test (839 tweets) sets. Οι συγγραφείς επιπλέον έβαλαν ετικέτες σε 8 διαφορετικούς στόχους από τα τρία σύνολα (Health Care Reform, Obama, Democrats, Republicans, Tea Party, Conservatives, Liberals, and Stupak). Ωστόσο, τόσο στα tweets όσο και στους στόχους αποδόθηκαν οι ίδιοι χαρακτηρισμοί. Το σύνολο HCR έχει χρησιμοποιηθεί για την αξιόλογη ταξινόμησης δύο κλάσεων [SSU+11], [SHA12.], [SHA12.], [SFH+14], καθώς και ταξινόμησης υποκειμενικότητας αφού αναγνωρίζει και τα ουδέτερα tweets.

4. **Obama - McCain Debate (OMD)**

Το Obama - McCain Debate (OMD) dataset δημιουργήθηκε από 3.238 tweets που εξάχθηκαν κατά το πρώτο τηλεοπτικό προεδρικό debate από τους Obama – McCain στις Ηνωμένες Πολιτείες το Σεπτέμβριο του 2008 από τους [SKC09]. Με χρήση του Amazon Mechanical

Turk αποδόθηκαν σ' αυτά τα tweets ετικέτες συναισθήματος (positive, negative, mixed, other) μετά από αξιολογήσεις τουλάχιστον τριών κριτών για το καθένα. Η συμφωνία των κριτών έχει υπολογιστεί σε 0,655, τιμή αρκετά ικανοποιητική. Το σύνολο OMD έχει χρησιμοποιηθεί σε εφαρμογές [SFH+14] τόσο Επιβλεπόμενης [HTT+13], [SSU+11], [SHA12.], όσο και Μη Επιβλεπόμενης Μάθησης [HTG+13], για την αξιολόγηση συναισθηματικής ταξινόμησης σε tweets [35].

5. Sentiment Strength Twitter Dataset (SS-Tweet)

Το Sentiment Strength Twitter Dataset (SS-Tweet) dataset αποτελείται από 4.242 tweets χειροκίνητα επισημασμένα με τιμές από -5 (extremely negative) μέχρι -1 (not negative) και 1(not positive) μέχρι 5 (extremely positive). Δημιουργήθηκε από τους Thelwall, Buckley, Paltoglou, [TBP12], με στόχο την αξιολόγηση του Sentistrength, μιας βασισμένης σε λεξικό εφαρμογής εκτίμησης έντασης συναισθήματος [36].

6. Sanders Twitter Dataset

Το Sanders Twitter Dataset dataset αποτελείται από 5.512 tweets για τα εξής τέσσερα θέματα Apple, Google, Microsoft, Twitter. Κάθε tweet έχει χαρακτηριστεί χειρωνακτικά από κάποιον κριτή ως positive, negative, neutral ή irrelevant σε σχέση με το κάθε θέμα. Από το χαρακτηρισμό προέκυψαν 654 negative, 2.503 neutral, 570 positive και 1.786 irrelevant tweets. Το σύνολο έχει χρησιμοποιηθεί για ταξινόμηση συναισθήματος και υποκειμενικότητας [MMP13], [LLG12], [DH13]. [37].

7. The Dialogue Earth Twitter Corpus (WA, WB, GASP)

Το The Dialogue Earth Twitter Corpus dataset αποτελείται από τρία υποσύνολα από tweets. Τα δύο πρώτα (WA, WB) περιέχουν 4.490 και 8.850 tweets σχετικά με τον καιρό, ενώ το τρίτο υποσύνολο (GASP) περιέχει 12.770 tweets σχετικά με τις τιμές του φυσικού αερίου. Αυτά τα σύνολα δημιουργήθηκαν ως μέρος του Dialogue Earth Project και επισημάνθηκαν χειρονακτικά από αρκετούς με τους χαρακτηρισμούς positive, negative, neutral, not related, can‟t tell (other). Τα σύνολα WAB και GASP έχουν χρησιμοποιηθεί στην αξιολόγηση απόδοσης ταξινομητών Μηχανική Μάθησης (π.χ. Naive Bayes, SVM, KNN) σε συναισθηματική ταξινόμηση tweets, [ATB+12].

8. SemEval Dataset (SemEval)

Το SemEval Dataset (SemEval) dataset δημιουργήθηκε για την ανάλυση συναισθήματος στο twitter στα πλαίσια του Semantic Evauation of System challenge (SemEval - 2013, Task 2). Το SemEval-2013 αποτελείται από 20000 tweets, τα οποία μοιράζονται σε σύνολα εκπαίδευσης, ανάπτυξης και ελέγχου. Όλα τα tweets έχουν χαρακτηριστεί χειροκίνητα από 5 κριτές του Amazon Mechanical Turk ως negative, positive θαη neutral. Οι κριτές, επίσης

αξιολογούν τα tweets ως υποκειμενικά ή αντικειμενικά. Το SemEval 2013 είδαμε να χρησιμοποιείτε από τους [OFM13]. Επιπλέον στο SemEval 2013 – Task2 το σύνολο δεδομένων χρησιμοποιήθηκε για αξιολόγηση των συστημάτων στον εντοπισμό υποκειμενικότητας σε επίπεδο έκφρασης, [MKZ13], [CE13] καθώς και σε επίπεδο tweet, [CRV+13], [Rem13]. Η ιστορία για το Semeval ξεκινάει από το Semeval-1 1998, Semeval-2 2001, Semeval-3 2004, SemEval-2007, SemEval-2010, SemEval-2012, SemEval-2013, SemEval-2014, SemEval-2015, SemEval-2016.

Από την Center for Research in Language [38], μπορούμε να δούμε μια ενδιαφέρουσα λίστα με κάποια από τα πιο γνωστά σώματα κειμένου (Corpora):

CHILDES	Child Language Description Exchange. Child language productions from a variety of researchers. http://childes.psy.cmu.edu/
WordNet 1.6	Lexical database (See http://www.cogsci.princeton.edu/~wn/)
Penn Treebank	Penn's Linguistic Data Consortium (LDC) collection, including Brown (Kucera-Francis); Wall Street Journal, and other sources; some text is parsed and can be searched wih the *tgrep* program. (See http://www.ldc.upenn.edu/)
North American News Text Corpus	Large (~350 million word) corpus of newswire text. (See http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T21)
Wall Street Journal 1987/parsed	~25 million word parsed text from WSJ (text from LDC; parsed version courtesy Eugene Charniak)
Spanish Language News Corpus	Large (~172 million word) corpus of newsire text. (See http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T9)
European Languages News Corpus	~100 million words of French, 90 million words of German, and 15 million words of Portuguese; newswire text. (Seehttp://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T11)
Hansard Parallel Text in English and French	Parallel English/French texts drawn from Canadian Parliament discussions. (Seehttp://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20)
CELEX	Lexical databases (word lemmas, phonology, morphology, frequency) for Dutch, German, and English. (See

	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14)
British National Corpus	(100 million word searchable corpus; Windows software for more extensive searching is also available) http://sara.natcorp.ox.ac.uk/lookup.html