Attitude, Sentiment and Emotion Detection

Alexandros Vrochidis
Department of Informatics, Aristotle
University of Thessaloniki
avrochid@csd.auth.gr

Theodoros Liapikos
Department of Informatics, Aristotle
University of Thessaloniki
tliapikos@csd.auth.gr

ΠΕΡΙΛΗΨΗ

Σε αυτό το άρθρο θα μελετηθεί η ανάλυση κειμένων με σκοπό την διεξαγωγή συμπερασμάτων σχετικά με την κατανόηση των συναισθημάτων του συγγραφέα. Θα μελετηθούν οι κατάλληλες τεχνικές για να επιτευχθεί ο σκοπός αυτός και επίσης θα γίνει αναφορά στα προβλήματα, που μπορεί να προκύψουν σε αντίστοιχες περιπτώσεις, αλλά και σε μεθόδους επίλυσης αυτών. Οι ηλεκτρονικές πηγές κειμένου παρουσίασαν τεράστια αύξηση με την έλευση του web 2.0, καθώς οι άνθρωποι έγιναν πιο πρόθυμοι να εκφράσουν και να μοιραστούν τις απόψεις στους δημόσια στο διαδίκτυο. Σε αυτό συμβάλλει καταλυτικά η εξέλιξη που υπάρχει στα μέσα κοινωνικής δικτύωσης, καθώς μέσω αυτών υπάρχει μεγαλύτερη διαφάνεια απόψεων, γεγονός που δίνει στους χρήστες κίνητρο για να δημιουργήσουν περιεχόμενο.

Ο τομέας της ανάλυσης συναισθημάτων απασχολεί ιδιαίτερα τον επιστημονικό αλλά και τον επιχειρηματικό κλάδο την τελευταία δεκαετία. Είναι γνωστός και ως εξόρυξη γνώμης και είναι ιδιαίτερα σημαντικός καθώς βοηθάει τις επιχειρήσεις να κατανοήσουν τη γνώμη του κοινού τους γι' αυτές. Η παρούσα έρευνα καλύπτει ένα σχετικά μικρό φάσμα της συνολικής επί του θέματος βιβλιογραφίας, αλλά παρουσιάζει τις τελευταίες τεχνολογίες και καινοτομίες που έχουν αναπτυχθεί. Θα αναλυθούν τεχνικές μηχανικής μάθησης καθώς και τεχνικές επεξεργασίας φυσικής γλώσσας, οι οποίες όταν συνδυαστούν επιφέρουν καλά αποτελέσματα. Τέλος θα αναλυθούν κάποια άλυτα ζητήματα τα οποία χρήζουν αντιμετώπισης.

Λέξεις Κλειδιά

εξόρυξη κειμένου; ανάλυση συναισθημάτων; εξόρυξη γνώσης; εξόρυξη χαρακτηριστικών;

1. ΕΙΣΑΓΩΓΗ

Η αυξανόμενη χρήση των δραστηριοτήτων στο διαδίκτυο οδηγεί στη δημιουργία δεδομένων με τεράστιο όγκο, τα οποία αναφέρονται ως Big Data. Αυτά τα δεδομένα μπορούν να αναλυθούν χρησιμοποιώντας συνδυασμό τεχνικών επεξεργασίας φυσικής γλώσσας και τεχνικών εξόρυξης δεδομένων από τον παγκόσμιο ιστό. Υπάρχει πολύ μεγάλος όγκος απόψεων πελάτων από διάφορες επιχειρήσεις, αλλά και απόψεις ανθρώπων σχετικά με μεγάλη ποικιλία ζητημάτων που αφορούν την κοινή γνώμη. Τα συναισθήματα, οι κριτικές και οι αναθεωρήσεις γίνονται όλο και πιο εμφανείς εξαιτίας του αυξανόμενου ενδιαφέροντος για το ηλεκτρονικό εμπόριο, το οποίο αποτελεί μια ιδιαίτερη πηγή έκφρασης και ανάλυσης απόψεων.

Οι περισσότερες από τις αγορές που πραγματοποιούνται σήμερα βασίζονται σε κριτικές που έχουν γραφτεί από χρήστες, που έχουν ήδη δοκιμάσει τα προϊόντα και έμειναν ευχαριστημένοι. Είναι λοιπόν ιδιαίτερα φανερό πως οι κριτικές καθοδηγούν τις τάσεις της αγοράς, γεγονός που καθιστά την ανάλυση τους ιδιαιτέρως σημαντική για τις επιχειρήσεις και όχι μόνο. Έτσι η κατασκευή ενός συστήματος, που θα μπορεί να ανιχνεύει, να αναλύει και να συνοψίζει τα συναισθήματα, που υπάρχουν γύρω από ένα θέμα, είναι μια καινούργια πρόκληση της δεκαετίας μας, ιδιαίτερα σημαντική και με υψηλή ζήτηση.

Την σημαντικότητα αυτού του ζητήματος, μπορεί να την καταλάβει κανείς αν αναλογιστεί το πόσο αχανές είναι το διαδίκτυο, με το 80% των παγκοσμίων δεδομένων να είναι αδόμητα. Τα περισσότερα από αυτά τα δεδομένα προέρχονται από δεδομένα κειμένου, όπως είναι οι αναφορές, τα άρθρα, τα emails και τα chats. Η ανάλυση συναισθημάτων λοιπόν βοηθάει οποιονδήποτε κάνει χρήση αυτής, να βγάλει κάποια συμπεράσματα στην τεράστια, αδόμητη ροή, που υπάρχει μέσω αυτοματοποιημένων διαδικασιών, γεγονός που θα βοηθήσει στην καλύτερη αποτελεσματικότητα σε πολλά ζητήματα.

Η επιστήμη λοιπόν, καλείται να αντιμετωπίσει την μεγάλη ροή πληροφορίας που υπάρχει, να την διαχειριστεί και να καταλήξει σε ασφαλή συμπεράσματα. Ο αυτοματισμός είναι η μόνη λύση που υπάρχει καθώς είναι αδιανόητο να διαχειριστεί κανείς τόσο μεγάλο όγκο δεδομένων με διαφορετικό τρόπο. Ο αυτοματισμός αυτός επιτυγχάνεται μέσω τεχνικών μηχανικής μάθησης. Γίνονται συνεχώς προσπάθειες για την βελτιστοποίηση αυτών των τεχνικών και αυτών των συστημάτων, με σκοπό την επίτευξη ακόμα καλύτερων αποτελεσμάτων, τα οποία θα οδηγούν με περισσότερη σιγουριά στην ασφαλή διεξαγωγή συμπερασμάτων.

Το υπόλοιπο άρθρο έχει οργανωθεί, με βάση το περιεχόμενο, ως εξής. Η Ενότητα 2 περιγράφει βασικές έννοιες που θα αναλυθούν, αλλά και κάποια έτοιμα εμπορικά προϊόντα που μπορούν να χρησιμοποιηθούν. Η Ενότητα 3 αποτελεί την κύρια ενότητα του άρθρου. Στην Ενότητα 3.1 περιγράφεται η προεπεξεργασία, που υφίστανται τα δεδομένα, πριν την ανάλυσή τους. Στην Ενότητα 3.2 γίνεται θεωρητική αναφορά της μεθοδολογίας που επικρατεί για την ανάλυση συναισθημάτων. Στην Ενότητα 3.3 γίνεται ανάλυση των συχνότερων προβλημάτων που εμφανίζονται στην ανάλυση συναισθημάτων. Στην Ενότητα 3.4 γίνεται αναλυτική παρουσίαση αλγοριθμικών πρακτικών προσεγγίσεων και των κατάλληλων μετρικών που χρησιμοποιούνται, αναφορικά με την ανάλυση συναισθημάτων. Τέλος η Ενότητα 4 ολοκληρώνει το άρθρο, παραθέτοντας τα σχετικά συμπεράσματα.

2. ΘΕΜΕΛΙΩΔΕΙΣ ΕΝΝΟΙΕΣ

2.1 Ορισμοί και επεξήγηση των εννοιών που θα χρησιμοποιηθούν

Ως ανάλυση συναισθημάτων ορίζεται η διαδικασία υπολογιστικής μελέτης των απόψεων και των συναισθημάτων των ανθρώπων ως προς μια οντότητα [13]. Σε αυτήν εντάσσεται και η κατηγοριοποίηση των απόψεων που εκφράζονται σε ένα κομμάτι κειμένου και αφορούν τις απόψεις ενός συγγραφέα σχετικά με μια οντότητα. Όταν εφαρμόζεται στον επιχειρηματικό τομέα, τα

κέρδη είναι πολλαπλά και τεράστια. Αρκεί ενδεικτικά να αναφερθεί πως, μέσω τέτοιων τεχνικών, μπορεί μια επιχείρηση να κατανοήσει το σεβασμό που αποπνέει προς τους πελάτες της και να βελτιώσει την εμπειρία των πελατών της. Επίσης, μπορεί να διακρίνει κάποια θέματα προτού λάβουν μεγάλες διαστάσεις και δημιουργηθεί κρίση επάνω στο ζήτημα.

Ένας ακόμη τομέας, όπου θέτει την ανάλυση συναισθημάτων ως σημαντικό χώρο ερευνών, είναι ο καθορισμός του πλάνου μιας επιχείρησης στηριγμένο στους πελάτες και τις απόψεις τους. Η βελτίωση των διαφημιστικών καμπανιών είναι δεδομένη, όταν μια επιχείρηση γνωρίζει τι λείπει από τα προϊόντα της και τι ζητούν οι πελάτες της. Ένας ιδιαίτερα ξεχωριστός και σημαντικός λόγος ύπαρξης της ανάλυσης συναισθημάτων είναι ο εντοπισμός κάποιων ατόμων, τα οποία ασκούν ιδιαίτερη επιρροή (influencers) σε άλλα άτομα, τα οποία αποτελούν ένα σεβαστό πλήθος. Τέλος, μπορεί να βοηθήσει στην καινοτομία καθώς θα γίνονται αποδεκτές πολλές ιδέες των πελατών, οι οποίες μπορούν να οδηγήσουν στην ανάπτυξη πολλών νέων καινοτόμων προϊόντων. Όλα τα παραπάνω γεγονότα είναι ιδιαίτερα σημαντικά και, παρότι αναφέρονται ενδεικτικά για τις επιχειρήσεις, δεν εφαρμόζονται μόνο σε αυτές, αλλά βρίσκουν εφαρμογές σε πολλούς τομείς, όπως η κατανόηση της τάσης των απόψεων σχετικά με ένα πολιτικό πρόσωπο ή ένα περιβαλλοντικό γεγονός.

Συνήθως η εξόρυξη γνώμης περιλαμβάνει κάποιες τεχνικές μηχανικής μάθησης, που εφαρμόζονται πάνω στα κείμενα, προκειμένου να δημιουργηθούν κατηγοριοποιήσεις. Η μηχανική μάθηση (machine learning) είναι ένας επιστημονικός τομέας που αναπτύσσει αλγορίθμους, ώστε να χρησιμοποιηθούν σε σύνολα δεδομένων, με κύριο πεδίο ενδιαφέροντος την πρόβλεψη (παλινδρόμηση), ταξινόμηση και ομαδοποίηση. Αυτές οι εργασίες διακρίνονται σε επιβλεπόμενες (supervised) και μη επιβλεπόμενες (unsupervised), ανάλογα με το αν υπάρχουν εκ των προτέρων γνώση για τα δεδομένα που πρόκειται να τις εφαρμόσουμε. Η μηχανική μάθηση χρησιμοποιείται ολοένα και περισσότερο στις μέρες μας. Οι κύριοι αλγόριθμοι που συναντώνται σε αυτόν τον τομέα είναι αρκετοί. Παρακάτω θα γίνει ιδιαίτερη εστίαση στους αλγορίθμους που χρησιμοποιούνται κατά κύριο λόγο στο συγκεκριμένο τομέα και δεν είναι άλλοι από τους Logistic Regression, Naïve Bayes, και Support Vector Machines. Για τη χρήση των αλγορίθμων μηχανικής μάθησης στην εξόρυξη γνώμης από γραπτά κείμενα, συνήθως απαιτείται τα δεδομένα που χρησιμοποιούνται, και μετά την συλλογή τους, να έχουν υποστεί κάποιου είδους προεπεξεργασία.

Στην προεπεξεργασία των δεδομένων συναντώνται πολλές τεχνικές επεξεργασίας φυσικής γλώσσας (Natural Language Processing, NLP). Με τον όρο επεξεργασία φυσικής γλώσσας, εννοείται η εφαρμογή υπολογιστικών μεθόδων και τεχνικών με σκοπό την ανάλυση της φυσικής γλώσσας που χρησιμοποιούν οι άνθρωποι. Είναι ένας τομέας της επιστήμης των δεδομένων, ο οποίος είναι ιδιαίτερα σημαντικός και αναπτυσσόμενος και βρίσκεις εφαρμογές σε μεγάλο πλήθος περιπτώσεων. Ο λόγος που αυτή η προεργασία είναι απαραίτητη δεν είναι άλλος από το γεγονός της ύπαρξης πολλών αδόμητων στοιχείων, τα οποία παρουσιάζουν τρομερά μεγάλη ανομοιομορφία. Έτσι, όταν κανείς συλλέγει δεδομένα, πρέπει να αποθηκεύει στο σύστημα του μόνο αυτά, τα οποία πιστεύει ότι θα τον βοηθήσουν να εξάγει συμπεράσματα και να εφαρμόζει τεχνικές καθαρισμού για την διαγραφή των πλεονασματικών. Επίσης, στην παραπάνω κατηγορία επεξεργασίας ανήκουν τεχνικές όπως ο διαχωρισμός των προτάσεων σε επιμέρους λέξεις καθώς και η αφαίρεση κάποιων καταλήξεων με σκοπό την επίτευξη καλύτερων αποτελεσμάτων. Στη συνέχεια αυτής της εργασίας θα παρουσιασθούν και θα αναλυθούν οι μέθοδοι με τις οποίες επιτυγχάνεται η προεπεξεργασία δεδομένων καθώς και η χρήση κάποιων εργαλείων που συνεισφέρουν στον τομέα της ανάλυσης.

2.2 Βασικά εργαλεία και εφαρμογές

Παρακάτω γίνεται μια αναφορά σε μερικά από τα εργαλεία που έχουν ήδη αναπτυχθεί και χρησιμοποιούνται σε μεγάλη κλίμακα από τις διάφορες εταιρείες που ενδιαφέρονται για την ανατροφοδότηση των προϊόντων τους και όχι μόνο. Υπάρχουν πάρα πολλά εργαλεία, που βοηθούν στην διεξαγωγή συμπερασμάτων σχετικά με τον συγκεκριμένο σκοπό και παρακάτω θα παρουσιαστούν μερικά από αυτά.

Ένα από τα γνωστά εργαλεία, στο συγκεκριμένο τομέα, αποτελεί το Quick Search, που παρέχει δωρεάν demo στους χρήστες τους και χρησιμοποιεί τα κοινωνικά δίκτυα (social networks) για την εξόρυξη των δεδομένων του. Επίσης περιλαμβάνει και ιστοσελίδες ενημέρωσης, φόρουμ καθώς και blogs. Επιτρέπει την κατηγοριοποίηση με βάση δημογραφικά χαρακτηριστικά, ώστε να γίνει εστίαση σε κάποια γεωγραφική περιοχή, γεγονός που το καθιστά ιδιαίτερα χρήσιμο. Το γραφικό περιβάλλον του είναι ιδιαίτερα φιλικό προς τον χρήστη, επιτρέποντάς του να εκμεταλλευτεί τις δυνάμεις και τις αδυναμίες του για μια επιτυχημένη στρατηγική μάρκετινγκ. Δυστυχώς, λόγω ειδικών περιορισμών από την εταιρία ανάπτυξης, δεν ήταν εφικτή η χρήση του με σκοπό την περαιτέρω ανάλυση του. Μέσα από αυτό το εργαλείο μπορεί να γίνει η εισαγωγή δύο ή περισσοτέρων brands και έπειτα η απεικόνιση των θετικών ή αρνητικών συναισθημάτων που προκύπτουν από τα δεδομένα του κόσμου, ο οποίος μιλάει γι' αυτά στις προαναφερθέντες πηγές. Το γραφικό

Εικόνα 1. Παρουσίαση του εργαλείου Quick Search

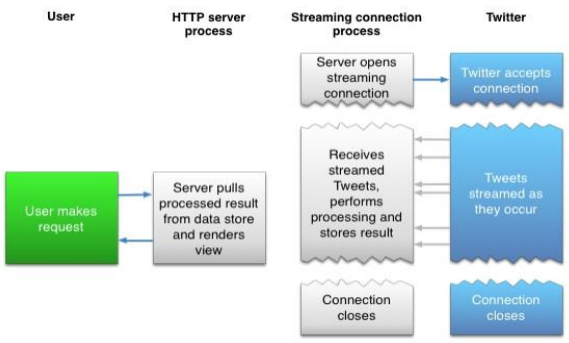
του περιβάλλον απεικονίζεται παρακάτω.

Ένα ακόμη ιδιαίτερο εργαλείο είναι το Hootsuite Insights, που αναφέρει πως συνδυάζει εκατό εκατομμύρια πηγές για να εξάγει τα δεδομένα που χρησιμοποιεί. Οι πηγές του είναι παρόμοιες με το παραπάνω εργαλείο, αλλά διαφέρει στην απεικόνιση των αποτελεσμάτων και στο γραφικό περιβάλλον. Δέχεται τους όρους που ο υποψήφιος χρήστης θέλει να μελετήσει και έπειτα παρουσιάζει τις αναφορές που έχουν γίνει γύρω από τον όρο, εκφρασμένες σε ποσοστά θετικών, αρνητικών και ουδέτερων αναφορών. Επιτρέπει την ύπαρξη κάποιων φίλτρων, όπως είναι του φύλου, της τοποθεσίας αλλά και της γλώσσας. Επιπροσθέτως δίνει την επιλογή στο χρήστη να πραγματοποιήσει εξόρυξη γενικότερων αναφορών, που είναι μόνο θετικές ή αρνητικές χωρίς να σχετίζονται όλες γύρω από έναν συγκεκριμένο όρο. Ο χρόνος εκτέλεσης του προγράμματος είναι ιδιαίτερα γρήγορος, γεγονός που το καθιστά ένα σημαντικό εργαλείο για μια επιχείρηση που έχει πολύ περιορισμένο χρόνο. Είναι διαθέσιμο σε πενήντα γλώσσες συμπεριλαμβανομένων των ελληνικών, γεγονός αρκετά σπάνιο καθώς θα πρέπει για κάθε γλώσσα να υπάρχουν όροι, με

τους οποίους θα γίνει η αναγνώριση της γλώσσας του κειμένου και να καταταχθεί στο ανάλογο συναίσθημα. Οι ποικιλία των γλωσσών είναι ένα θέμα που έχει προκύψει στον συγκεκριμένο τομέα και φαίνεται πως στη συγκεκριμένη περίπτωση έχει λυθεί.

Ένα άλλο εργαλείο για ανάλυση συναισθημάτων είναι το Sentiment Analyzer της Daniel Soper. Το σημαντικότερο του προτέρημα είναι ότι αποτελεί ένα εργαλείο, το οποίο μπορεί κάθε χρήστης να το χρησιμοποιήσει δωρεάν. Οι απαιτήσεις που ικανοποιεί επομένως είναι λίγο περιορισμένες, όπως επίσης και το γραφικό του περιβάλλον. Για να λειτουργήσει έχει εκπαιδευτεί με περισσότερα από 8000 δείγματα κειμένων, γεγονός που το καθιστά αρκετά καλό σε ακρίβεια. Μπορεί να δεχτεί αρκετά μεγάλα κείμενα και έπειτα, σε πολύ μικρό χρονικό διάστημα, να τα αναλύσει και να δείξει τον γενικό βαθμό του συναισθήματος που αυτά εκφράζουν. Η κλίμακα του είναι από το -100, το οποίο δείχνει το αρνητικό συναίσθημα σε όλο το μήκος του κειμένου και φτάνει στο +100, το οποίο αναφέρεται σε πλήρες θετικές αναφορές κατά μήκους του κειμένου. Σαν εργαλείο αποδίδει πολύ υψηλά όταν γίνεται χρήση της αγγλικής γλώσσας, αλλά απέδωσε εξίσου καλά και σε δοκιμές που έγιναν στην ελληνική γλώσσα, αναγνωρίζοντας λέξεις με αρνητικό αντίκτυπο. Αν σκεφτεί κανείς ότι διανέμεται δωρεάν στο κοινό, είναι ένα εξαιρετικό εργαλείο για απλές εργασίες. Το αρνητικό του είναι ότι δεν μπορεί να δεχτεί ροή δεδομένων και να την κατατάξει σε επιμέρους κατηγορίες. Σε τέτοιες περιπτώσεις πρέπει να προτιμηθούν τα δύο προηγούμενα εργαλεία.

Πολλά micro-blogging sites παρέχουν κάποια εργαλεία Application Programming Interface (API), τα οποία βοηθούν τους χρήστες που βρίσκονται σε αυτά να εξάγουν δεδομένα και να τα χρησιμοποιούν για ανάλυση. Πολύ καλή δουλειά στο συγκεκριμένο τομέα φαίνεται να γίνεται από το μέσο κοινωνικής δικτύωσης του Twitter, το οποίο προσφέρει δωρεάν, σε κάποιες περιπτώσεις, τα αντίστοιχα εργαλεία, που έχει αναπτύξει γι' αυτό το σκοπό. Μάλιστα έχουν αναπτυχθεί δύο πολύ αξιόλογα εργαλεία, τα οποία διαφέρουν ως προς την στατικότητα των δεδομένων, τα οποία θέλει να μελετήσει ο πιθανός ερευνητής. Το πρώτο εργαλείο ονομάζεται Twitter REST API και αφορά τη συλλογή δεδομένων, τα οποία είναι στατικά και δεν έχουν ροή, όπως είναι τα στοιχεία των χρηστών αλλά και κάποια tweets, τα οποία έχουν γίνει σε παρελθοντική χρονική στιγμή. Για συνεχείς ροές δεδομένων υπάρχει το αντίστοιχο Twitter Streaming API, μέσω του οποίου μπορεί κάθε χρήστης, που του δίνεται η άδεια, να ανακτήσει ροές δεδομένων, μέσω μιας διαδικασίας στην οποία γίνεται πρώτα μια αίτηση στον ανάλογο server και έπειτα ο server με τη σειρά του ανοίγει μια σύνδεση με το Twitter. Αν αυτή η σύνδεση γίνει δεκτή, τότε ξεκινάει η καταγραφή των tweets που γίνονται από τη στιγμή που πραγματοποιήθηκε η σύνδεση, μέχρι τη στιγμή που η σύνδεση θα διακοπεί. Αφού παραληφθούν τα δεδομένα που χρειάζονται ο server τα δίνει πίσω στο χρήστη κλείνοντας έτσι την συνεδρία. Συνήθως η δωρεάν μορφή αυτών των εφαρμογών γίνεται για συγκεκριμένο αριθμό δεδομένων



Εικόνα 2. Παρουσίαση της διαδικασίας του Twitter Streaming API

εξόρυξης, που φτάνει στην περίπτωση του Twitter τα 10.000 δεδομένα. Στην περίπτωση του streaming δεν υπάρχει αυτός ο περιορισμός, αλλά η ροή μπορεί να σταματήσει οποτεδήποτε χωρίς να έχει δοθεί αντίστοιχη εντολή, γεγονός που εντάσσεται στα αρνητικά στοιχεία της. Η διαδικασία που ακολουθείται στην εφαρμογή εμφανίζεται στην Εικόνα 2.

3. ΜΟΝΤΕΛΑ

3.1 Προεπεξεργασία των δεδομένων

Στα πλαίσια της εξόρυξης κειμένου υπάρχει ιδιαίτερη ανάγκη για εφαρμογή μεθόδων, οι οποίες θα μετατρέψουν τα δεδομένα στην κατάλληλη μορφή έτσι ώστε να είναι δυνατόν έπειτα να γίνουν διαχειρίσιμα από τους διάφορους αλγόριθμους που θα εφαρμοστούν σε αυτά. Η ανάγκη της προεπεξεργασίας των δεδομένων έγκειται στο γεγονός, πως αυτά αποτελούν μια μεγάλη αδόμητη μάζα. Πλέον σπάνια υπάρχουν στατικά δεδομένα καθώς, στις περισσότερες περιπτώσεις, συναντώνται μεγάλες ροές δεδομένων που δεν σταματούν σε καμία χρονική στιγμή. Το γεγονός αυτό καθιστά επιτακτική την ανάγκη ταχύτερης εφαρμογής των αλγορίθμων με όσο το δυνατόν καλύτερα ποσοστά επιτυχίας [1].

Οι τεχνικές επεξεργασίας κειμένου έχουν κάποια κοινά χαρακτηριστικά μεταξύ τους και αυτό που τις διαφοροποιεί από τις τεχνικές επεξεργασίας φυσικής γλώσσας είναι η χρήση του κειμένου όχι ως μια ενιαία οντότητα αλλά εστιάζοντας κυρίως πάνω στις ίδιες τις λέξεις. Κάθε λέξη αποτελεί μια μονάδα ανάλυσης και μπορεί να μετρηθεί ως προς την συχνότητα της στο κείμενο και όχι μόνο. Από την άλλη μεριά οι τεχνικές επεξεργασίας της φυσικής γλώσσας υπολογίζουν το υποκείμενο καθώς και τα μεταδεδομένα που μπορεί να περιέχονται [4]. Δεν υπολογίζουν δηλαδή μόνο τη λέξη αυτή καθ' αυτή αλλά και το περιβάλλον γύρω της, προκειμένου να βγάλουν συμπεράσματα. Μέσω τέτοιων διεργασιών, οι οποίες εφαρμόζονται αμέσως μετά τη συλλογή των δεδομένων, μπορούν να επιτευχθούν αποτελέσματα ανώτερα αυτών που θα επιτυγχάνονταν χωρίς την εφαρμογή αυτών. Για παράδειγμα στην πρόταση 'δεν θα πρότεινα ποτέ αυτήν την ταινία' οι τεχνικές επεξεργασίας κειμένου θα έδιναν ένα θετικό αντίκτυπο, λόγω της θετικής λέξης 'προτείνω', ενώ στην περιοχή της επεξεργασίας φυσικής γλώσσας θα ληφθούν υπ' όψη τα συμφραζόμενα και θα οριστεί η κριτική ως αρνητική. Παρακάτω θα αναλυθούν μερικές γενικότερες τεχνικές προεπεξεργασίας δεδομένων, οι οποίες συνεισφέρουν θετικά τις περισσότερες φορές, που χρησιμοποιούνται.

Η πρώτη τεχνική που θα αναλυθεί είναι αυτή του tokenization. Χρησιμοποιείται για να σπάσει ένα κείμενο σε προτάσεις, λέξεις ή σύμβολα αφαιρώντας τα σημεία στίξης [9]. Στις περισσότερες περιπτώσεις γίνεται η είσοδος ενός κειμένου και έπειτα σπάει κάθε πρόταση του κειμένου σε επιμέρους λέξεις. Όταν γίνεται ο διαχωρισμός του κειμένου σε λέξεις είναι πολύ πιο εύκολα διαχειρίσιμο, καθώς ακολουθεί την τακτική του διαίρει και βασίλευε. Το ότι ένα κείμενο διαιρείται σε λέξεις δε σημαίνει φυσικά ότι δεν λαμβάνονται υπόψη οι προηγούμενες και οι επόμενες λέξεις που εξετάζονται σε κάθε περίπτωση, καθώς όπως φάνηκε στο προηγούμενο παράδειγμα, παίζουν ιδιαίτερο ρόλο.

Στις περισσότερες τεχνικές προεπεξεργασίας, που συναντώνται στη βιβλιογραφία, γίνεται εύκολα διακριτό πως σχεδόν κάθε φορά χρησιμοποιείται η αφαίρεση των σημείων στίξης. Αυτό συμβαίνει λόγω του γεγονότος ότι τα σημεία στίξης δεν έχουν κάτι ιδιαίτερο να προσφέρουν στον ερευνητή. Δεν διευκολύνουν στη διεξαγωγή κάποιου συμπεράσματος και η αφαίρεση τους οδηγεί σε μικρότερο όγκο δεδομένων για επεξεργασία και επομένως σε

καλύτερες ταχύτητες και υψηλότερα ποσοστά ακρίβειας. Γι' αυτό το λόγο στις περισσότερες περιπτώσεις όπου αφαιρούνται τα σημεία στίξης παρατηρείται βελτίωση της απόδοσης των αλγορίθμων.

Μια τεχνική, που τις περισσότερες φορές επιφέρει θετικά αποτελέσματα, αλλά δε συμβαίνει για όλες τις περιπτώσεις, είναι το stemming. Σε αυτήν την τεχνική εφαρμόζεται η διαδικασία της μείωσης γραμμάτων από κάθε λέξης κρατώντας μόνο την ρίζα της λέξης. Μελετάται από το 1960 σαν τεχνική αλγορίθμων και εμφανίζει θετικά αποτελέσματα, κυρίως όταν αφορά εύρεση παρόμοιων λέξεων, καθώς ο πυρήνας της κάθε λέξης έχει ιδιαίτερη σημασία σε τέτοιες περιπτώσεις. Στο stemming μπορεί να αφαιρεθεί είτε η κατάληξη μιας λέξης είτε το πρόθεμα της. Μερικές φορές δεν παρατηρείται διαφορά στην απόδοση με τη χρήση αυτής της τεχνικής ενώ μερικές φορές επιδρά αρνητικά. Αυτός είναι και ο λόγος, που θα πρέπει να χρησιμοποιείται μετά από σκέψη, ανάλογα με το σκοπό που πρέπει να επιτευχθεί.

Στις τεχνικές που έχουν ήδη αναφερθεί έρχεται να προστεθεί και το Part Of Speech (POS) tagging, το οποίο συμβάλει στην γραμματική αποσαφήνιση των λέξεων σε κατηγορίες όπως τα ρήματα και τα επίθετα. Αυτή η κατηγοριοποίηση γίνεται με βάση τις γειτονικές και τις σχετικές λέξεις σε μια φράση. Χρησιμοποιεί αλγορίθμους βασισμένους σε ένα σύνολο περιγραφικών ετικετών, για να συνδέσει διακριτούς όρους αλλά και να βρει κάποια κρυμμένα τμήματα του λόγου, όπως τα απαρέμφατα και τα αντίστοιχα μέρη του λόγου. Η αναγνώριση διαφορετικών μερών του λόγου σε ένα κείμενο είναι ιδιαίτερα σημαντική για την επεξεργασία της φυσικής γλώσσας και επομένως πρέπει να γίνονται παρόμοιες διαδικασίες με σκοπό τη βελτιστοποίηση.

Όπως έχει ήδη αναφερθεί, όλες οι τεχνικές προεπεξεργασίας λαμβάνουν μέρος στη διαδικασία της ανάλυσης συναισθημάτων προκειμένου να υπάρξει μια δόμηση των δεδομένων που έχουν εξορυχθεί. Μια τεχνική που συμβάλει σε αυτήν την ενέργεια είναι η εξαγωγή οντότητας (entity extraction) η οποία είναι μια τεχνική εξαγωγής πληροφοριών που αναφέρεται στη διαδικασία ταυτοποίησης και κατηγοριοποίησης βασικών στοιχείων ενός κειμένου σε επιμέρους προκαθορισμένες κατηγορίες. Έτσι προκύπτουν δομημένα στοιχεία, έτοιμα για μηχανική ανάγνωση και εξαγωγή συμπερασμάτων. Οι κατηγορίες στις οποίες εντάσσονται μπορεί να αφορούν οντότητες ανθρώπων, τοποθεσιών, αλλά και χρόνο, ποσά πληρωμών, τηλέφωνα κ.α. Έτσι κάθε μηχάνημα που θα διαβάζει τις σχετικές πληροφορίες θα είναι σε θέση να γνωρίζει που ανήκει η κάθε οντότητα και που αναφέρεται. Επομένως, είναι μια πολύ χρήσιμη τεχνική που χρησιμεύει ως σημείο εκκίνησης για περαιτέρω ανάλυση και διερεύνηση.

Μια παρόμοια τεχνική με τη προηγούμενη είναι η εξαγωγή συσχέτισης η οποία παίζει πολύ σημαντικό ρόλο στην εξαγωγή δομημένων πληροφοριών από αδόμητες πηγές. Είναι μια τεχνική που κάνει εμφανή τη διαφορά μεταξύ της επεξεργασίας φυσικής γλώσσας και της απλής επεξεργασίας κειμένου. Δείχνει την εξέλιξη που υπάρχει και μέσω αυτής μπορεί κάποιος να βρει συσχέτιση μεταξύ δύο οντοτήτων. Για να εφαρμοστεί χρησιμοποιεί μοντέλα μηχανικής μάθησης προκειμένου να είναι σε θέση να εντάξει σε κάποια σχέση δύο οντότητες. Για παράδειγμα στην πρόταση 'Ο Ομπάμα είναι πρόεδρος τον Ηνωμένων Πολιτειών της Αμερικής' πρέπει να υπάρξει συσχέτιση προέδρου μεταξύ Αμερικής και Ομπάμα. Για να συμβεί αυτό έχει εκπαιδευτεί με αντίστοιχα κείμενα ένα μοντέλο μέσω μηχανικής μάθησης. Αφού πραγματοποιηθεί η εκπαίδευση τότε είναι εφικτή η εξαγωγή συμπερασμάτων και για παρόμοιες περιπτώσεις.

Σε μεγάλες ροές δεδομένων, όπου ο όγκος είναι τεράστιος συμβάλουν τεχνικές εξαγωγής χαρακτηριστικών (feature extraction), οι οποίες δέχονται ένα αρχικό σύνολο δεδομένων και παράγουν χαρακτηριστικά, τα οποία πρέπει να είναι όσο το δυνατόν πιο χρήσιμα για τη εξαγωγή συμπερασμάτων. Έτσι έχουμε μείωση περιττών πληροφοριών και καλύτερες ερμηνείες. Τέτοιες τεχνικές μειώνουν τις διαστάσεις των προβλημάτων καθώς μειώνονται τα αδόμητα χαρακτηριστικά σε διαχειρίσιμες ομάδες και παράλληλα εξακολουθούν να περιγράφουν το αρχικό σύνολο δεδομένων με την ίδια συνέπεια. Συμβάλουν στην ταχύτητα εκτέλεσης των αλγορίθμων που θα διαχειριστούν τα δεδομένα μετά την εφαρμογή των τεχνικών αυτών και πολλές φορές συμβάλουν και στην εκτέλεση καθώς χωρίς τεχνικές τέτοιου είδους η επεξεργασία πληροφορίας σε μερικές περιπτώσεις δεν οδηγεί στην απόδοση καλών αποτελεσμάτων.

Ο διαχωρισμός αυτών των χαρακτηριστικών είναι ιδιαίτερα σημαντικός και αποτελεί ένα είδος τεχνικής από μόνος του. Στην περίπτωση του κοινωνικού μέσου Twitter για παράδειγμα υπάρχουν 4 διαφορετικές κατηγορίες χαρακτηριστικών, που πρέπει να κατηγοριοποιηθούν [7]. Η πρώτη από αυτές είναι τα σημασιολογικά χαρακτηριστικά (semantic features) δηλαδή τα χαρακτηριστικά που θα αποκαλύψουν θετικό ή αρνητικό συναίσθημα. Η δεύτερη αφορά τα συντακτικά χαρακτηριστικά (syntactic features) και εμπλέκονται τα n-grams, που θα αναφερθούν παρακάτω, καθώς και τεχνικές, όπως η POS, που αναφέρθηκε προηγουμένως. Σκοπός της κατηγορίας είναι η συντακτική απεικόνιση των λέξεων. Τρίτη κατηγορία είναι τα στιλιστικά χαρακτηριστικά (stylistic features) τα οποία αναφέρονται στην ιδιομορφία που παρουσιάζει η κάθε γραφή και επαφίεται στον ανάλογο συγγραφέα. Σε αυτά εντάσσονται κάποιες λέξεις που μπορεί να χρησιμοποιεί ο συγγραφέας σε δική του διάλεκτο (slang) καθώς και κάποια emoticons, τα οποία είναι τόσο χρήσιμα για τη διεξαγωγή συμπερασμάτων, που μπορούν από μόνα τους να συνεισφέρουν σε μια ανάλυση συναισθημάτων. Οι τρεις παραπάνω κατηγορίες μπορούν να υπάρξουν οπουδήποτε. Η τέταρτη κατηγορία αφορά το Twitter, αλλά ανάλογες κατηγορίες υπάρχουν σε κάθε μέσο κοινωνικής δικτύωσης και όχι μόνο. Αυτή η κατηγορία ονομάζεται ειδικά χαρακτηριστικά (specific features) και αναφέρεται σε tweets, retweets, hashtags, usernames και αντίστοιχες οντότητες των κοινωνικών μέσων.

Η μοναδική τεχνική που αναφέρθηκε και χρήζει ανάλυσης είναι τα n-grams. Αυτά είναι μια συνεχόμενη ακολουθία στοιχείων, τα οποία προέρχονται από ένα δείγμα κειμένου. Χωρίζονται σε bigram, trigram κλπ., αναλόγως με το πόσες λέξεις αφορούν. Είναι αυτά στα οποία επαφίεται η εξέλιξη της απλής εξόρυξης κειμένου σε τεχνικές επεξεργασίας φυσικής γλώσσας. Στην ουσία μέσω τέτοιων τεχνικών οι οντότητες λαμβάνουν υπόψη τις λέξεις πριν και μετά από αυτές. Το μέγεθος του n αναφέρεται στον αριθμό των λέξεων που θα προσμετρήσουν. Για παράδειγμα σε ένα pentamer θα προσμετρήσουν οι δύο λέξεις πριν και μετά την λέξη που θα γίνει η εστίαση. Πολλές φορές υπάρχει συνδυασμός n-grams, ο οποίος σε κάποιες περιπτώσεις επιφέρει καλύτερα συμπεράσματα. Έτσι, μπορούνε να βγαίνουν συνδυαστικά συμπεράσματα και όχι μεμονωμένα για κάθε λέξη. Τα πλεονεκτήματα αυτής της μεθόδου εξηγήθηκαν προηγουμένως και εστιάζουν στην σωστή κατανόηση του περιεχομένου, βάσει της συνολικής εικόνας και όχι μόνο μιας λέξης.

3.2 Μεθοδολογία ανάλυσης συναισθημάτων

Οι τεχνικές ανάλυσης συναισθημάτων χωρίζονται σε δύο βασικά επίπεδα μελέτης. Το ένα επίπεδο εξετάζει το χαρακτηρισμό ενός

κειμένου ως θετικό, αρνητικό ή ουδέτερο με βάση το περιεχόμενό του, ενώ το δεύτερο επίπεδο εστιάζει σε μεμονωμένα συναισθήματα τα οποία χαρακτηρίζουν το κείμενο (πχ χαρά, λύπη, θύμος κ.α.). Επίσης θα αναφερθούν μερικά προβλήματα που σχετίζονται με την ανάλυση συναισθημάτων, καθώς και τρόποι, με τους οποίους η επίλυση τους γίνεται πιο εύκολη.

Η ανάλυση συναισθημάτων μπορεί να γίνει σε τέσσερα διαφορετικά επίπεδα όσον αφορά το μέγεθος του κειμένου που εξετάζεται και κατηγοριοποιείται κάθε φορά. Το πρώτο επίπεδο αφορά την ανάλυση σε επίπεδο λέξεων. Κάθε λέξη θεωρείται ως μοναδική οντότητα και αποδίδεται σε αυτήν ένας χαρακτηρισμός συναισθημάτων που επιφέρει. Σε τέτοιο επίπεδο βρίσκουν εφαρμογή παλαιότερες τεχνικές εξόρυξης γνώσης από δεδομένα κειμένου χωρίς εφαρμογή τεχνικών επεξεργασίας φυσικής γλώσσας, όπου χρησιμοποιούνται συνήθως προτάσεις ολόκληρες στην κατηγοριοποίηση.

Αυτό είναι και το δεύτερο επίπεδο. Κάθε λέξη δεν αποτελεί ξεχωριστή οντότητα. Ο αναλυτής κατηγοριοποιεί με βάση το συναίσθημα κάθε πρόταση στο σύνολο της. Αυτό πολλές φορές αποδίδει καλύτερα αποτελέσματα, καθώς αποφεύγονται περιπτώσεις όπου μια λέξη θα αναιρεί και θα αντιστρέφει το νόημα της αμέσως επόμενης λέξης. Για παράδειγμα στη πρόταση ‘δεν μου άρεσε’ θα υπάρξει αρνητικός χαρακτηρισμός όταν μετρηθεί σαν πρόταση, ενώ όταν μετρηθεί κατά λέξη θα υπάρξει το ‘δεν’ σαν αρνητικό και το ‘άρεσε’ σαν θετικό, οπότε θα χαρακτηριστεί ως κάτι ουδέτερο.

Το τρίτο επίπεδο κατηγοριοποίησης είναι σε επίπεδο κειμένου. Σε αυτό το επίπεδο ένα κείμενο κατηγοριοποιείται σαν ένα ενιαίο σύνολο. Αυτό γίνεται με βάση το αν οι λέξεις που χρησιμοποιεί επί το πλείστον έχουν θετική ή αρνητική σημασία.

Αντίστοιχο είναι και το τέταρτο και τελευταίο επίπεδο, το οποίο στην ουσία μελετάει κείμενα τα οποία αναφέρονται σε μια συγκεκριμένη οντότητα, όπως για παράδειγμα ένα κινητό τηλέφωνο. Σε αυτήν την κατηγορία μελετάται το κείμενο όχι ως προς το σύνολο του, αλλά ως προς την στάση του απέναντι στην οντότητα. Αν είναι περισσότερες οι θετικές αναφορές στην οντότητα, τότε το κείμενο κατατάσσεται ως θετικό. Δεν έχει λοιπόν σημασία αν το κείμενο συνολικά περιέχει περισσότερες λέξεις αρνητικές. Αυτές που μετράνε στο αποτέλεσμα είναι μόνο αυτές που αναφέρονται σε συγκεκριμένη οντότητα, για την οποία διεξάγεται έρευνα. Παρότι έχουν προταθεί πολλά παρόμοια συστήματα, κυρίως σε ερευνητικό επίπεδο, δεν υπάρχει καθιερωμένος τρόπος σύνθεσης ενός αντίστοιχου συστήματος, αλλά υπάρχουν καθιερωμένα μέτρα αξιολόγησης.

Το παραπάνω επίπεδο λοιπόν λύνει πολλά θέματα που προκύπτουν από τις μεθόδους της ανάλυσης πολικότητας (polarity analysis). Σε αυτές τις μεθόδους μετρούνται μόνο οι θετικοί, αρνητικοί και ουδέτεροι όροι κάθε πρότασης. Αυτό από μόνο του δεν είναι επαρκές για να δώσει μια ικανοποιητική εικόνα στον ερευνητή ή την επιχείρηση, που ενδιαφέρεται να έχει μια εικόνα των σχόλιων που την αφορούν. Αυτό συμβαίνει γιατί δεν θα είναι σε θέση να γνωρίζει ποια κομμάτια είναι αρνητικά φορτισμένα και ποια θετικά ώστε να δοθεί η ανάλογη βαρύτητα σε αυτά. Επίσης σε περιπτώσεις όπου υπάρχει στην ίδια πρόταση αρνητική αναφορά σε μια οντότητα και θετική αναφορά σε άλλη οντότητα θα υπάρχει κατηγοριοποίηση ως κάτι ουδέτερο. Αυτό σαν γεγονός δίνει λάθος εικόνα, καθώς χάνεται πληροφορία. Πρέπει ο ερευνητής ή η επιχείρηση να είναι σε θέση να ξέρει ακριβώς τι είναι χαρακτηρισμένο αρνητικά και τι θετικά. Σε αυτό έρχεται να συμβάλει η παραπάνω μέθοδος που μελετάει τις αναφορές σχετικά με κάθε οντότητα ξεχωριστά.

Για μια επιτυχημένη ανάλυση συναισθήματος πρέπει να υπάρχουν σωστά λεξικά συναισθήματος (sentiment lexicon). Ένα λεξικό συναισθήματος είναι μια λίστα λέξεων ή φράσεων, που συνήθως χρησιμοποιούνται για να εκφράσουν ένα θετικό ή αρνητικό συναίσθημα [8]. Οι περισσότερες τεχνικές που εφαρμόζονται βασίζονται σε λεξικά που έχουν φτιαχτεί προηγουμένως γι’ αυτό το σκοπό. Επομένως, αν τα λεξικά αυτά δεν είναι φτιαγμένα σωστά, οποιαδήποτε ανάλυση θα αποτύχει, ασχέτως των μεθόδων που θα χρησιμοποιηθούν. Η ανάπτυξη τέτοιων λεξικών αποτελεί σημαντικό ζήτημα και απασχολεί πολλούς ερευνητές.

Τα λεξικά βασίζονται κυρίως σε βάρη που δίνουν σε κάθε λέξη, τα οποία είναι είτε θετικά, είτε αρνητικά ανάλογα με το νόημά τους. Επίσης βρίσκουν λέξεις συνώνυμες και αντώνυμες και δίνουν βάρη αντίστοιχα με τη λέξη που υπάρχει ομοιότητα. Αν είναι αντώνυμες τότε δίνεται το αντίστοιχο αρνητικό βάρος. Έτσι μπορούν να επεκταθούν λεξικά που έχουν μόνο λίγες λέξεις με δοσμένο βάρος. Να αναφερθεί στο σημείο αυτό πως όσες λέξεις δεν υπάρχουν στο λεξικό θεωρούνται ως ουδέτερες και δεν μετρούνται στα αποτελέσματα. Μια ακόμη τεχνική ανάπτυξης τέτοιων λεξικών είναι με τη χρήση σημασιολογικών όρων αλλά και συντακτικών όρων, όπως το συζευκτικό ‘και’. Σε αυτή την περίπτωση χρησιμοποιείται ένα σύστημα που ονομάζεται Pointwise Mutual Information (PMI), που είναι σύστημα χρήσης σημείων αλληλοκατανόησης. Έχοντας ένα λεξικό συναισθημάτων που έχει λίγους όρους μπορεί κάποιος να το επεκτείνει χρησιμοποιώντας νέα κείμενα και λαμβάνοντας υπόψη τα σημασιολογικά χαρακτηριστικά τους. Λέξεις που συνδέονται με ‘και’ θα έχουν παρόμοια σημασία και ίσως θα εκφράζουν παρόμοιο συναίσθημα. Το ίδιο θα συμβαίνει και με λέξεις που έχουν την ίδια συντακτική και σημασιολογική ρίζα.

Μια άλλη τεχνική, που εφαρμόζεται σε θέματα ανάλυσης συναισθημάτων, είναι η ανάλυση υποκειμενικότητας (subjectivity analysis). Ο σκοπός είναι να βρεθεί εάν μια πρόταση είναι αντικειμενική και ισχύει σε γενικό βαθμό όπως ότι ‘το καλοκαίρι έχει συνήθως ηλιόλουστο καιρό’, ή εάν είναι υποκειμενική και εκφέρει τη γνώμη του συγγραφέα, χωρίς αυτό να ισχύει για το γενικό σύνολο [2], π.χ. μια κριτική όπου αναφέρει ‘Το χειρότερο κινητό που έχω δοκιμάσει’. Για να επιτευχθεί αυτό, ένα κείμενο χωρίζεται σε προτάσεις και έπειτα μέσω αλγορίθμων μηχανικής μάθησης κατατάσσονται οι προτάσεις ανάλογα αν αυτές αφορούν μια γνώμη συγγραφέα ή είναι γενικές και ισχύουν για κάθε περίπτωση εκφέροντας αντικειμενικότητα.

Μέσω των προηγούμενων τεχνικών δίνεται η δυνατότητα της ανάλυσης και σύγκρισης προτάσεων. Η ανάλυση μπορεί αν γίνει μέσω δύο τρόπων: απευθείας ανάλυση και συγκριτική ανάλυση [10]. Στην πρώτη περίπτωση το αποτέλεσμα είναι θετικό, αρνητικό ή ουδέτερο, ενώ στη δεύτερη περίπτωση εκφράζεται η αρέσκεια η δυσαρέσκεια κάποιου, με τη μορφή σύγκρισης με μια άλλη οντότητα. Ένα τέτοιο παράδειγμα είναι ‘Η κάμερα Χ έχει πολύ χειρότερη ποιότητα από την κάμερα Υ’. Είναι φανερό πως πρέπει να δίνεται τέτοια δυνατότητα σε μια επιχείρηση, ώστε να μπορεί να καθορίσει ποια αντικείμενα προτιμώνται έναντι των προϊόντων της.

Φυσικά της παραπάνω ανάλυσης προηγείται ο εντοπισμός της γνώμης και η εξαγωγή της. Αν, π.χ., κάποιος θέλει να δει γνώμες σχετικά με ένα διφορούμενο θέμα πρέπει να εφαρμόσει τεχνικές για να βρει άρθρα και προτάσεις για το συγκεκριμένο θέμα και έπειτα να εφαρμόσει τεχνικές για την ανάλυση συναισθημάτων, ώστε να μπορεί να διεξάγει κάποια συμπεράσματα. Τα παραπάνω είναι γνωστά και ως opinion search and retrieval.

Ένας ακόμη ιδιαίτερος τομέας είναι αυτός της εύρεσης των ανεπιθύμητων κειμένων (spam) [3]. Δεν αποτελεί έκπληξη άλλωστε πως πολλοί άνθρωποι έχουν αρχίζει να 'παίζουν' με τα συστήματα των κοινωνικών δικτύων δημιουργώντας παραπληροφοριτικό κείμενο ή/και αναπαράγοντάς το. Μέσω του opinion spam γίνεται αναφορά σε ψεύτικες αναφορές, οι οποίες δεν προσδίδουν την άποψη του συγγραφέα και προσπαθούν σκόπιμα να παραπλανήσουν τους αναγνώστες. Πολλές φορές προέρχονται από αυτοματοποιημένα συστήματα δίνοντας θετικές κριτικές σε προϊόντα με σκοπό την προώθησή τους. Φυσικά υπάρχουν και περιπτώσεις που μπαίνουν αυτοματοποιημένα, αρνητικά σχόλια στον ανταγωνισμό. Η εύρεση αυτών των σχολίων είναι μια δύσκολη διαδικασία, που είναι όμως πολύ χρήσιμη. Για να αντιμετωπιστεί γίνεται χρήση της χρησιμότητας των κριτικών, η οποία στην ουσία δείχνει πόσοι χρήστες βρήκαν χρήσιμη μια κριτική. Αυτό μπορεί να στηρίζεται είτε σε θετικές ψήφους της κριτικής, είτε απλά σε πολλές αναγνώσεις. Οι ποιοτικές απόψεις λοιπόν έχουν μεγαλύτερο βάρος από τις άλλες και έτσι μετριάζεται κάπως το συγκεκριμένο θέμα.

Η παραπάνω διαδικασία ονομάζεται μέτρηση χρησιμότητας (usefulness measurement) και υπάρχουν αρκετοί τρόποι ανάπτυξής της εκτός από αυτούς που προαναφέρθηκαν [9]. Ο προσδιορισμός διαφόρων μετρικών της χρησιμότητας είναι καθοριστικός σε αυτόν τον τομέα. Χρησιμοποιούνται τεχνικές που προαναφέρθηκαν για να μετρηθεί η υποκειμενικότητα της κριτικής, καθώς και η μέτρηση των ορθογραφικών σφαλμάτων. Επίσης γίνεται μέτρηση της μέσης χρησιμότητας παλαιοτέρων κριτικών προκειμένου να δοθεί έμφαση στην αξία μιας νέας κριτικής. Για να ταξινομηθούν οι κριτικές σε spam και κανονικές χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης, οι οποίοι θα μελετηθούν στη συνέχεια της εργασίας.

Σε πολλές πηγές, από τις οποίες γίνεται εξόρυξη δεδομένων, υπάρχουν δεδομένα τα οποία αναφέρονται σε σαρκασμό. Η εύρεση αυτών ονομάζεται sarcasm detection και απαιτεί βαθύτερη κατανόηση της φυσικής γλώσσας [12]. Ο σαρκασμός μπορεί να μετατρέψει πολύ εύκολα μια θετική κριτική της οποίας ο συγγραφέας ήθελε να σαρκάσει κάποιο γεγονός, σε μια αρνητική, χωρίς αυτό να αποδίδεται σωστά σαν αποτέλεσμα εξόρυξης συναισθήματος. Αυτός είναι και ο λόγος που η εύρεση τέτοιων σχολίων είναι ιδιαιτέρως σημαντική. Για να εντοπιστούν χρησιμοποιούνται μέθοδοι βαθιάς μάθησης (deep learning) όπως τα νευρωνικά δίκτυα. Για την εκπαίδευση των αλγορίθμων χρησιμοποιούνται κείμενα, που περιέχουν σαρκαστικές προτάσεις μέσα τους και με βάση αυτά προπονούνται και είναι έτοιμα να κάνουν εύρεση σαρκασμού σε νέα κείμενα.

3.3 Προβλήματα ανάλυσης συναισθημάτων

Εκτός από αυτά, που αναφέρθηκαν προηγουμένως, υπάρχουν κάποια ακόμα προβλήματα, τα οποία χρήζουν αντιμετώπισης. Το πρώτο που μελετάται είναι αυτό της εξόρυξης της οντότητας και του χρόνου που έχει γραφτεί το κείμενο. Η μελέτη της εξόρυξης των οντοτήτων έχει μελετηθεί παραπάνω. Για την εξόρυξη της χρονικής στιγμής που έχει γραφτεί κάθε κείμενο υπάρχουν μέθοδοι στα περισσότερα μέσα κοινωνικής δικτύωσης και θα πρέπει να χρησιμοποιούνται. Αυτό θα πρέπει να γίνεται, επειδή η χρονική στιγμή που γράφεται ένα κείμενο παίζει καθοριστικό ρόλο στην ανάλυση των συμπερασμάτων που προκύπτουν από αυτό. Σαν πρόβλημα αντιμετωπίζεται εύκολα καθώς πλέον πολλά μέσα παρέχουν εργαλεία που σχετίζονται με το χρόνο.

Το δεύτερο πρόβλημα, που συναντάται συχνά, αφορά την ομαδοποίηση των εκφράσεων που σχετίζονται με τις ίδιες πτυχές [2]. Σε αυτήν την κατηγορία ανήκουν λέξεις διαφορετικά εκφρασμένες, οι οποίες όμως σημαίνουν ακριβώς το ίδιο πράγμα. Ένα τέτοιο παράδειγμα είναι π.χ. ο όρος 'φωτογραφία' και ο όρος 'εικόνα'. Είναι πολύ σημαντική η αναγνώριση και των δύο ως όμοια αντικείμενα για κάθε μοντέλο που αναπτύσσεται. Για να λυθεί χρησιμοποιούνται μετρήσεις ομοιότητας των συμβολοσειρών καθώς και των συνωνύμων. Τα πειράματα που γίνονται για την επίλυση αυτών των θεμάτων είναι αρκετά ελπιδοφόρα και φαίνεται πως δουλεύουν σε αρκετά ικανοποιητικό βαθμό.

Παρόμοιο είναι και το πρόβλημα των εκφράσεων που δεν μπορούν να καταταχθούν θετικά ή αρνητικά με σιγουριά, καθώς δεν εκφράζουν πάντα το ίδιο συναίσθημα. Το επίθετο 'όμορφη' για παράδειγμα είναι θετικό. Δεν είναι όλα τα επίθετα έτσι όμως. Το επίθετο 'μικρό' για παράδειγμα μπορεί να είναι θετικό για μια αποθηκευτική μονάδα USB αλλά αρνητικό για μια οθόνη κινητού. Επίσης πολλά επίθετα δεν αναφέρονται στο ίδιο γεγονός κάθε φορά. Το επίθετο 'heavy' για παράδειγμα αναφέρεται στο βάρος αλλά και στην κυκλοφορία ως 'heavy traffic'. Είναι σημαντικό να αντιστοιχίζονται σωστά οι εκφράσεις με τα νοήματα. Για να επιτευχθεί αυτό χρησιμοποιούνται λίστες με συνδυασμούς αυτών των επιθέτων και γίνεται εκπαίδευση με τη χρήση αυτών.

Η ομαδοποίηση εκφράσεων με τις ίδιες αντιστοιχήσεις είναι επίσης πρόβλημα παρόμοιο με τα προηγούμενα. Σε αυτό το πρόβλημα όμως δίνεται έμφαση σε μέρη του λόγου όπως οι αντωνυμίες. Ονομάζεται σαν πρόβλημα coreference resolution. Για παράδειγμα υπάρχει η πρόταση 'Η Χ κάμερα είναι πολύ καλύτερη από την Υ κάμερα. Είναι φθηνότερη κιόλας'. Η δεύτερη πρόταση μπορεί να επιφέρει συναίσθημα καθώς αναφέρεται και αυτή θετικά στην κάμερα Χ. Πρέπει επομένως να εντοπισθεί και να ληφθεί στα αποτελέσματα. Για την επίλυση αυτού του προβλήματος χρησιμοποιούνται παρόμοιες τεχνικές με τα προηγούμενα προβλήματα, όπου υπάρχουν λίστες συνδυασμών με τις οποίες τα μοντέλα εκπαιδεύονται.

Τελευταίο, και ίσως από τα σημαντικότερα προβλήματα που θα αναλυθούν, είναι αυτό που ονομάζεται cross lingual opinion mining και αναφέρεται στην ανάλυση συναισθημάτων από πηγές που συνδυάζουν διαφορετικές γλώσσες. Είναι ένα πρόβλημα, το οποίο αν και λύνεται σε πολλές περιπτώσεις, εξακολουθεί να υπάρχει σε αρκετά μεγάλο βαθμό. Οι περισσότερες εργασίες που αφορούν ανάλυση συναισθημάτων γίνονται στα αγγλικά, καθώς εκεί υπάρχουν τα μεγαλύτερα λεξιλόγια για εκπαίδευση των συστημάτων. Γλώσσες όπως τα κινέζικα, που έχουν πολύ μεγάλο πλήθος συμβόλων αντιμετωπίζουν πολλά προβλήματα σε αυτόν τον τομέα. Η λύση είναι η έρευνα και η ανάπτυξη λεξιλογίων σε κάθε γλώσσα. Μέχρι τότε χρησιμοποιείται μετάφραση που γίνεται από υπολογιστές. Είτε μεταφράζονται τα λεξιλόγια και προσαρμόζονται σε κάθε γλώσσα, είτε μεταφράζονται οι πηγές στα αγγλικά και από εκεί γίνεται η επεξεργασία. Είναι ένα πρόβλημα ιδιαίτερα σημαντικό για πολλές επιχειρήσεις, οι οποίες θέλουν να συνδυάζουν τις διαφορετικές γνώμες των πελατών τους σε πολλές διαφορετικές χώρες ταυτόχρονα.

3.4 Πρακτικές προσεγγίσεις επίλυσης

Για να προσεγγίσει κανείς το πρόβλημα της ανάλυσης συναισθημάτων υπάρχουν τρεις διαφορετικές μέθοδοι, που μπορεί να ακολουθήσει, προκειμένου να επιτύχει θετικά αποτελέσματα. Η πρώτη μέθοδος είναι αυτή των αλγορίθμων της μηχανικής μάθησης. Η δεύτερη μέθοδος είναι η χρήση έτοιμων λεξικών, ενώ υπάρχουν και υβριδικές προσεγγίσεις, οι οποίες συνδυάζουν τις δύο προηγούμενες μεθόδους. Παρακάτω θα αναλυθούν εκτενέστερα οι σχετικές μέθοδοι.

3.4.1 Μέθοδοι Μηχανικής Μάθησης

Μηχανική Μάθηση είναι ένας επιστημονικός τομέας που αναπτύσσει αλγορίθμους, ώστε να χρησιμοποιηθούν σε σύνολα δεδομένων, με κύριο αντικειμενικό σκοπό την πρόβλεψη (παλινδρόμηση), ταξινόμηση και ομαδοποίηση. Αυτές οι διεργασίες διακρίνονται σε δύο κύριες κατηγορίες: την υπό επίβλεψη (supervised) και την χωρίς επίβλεψη (unsupervised) Μηχανική Μάθηση. Στην επιβλεπόμενη μάθηση θεωρείται ότι υπάρχει ένα πεπερασμένο σύνολο κλάσεων, στις οποίες το έγγραφο μπορεί να ενταχθεί. Στην απλούστερη έκδοση υπάρχουν μόνο δύο κλάσεις, η θετική και η αρνητική. Μια απλή επέκταση θα μπορούσε να περιέχει και μια ουδέτερη κλάση ή να περιέχει μια διακριτή αριθμητική κλίμακα (π.χ. από το 1 έως το 5) στην οποία θα πρέπει να ενταχθεί το έγγραφο. Για κάθε μία από τις πιθανές κλάσεις του εγγράφου υπάρχουν και τα ανάλογα δεδομένα εκπαίδευσης.

Με βάση αυτά τα δεδομένα το σύστημα μαθαίνει ένα μοντέλο ταξινόμησης, με χρήση ενός από τους συνηθισμένους αλγορίθμους ταξινόμησης, όπως οι SVM, NB, Logistic Regression, κ.α., οι οποίοι θα αναλυθούν παρακάτω. Στη συνέχεια το μοντέλο χρησιμοποιείται για την ταξινόμηση νέων εγγράφων. Όταν το έγγραφο ταξινομείται σε μια αριθμητική κλίμακα πεπερασμένου εύρους, τότε μπορεί να χρησιμοποιηθεί παρεμβολή (Regression). Οι παραπάνω μέθοδοι παρουσιάζουν καλά αποτελέσματα ακρίβειας (Accuracy) ακόμα και στις περιπτώσεις που τα έγγραφα αναπαρίστανται ως απλοί "σάκοι λέξεων" (Bag of Words). Παράλληλα μπορούν να χρησιμοποιηθούν πιο αναβαθμισμένοι τρόποι αναπαράστασης, όπως TF-IDF, POS σήμανση, sentiment lexicons, αλλά και όλες οι μέθοδοι που αναφέρθηκαν στα προηγούμενα κεφάλαια, οι οποίες αυξάνουν την απόδοση της ταξινόμησης.

Οι προσεγγίσεις μη επιβλεπόμενης μηχανικής μάθησης για την ανάλυση συναισθημάτων των εγγράφων, βασίζονται στον εντοπισμό και προσδιορισμό του σημασιολογικού προσανατολισμού (semantic orientation) συγκεκριμένων φράσεων μέσα στο έγγραφο. Ο μέσος προσανατολισμός των φράσεων του εγγράφου (σε σχέση με κάποιο προκαθορισμένο όριο) καθορίζει τελικά και την ταξινόμηση του συνολικού εγγράφου ως θετικό ή αρνητικό.

Υπάρχουν δύο κύριες προσεγγίσεις για την επιλογή των ελεγχόμενων φράσεων: Η πρώτη είναι να επιλέγονται βάσει κάποιων προκαθορισμένων POS μοτίβων (POS patterns), ενώ η δεύτερη αφορά τη χρήση λεξικών που περιέχουν συναισθηματικά κατηγοριοποιημένες λέξεις και φράσεις.

Ο κλασικός τρόπος προσδιορισμού του σημασιολογικού προσανατολισμού μιας φράσης P είναι ο υπολογισμός της διαφοράς του PMI (Pointwise Mutual Information) ανάμεσα στην φράση και σε κάποιες επιλεγμένες λέξεις W, που ανήκουν συναισθηματικά (sentiment words) στις χρησιμοποιούμενες κλάσεις. Π.χ. αν αναφερόμαστε σε κριτική ταινιών η λέξη ‘υπέροχη’ θα μπορούσε να θεωρηθεί ότι αντιστοιχεί στη θετική κλάση ενώ η λέξη ‘απαίσια’ ή ‘μέτρια’ θα μπορούσε να θεωρηθεί ότι αντιστοιχεί στην αρνητική κλάση. Ο όρος PMI(P,W) υπολογίζει τη στατιστική εξάρτηση ανάμεσα στη φράση P και στις λέξεις W, εξετάζοντας ουσιαστικά της συχνότητα συν εμφάνισής τους σε ένα δεδομένο corpus ή και στο WEB (με χρήση κάποιας μηχανής αναζήτησης). Ο σημασιολογικός προσανατολισμός δείχνει με ποια λέξη (και την αντίστοιχη κλάση) συγγενεύει σημασιολογικά περισσότερο η εξεταζόμενη φράση P. Τα παραπάνω χρησιμοποιούμενα εργαλεία είναι ιδιαίτερα αναπτυγμένα κυρίως για την Αγγλική γλώσσα. Για το σκοπό αυτό πολλοί ερευνητές, προκειμένου να αναλύσουν έγγραφα σε διαφορετική γλώσσα, αρχικά τα μεταφράζουν (ίσως και με χρήση αυτόματων WEB μεταφραστών) και στη συνέχεια ακολουθούν τη διαδικασία ανάλυσης που ακολουθείται για τα έγγραφα στην Αγγλική γλώσσα. Παρακάτω θα αναλυθούν οι τρεις αλγόριθμοι που χρησιμοποιούνται ως επί το πλείστον καθώς και μια καινούργια τάση.

Στον συγκεκριμένο τομέα φαίνεται να λειτουργούν άψογα τα Support Vector Machines (SVM), τα οποία αποδίδουν πολύ καλά κάθε φορά που η είσοδος τους αφορά κείμενα. Ανήκουν σαν μοντέλα στην επιβλεπόμενη μάθηση και χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση [5]. Στην περίπτωση των κειμένων χρησιμοποιούνται για να ταξινομήσουν τις επιμέρους κριτικές σε θετικές, αρνητικές και ουδέτερες. Αφού δοθεί στον αλγόριθμο ένα σύνολο εκπαιδευτικών παραδειγμάτων, που αφορούν και τις τρεις προαναφερθείσες κλάσεις ο SVM ξεκινάει να δημιουργεί ένα μοντέλο, που εκχωρεί νέα παραδείγματα σε κάθε κατηγορία. Το βασικό χαρακτηριστικό του είναι η αναπαράσταση των παραδειγμάτων ως σημεία χαρτογραφημένα έτσι ώστε τα παραδείγματα των διαφορετικών κατηγοριών να χωρίζονται από ένα σαφές κενό, το οποίο πρέπει να είναι όσο το δυνατόν ευρύτερο. Στη συνέχεια τα νέα παραδείγματα χαρτογραφούνται επίσης στο χώρο και προβλέπεται σε ποια πλευρά ανήκουν.

Τα πλεονεκτήματα των SVM είναι πως μεταχειρίζονται πολύπλοκα μη γραμμικά προβλήματα, με ιδιαίτερα καλή απόδοση στα κείμενα. Χρησιμοποιούν απλούς γραμμικούς αλγόριθμους, είναι ανθεκτικοί στην υπερμοντελοποίηση, έχοντας παράλληλα χαμηλό κόστος. Στα αρνητικά τους βρίσκεται το γεγονός πως έχουν μεγάλες απαιτήσεις μνήμης και δεν είναι ερμηνεύσιμα μοντέλα. Επίσης δεν υπάρχει μεθοδολογία για την επιλογή της συνάρτησης και των παραμέτρων του πυρήνα. Τέλος, σε περιπτώσεις ταξινόμησης σε πολλαπλές κλάσεις το πρόβλημα διατυπώνεται σαν συνδυασμός δυαδικών προβλημάτων, γεγονός που κάποιες φορές προκαλεί δυσκολίες στον υποψήφιο ερευνητή.

Ένας ακόμη αλγόριθμος που αποδίδει σε πολλές έρευνες του συγκεκριμένου τομέα είναι ο Naïve Bayes. Ανήκει και αυτός στην κατηγορία των αλγορίθμων μηχανικής μάθησης υπό επίβλεψη. Είναι ένας αλγόριθμος, ο οποίος βασίζεται στο θεώρημα του Bayes, κάνοντας ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Μελετήθηκε εκτενώς τη δεκαετία του 1960 και παραμένει μέχρι σήμερα μια βασική μέθοδος, η οποία χρησιμοποιείται για την κατηγοριοποίηση κειμένων. Χρειάζεται κατάλληλη προεπεξεργασία για να καταφέρει να είναι ανταγωνιστικός, σχετικά με τους άλλους αλγορίθμους. Ο ταξινομητής που χρησιμοποιεί μαθαίνει από ένα σύνολο δεδομένων εκπαίδευσης, τα οποία περιλαμβάνουν τα κείμενα, από τα οποία θέλουμε να εξάγουμε τα συναισθήματα και δίνει στο κάθε νέο αντικείμενο που εισέρχεται για να ταξινομηθεί, μια πιθανότητα να ανήκει στην κάθε κατηγορία [11]. Οι πιθανότητας αυτές προκύπτουν βάση του θεωρήματος Bayes το οποίο παρατίθεται παρακάτω.

$$P(h|\mathrm{D}) = \frac{P(D|\mathrm{h})P(h)}{\mathrm{P(D)}}$$

Εικόνα 3. Παρουσίαση του θεωρήματος Bayes

Με το σύμβολο h αναφέρεται σε ένα χώρο υποθέσεων και με το σύμβολο D στα δεδομένα, τα οποία έχουν παρατηρηθεί. Μια τέτοια υπόθεση ονομάζεται μέγιστη υπό συνθήκη υπόθεση. Ο όρος P(h|D) ονομάζεται πιθανοφάνεια των δεδομένων D με βάση

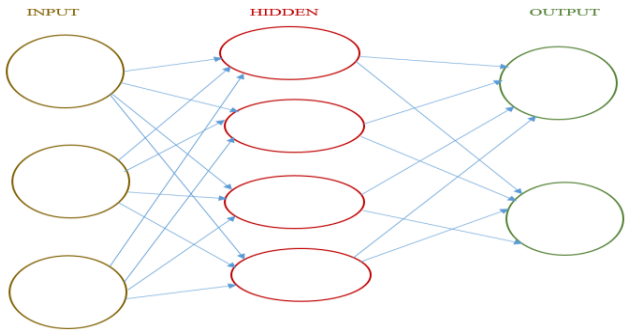
την υπόθεση h και κάθε υπόθεση που μεγιστοποιεί τον όρο αυτό ονομάζεται υπόθεση μέγιστης πιθανοφάνειας. Στα μειονεκτήματα του αλγορίθμου βρίσκεται η υψηλή πολυπλοκότητα, που παρουσιάζει καθώς είναι γραμμική σχετικά με το αριθμό των υποψηφίων υποθέσεων με κάποιες ειδικές περιπτώσεις, στις οποίες το κόστος μειώνεται.

Ο τρίτος αλγόριθμος που μελετάται είναι αυτός του Logistic Regression. Πρόκειται για ένα στατιστικό μοντέλο που στη βασική του μορφή χρησιμοποιεί μια λογική λειτουργία προκειμένου να καταφέρει να μοντελοποιήσει μια δυαδική εξαρτώμενη μεταβλητή, παρά το γεγονός πως υπάρχουν πολύ περισσότερες και πολύπλοκες εκτάσεις. Πολλές φορές χρησιμοποιεί μια συνάρτηση, η οποία ονομάζεται σιγμοειδής και είναι αυτή που καθορίζει τα αποτελέσματα. Η κάθε ανεξάρτητη μεταβλητή X1 συμμετέχει στο αποτέλεσμα του αλγορίθμου πολλαπλασιαζόμενη με έναν ειδικό συντελεστή A1. Αυτό σημαίνει ότι η μεταβολή κατά μια μονάδα της X1 μεταβάλει το αποτέλεσμα κατά A1. Όταν η έξοδος αφορά δυαδική μεταβλητή η τιμή της αντιστοιχεί στην πιθανότητα της καθεμίας κλάσης. Είναι ένα μοντέλο που χρησιμοποιείται αρκετά από τους ερευνητές καθώς τα χαρακτηριστικά του ταιριάζουν σε αρκετές έρευνες.

Μερικά από τα πλεονεκτήματα του είναι πως αποδίδει πολύ καλά εξαρτημένες μεταβλητές που είναι κατηγορικές. Οι ανεξάρτητες μπορούν να λάβουν οποιαδήποτε μορφή. Επίσης, μπορούν να υπολογιστούν εύκολα οι γραμμικοί συνδυασμοί των μεταβλητών, παρουσιάζοντας έτσι μια απλότητα. Από την άλλη πλευρά, στα μειονεκτήματα του είναι η δυσκολία στην παραμετροποίηση, προκειμένου να έρθει στη μορφή που χρειάζεται ο ερευνητής.

Η αλγόριθμοι που αφορούν τις τελευταίες τάσεις στο θέμα της ανάλυσης συναισθημάτων σχετίζονται γύρω από τη βαθιά μάθηση (deep learning). Είναι ένας τομέας που έχει έρθει στο προσκήνιο τα τελευταία χρόνια και χρησιμοποιείται ευρέως, παρότι προϋπήρχε εδώ και αρκετά χρόνια. Ένας σημαντικός λόγος είναι ότι επιτρέπει μάθηση με εποπτεία και χωρίς εποπτεία, γεγονός πολύ χρήσιμο για τους ερευνητές. Βασίζεται σε τεχνητά νευρωνικά δίκτυα, τα οποία αναλύονται παρακάτω.

Τα νευρωνικά δίκτυα είναι κυκλώματα διασυνδεδεμένων νευρώνων. Τα τεχνητά νευρωνικά δίκτυα είναι μαθηματικά μοντέλα, τα οποία είναι εμπνευσμένα από τα βιολογικά νευρικά μοντέλα, που χρησιμοποιούν νευρώνες. Οι νευρώνες αποτελούν δομή του ανθρώπινου εγκεφάλου. Αποτελούν έναν αποτελεσματικό τρόπο επίλυσης σύνθετων προβλημάτων. Η μέθοδος που χρησιμοποιούν, λύνει το πρόβλημα κομμάτι-κομμάτι και έχουν τη δυνατότητα έκφρασης μεγάλης ποικιλίας μη γραμμικών επιφανειών χρησιμοποιώντας έναν αριθμό προτύπων εισόδου και εξόδου, τα οποία επιλέγονται σε παγκόσμια κλίμακα. Χρησιμοποιούνται πολύ συχνά για εξόρυξη δεδομένων και για πρόβλεψη συγκεκριμένων μεταβλητών σε μελλοντική χρονική στιγμή. Ο σκοπός τους είναι να βρεθεί ένα μοτίβο ή μια σχέση μεταξύ φαινομενικά στοχαστικών ή άσχετων μεταξύ τους εισροών. Αποτελείται από απλούς υπολογιστικούς κόμβους οι οποίοι είναι διασυνδεδεμένοι μεταξύ τους και ονομάζονται νευρώνες. Το σύστημα από το οποίο έχει εμπνευστεί και προσπαθεί να προσομοιώσει είναι αυτό του Κεντρικού Νευρικού Συστήματος. Στο αρχικό επίπεδο βρίσκονται οι νευρώνες που ονομάζονται νευρώνες εισόδου και απλά αφορούν την είσοδο των στοιχείων που δίνονται στο δίκτυο, χωρίς να πραγματοποιούν κάποιον υπολογισμό. Στο τέλος κάθε νευρωνικού δικτύου, υπάρχουν οι νευρώνες εξόδου, οι οποίοι είναι υπεύθυνοι για την τελική έξοδο των μεταβλητών του δικτύου, δηλαδή παράγουν το τελικό αποτέλεσμα. Η τρίτη κατηγορία νευρώνων είναι αυτοί των υπολογιστικών νευρώνων ή κρυμμένων νευρώνων, η δουλεία των οποίων είναι να πολλαπλασιάζουν κάθε είσοδο που δέχονται με το αντίστοιχο συναπτικό βάρος και να υπολογίζουν ένα τελικό γινόμενο, που αποτελεί και την έξοδο του νευρώνα. Η τυπική μορφή ενός δικτύου παρουσιάζεται παρακάτω.



Εικόνα 4. Παρουσίαση της δομής ενός νευρωνικού δικτύου

Ο μαθηματικός τύπος που βρίσκεται πίσω από τους κρυμμένους νευρώνες και χρησιμοποιείται για τις πράξεις θα αναφέρεται παρακάτω. Στον τύπο αυτόν, με το δείκτη k δηλώνεται ο εκάστοτε νευρώνας. Με το γράμμα w το βάρος του νευρώνα. Τα γράμματα j, m δείχνουν τον j, m(οστό) νευρώνα αντίστοιχα, αποτελώντας δείκτες. Ως X_j, συμβολίζεται η είσοδος του j-οστού νευρώνα. Ο παρακάτω τύπος μας δείχνει πως η έξοδος του k νευρώνα, είναι ίση με το άθροισμα των γινομένων των εισόδων του νευρώνα επί των αντιστοίχων βαρών τους. Σαν τιμή εισόδου το κατώφλι έχει πάντα την μονάδα και εάν το συνολικό άθροισμα από τις υπόλοιπες εισόδους του νευρώνα είναι μεγαλύτερης κλίμακας από τη συγκεκριμένη τιμή, τότε ο νευρώνας ενεργοποιείται. Σε αντίθετη περίπτωση ο νευρώνας παραμένει ανενεργός. Προκειμένου να γνωρίζει ένας νευρώνας πότε θα παραμείνει ανενεργός ή πότε θα ενεργοποιηθεί, εφαρμόζεται μια ειδική συνάρτηση η οποία ονομάζεται βηματική συνάρτηση ενεργοποίησης. Υπάρχουν αρκετές συναρτήσεις, οι οποίες είναι υπεύθυνες για την ενεργοποίηση της τιμής του κατωφλιού.

$$W_k = \sum_{j}^{m} W_{kj} * X_j$$

Εικόνα 5. Παρουσίαση του τύπου που χρησιμοποιείται στα νευρωνικά δίκτυα

Στα πλεονεκτήματα που παρουσιάζουν τα νευρωνικά δίκτυα βρίσκεται η ευκολία και η αποδοτική μάθηση μέσω παραδειγμάτων, την οποία αυτά παρουσιάζουν. Υπάρχει μεγάλη ανοχή σε σφάλματα, γεγονός πολύ χρήσιμο για κάθε δοκιμή. Επίσης παρουσιάζουν εξαιρετικά αποτελέσματα στην αναγνώριση προτύπων. Στα μειονεκτήματα τους παρουσιάζεται η υψηλή πολυπλοκότητα τους σε συνδυασμό με των αυξημένο χρόνο εκτέλεσης τους. Ο χρόνος που χρειάζονται για να εκτελεσθούν δεν είναι αποτρεπτικός φυσικά, απλά είναι μεγαλύτερος από άλλους αλγορίθμους, οι οποίοι προαναφέρθηκαν και μπορούν να χρησιμοποιηθούν στην ανάλυση συναισθημάτων. Ένα ακόμη

μεγάλο αρνητικό χαρακτηριστικό των αλγορίθμων των νευρωνικών δικτύων είναι πως δεν παρουσιάζουν εύκολη ερμηνευσιμότητα των αποτελεσμάτων. Αυτό το γεγονός κοστίζει πολύ σε κάποιες περιπτώσεις. Στην περίπτωση της ανάλυσης συναισθημάτων δεν κοστίζει ιδιαίτερα καθώς τα μοντέλα προορίζονται για χρήση ερευνητών και εταιριών και επομένως μπορεί να θυσιαστεί η ερμηνευσιμότητα προκειμένου να επιτευχθούν καλύτερα αποτελέσματα.

Μια τεχνική που χρησιμοποιείται από τους σύγχρονους επιστήμονες είναι το Word2vec. Το Word2vec λοιπόν είναι μια ομάδα σχετικών μοντέλων, τα οποία χρησιμοποιούνται για την παραγωγή ενσωματώσεων (embeddings) των λέξεων. Αυτά τα μοντέλα είναι νευρωνικά δίκτυα δύο επιπέδων. Λαμβάνει ως είσοδο ένα μεγάλο κορμό κειμένου και παράγει ένα χώρο διανυσμάτων τα οποία συνήθως είναι μερικών εκατοντάδων διαστάσεων με κάθε μοναδική λέξη, να αντιστοιχεί στον αντίστοιχο διάνυσμα στο χώρο. Αφού οι λέξεις τοποθετηθούν στο χώρο είναι εύκολο να βρεθούν αυτές που βρίσκονται σε στενή εγγύτητα μεταξύ τους και άρα είναι παρόμοιες. Είναι μια τεχνική πολύ χρήσιμη με πολύ καλά αποτελέσματα στις έρευνες όπου χρησιμοποιείται.

3.4.2 Μετρικές Αξιολόγησης

Ένας ακόμα πολύ σημαντικός τομέας που πρέπει να αναλυθεί είναι αυτός των μετρικών. Πρέπει να μελετηθεί, με ποιόν τρόπο θα αξιολογηθούν τα αποτελέσματα, προκειμένου να βρεθεί ο καταλληλότερος αλγόριθμος για την κάθε περίπτωση. Παρακάτω θα αναλυθούν μερικές από τις μετρικές που χρησιμοποιούνται κατά κόρον σε θέματα μηχανικής μάθησης, που αφορούν την ανάλυση των συναισθημάτων. Μια κύρια μετρική που χρησιμοποιείται είναι αυτή της ακρίβειας (Accuracy) η οποία αναφέρεται στο λόγο του αριθμού των σωστών προβλέψεων προς το συνολικό αριθμό δειγμάτων εισόδου. Λειτουργεί καλά μόνο αν υπάρχει παραπλήσιος αριθμός δειγμάτων σε κάθε κατηγορία και όχι όταν υπάρχει ανισορροπία μεταξύ αυτών. Στη δεύτερη περίπτωση προτιμώνται άλλες μετρικές.

Μια ακόμη βασική μετρική που πρέπει να αναφερθεί είναι αυτή του Precision. Αυτή η μετρική μετρά τον αριθμό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό θετικών αποτελεσμάτων που προβλέπει ο ταξινομητής. Ο λόγος που την περιγράφει είναι τα σωστά θετικά αποτελέσματα (TP) προς τα σωστά θετικά συν τα λάθος θετικά (FP).

Παράλληλα υπάρχει μια ακόμη μετρική που αφορά την ανάκληση (Recall). Σε αυτή τη μετρική μετριούνται τα σωστά θετικά αποτελέσματα (TP) ως ποσοστό των συνολικών πραγματικά θετικών δειγμάτων (TP+FN).

Η μετρική στην οποία βασίζονται περισσότερο οι ερευνητές συνήθως είναι αυτή που ονομάζεται F1 Score. Η βαθμολογία F1 είναι μια αρμονική μέση τιμή μεταξύ ακρίβειας (Precision) και ανάκλησης (Recall). Συνδυάζει έτσι τις παραπάνω δύο μετρικές για να παράγει τα αποτελέσματά της. Μαθηματικά εκφράζεται με τον παρακάτω τύπο.

$$\mathrm{F1} = 2 * \frac{1}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}}$$

Εικόνα 6. Παρουσίαση του τύπου της μετρικής F1

3.4.3 Λεξικά Συναισθημάτων

Όπως αναφέρθηκε προηγουμένως, εκτός από τη μηχανική μάθηση, η οποία χρησιμοποιείται κατά κόρον, υπάρχει και η ανάλυση συναισθημάτων με τη χρήση λεξικών. Σε αυτές τις τεχνικές χρησιμοποιούνται λεξικά συναισθημάτων, τα οποία περιέχουν λέξεις, οι οποίες προσδίδουν ένα θετικό, αρνητικό ή ουδέτερο νόημα στις προτάσεις στις οποίες εμπλέκονται. Κάθε λέξη δέχεται ένα σκορ με βάση το αν ανήκει σε αυτά τα λεξικά. Το σκορ της κάθε λέξη δείχνει την ένταση των εκφράσεων των συναισθημάτων που αυτή προσδίδει. Πολλές φορές αυτά τα σκορ είναι κανονικοποιημένα και οι τιμές τους βρίσκονται μεταξύ μια κλίμακας ανάμεσα στο 0 και το 1. Επίσης μέσα στο κείμενο υπάρχουν συχνά επιρρήματα ή άλλες συναφής φράσεις, οι οποίες ενισχύουν ή αποδυναμώνουν την έννοια της λέξης στην οποία αναφέρονται. Για να συμπεριληφθούν και αυτές οι λέξεις στη συνολική βαθμολογία συναισθήματος χρησιμοποιείται ο παρακάτω τύπος [6]. Με το σύμβολο et_{ir} συμβολίζεται το συνολικό συναισθηματικό σκορ μιας λέξης, ενώ με το σύμβολο $intens_j$ συμβολίζεται η λέξη που επηρεάζει τη λέξη που μελετάται.

$$WScore(et_{ir}) = \| \left(1 + score\left(intens_j\right)\right) * \mathrm{score}(et_{ir}) \|$$

Εικόνα 7. Παρουσίαση του τύπου για την συμπερίληψη λέξεων που επηρεάζουν το νόημα

Αφού λοιπόν οι λέξεις πάρουν ένα βαθμό που τις χαρακτηρίζει γίνεται μέτρηση του συνολικού βαθμού του κειμένου και έπειτα αυτό χαρακτηρίζεται ως θετικό, αρνητικό ή ουδέτερο. Σε αυτές τις τεχνικές δεν υπάρχει χρήση τεχνικών μηχανικής μάθησης. Επειδή βασίζονται αποκλειστικά στα λεξικά, πρέπει πρώτα να φτιαχτούν σωστά λεξικά με αντιπροσωπευτικές λέξεις μέσα τους, οι οποίες θα συμπεριλαμβάνουν όλα τα συναισθήματα που μπορεί κανείς να συναντήσει σε ένα κείμενο. Είναι μια καλή μέθοδος, στην οποία βασίζονται και τα περισσότερα μοντέλα μηχανικής μάθησης προκειμένου να κατατάξουν τα κείμενα εκπαίδευσής τους σε επιμέρους κατηγορίες. Βοήθησε πολύ στην εξέλιξη του τομέα της ανάλυσης συναισθημάτων δίνοντας μια γερή εκκίνηση στις ανάλογες εφαρμογές. Υπάρχουν αρκετά προγράμματα διαθέσιμα στο διαδίκτυο, τα οποία βασίζονται σε αυτήν την τεχνική και παράγουν αποτελέσματα χωρίς να χρησιμοποιούν πολύπλοκους αλγόριθμους από πίσω. Είναι αρκετά γρήγορα προγράμματα και παρέχουν αρκετά αξιόπιστα αποτελέσματα, χωρίς βέβαια να είναι πάντα τα καλύτερα δυνατά από άποψη ακρίβειας και αξιοπιστίας.

3.4.4 Υβριδικές μέθοδοι

Η τρίτη κατηγορία είναι αυτή της υβριδικής προσέγγισης, η οποία χρησιμοποιείται κατά κόρον από ερευνητές και φαίνεται να αποδίδει εξαιρετικά καλά αποτελέσματα. Αφού λοιπόν βρεθούν οι καταλληλότεροι αλγόριθμοι μηχανικής μάθησης, που σχετίζονται με το κάθε πρόβλημα ξεχωριστά ξεκινάει η διαδικασία εξαγωγής συναισθημάτων με βάση τα λεξικά. Είναι μια τεχνική, η οποία συναντάται σε πολύ μεγάλο πλήθος ερευνών. Υπάρχουν παραδείγματα [6], στα οποία φαίνεται πως αυτή η τεχνική υπερτερεί έναντι των άλλων και προσδίδει εξαιρετικά αποτελέσματα στην ανάλυση συναισθημάτων. Τα λεξικά χρησιμοποιούνται λοιπόν για την εκπαίδευση των μοντέλων της μηχανικής μάθησης και έτσι αυτοί μπορούν με περισσότερη ακρίβεια και ευκολία να κατατάξουν τα κείμενα στις κατηγορίες συναισθημάτων, που τους ζητούνται.

Φυσικά για να λειτουργήσουν σωστά οι αλγόριθμοι της μηχανικής μάθησης, πρέπει ο αριθμός κειμένων για κάθε συναίσθημα να είναι επαρκής ώστε να οδηγήσει σε αξιόπιστα συμπεράσματα. Σε κάθε άλλη περίπτωση δεν θα αποδίδει με

επιτυχία το μοντέλο. Είναι πολύ σημαντικό κάθε συναίσθημα να περιέχει όμοια παραδείγματα σε πλήθος με τα άλλα συναισθήματα. Αν, για παράδειγμα, οι κλάσεις του φόβου παρουσιάζουν μεγαλύτερο πλήθος από τις άλλες κλάσεις, είναι φυσιολογικό να έχει ένα κείμενο περισσότερες πιθανότητες να καταταχθεί εκεί, ακόμη και αν δεν αφορά συναισθήματα φόβου. Ένα καθορισμένο σύνολο αντιπροσωπευτικών λέξεων για κάθε συναίσθημα μπορεί να επηρεάσει σημαντικά τη διαδικασία ανίχνευσης συναισθημάτων λοιπόν. Γι' αυτό το λόγο θα πρέπει να εφαρμόζονται πολύ προσεκτικά, συγκεκριμένες ενέργειες προεπεξεργασίας των λεξικών και των δεδομένων προκειμένου να αποφεύγονται τέτοια προβλήματα.

Για να επιτευχθεί με μεγάλα ποσοστά επιτυχίας ο παραπάνω στόχος πρέπει, εκτός από την προεπεξεργασία, να γίνει και σωστός σχηματισμός των λεξικών. Πρέπει να προσέξει ένας ερευνητής να μη παίρνει δεδομένα μόνο από χρήστες που συμπαθούν μια εταιρεία, καθώς τα δεδομένα του, όπως είναι λογικό, θα είναι ως επί το πλείστον θετικά. Το ίδιο θα συμβεί και εάν ο χρήστης λάβει ως δείγμα εταιρείες που έχουν αρνητική προβολή στην αγορά. Τότε τα αρνητικά συναισθήματα θα υπερτερήσουν. Γι' αυτό το λόγο είναι καλό να υπάρχει μεγάλη κλίμακα, χρονική, τοπική και εύρους συναισθημάτων ώστε να φτιαχτούν σωστά λεξικά. Η χρήση πλατφορμών crowdsourcing (π.χ. Crowd-Flower) βοηθάει ιδιαίτερα στη δημιουργία ενός πιο σωστού συνόλου δεδομένων.

4. Συμπεράσματα

Ο τομέας της ανάλυσης των συναισθημάτων είναι συνεχώς αναπτυσσόμενος, παρουσιάζοντας μεγάλο ερευνητικό ενδιαφέρον την τελευταία δεκαετία, τόσο από επιχειρήσεις όσο και από ερευνητές. Κάθε επιχείρηση που νοιάζεται για την ανατροφοδότηση των αποτελεσμάτων της, πρέπει αν χρησιμοποιεί ένα τέτοιο σύστημα προκειμένου να γνωρίζει τους τομείς που πρέπει να βελτιώσει. Έχουν αναπτυχθεί και εξελιχθεί εξαιρετικά εργαλεία επί του θέματος και συνεχώς αναπτύσσονται νέα. Τα λεξικά, τα οποία χρησιμοποιούνται αναπτύσσονται και αυτά με αρκετά καλούς ρυθμούς, παρουσιάζοντας όμως κάποιες ελλείψεις στη χρήση τους.

Μια από τις σημαντικότερες ελλείψεις στο χώρο είναι η ύπαρξη ποιοτικών λεξικών συναισθημάτων κατά κύριο λόγω μόνο στην αγγλική γλώσσα. Αυτό το γεγονός οδηγεί ερευνητές από άλλες χώρες να μεταφράσουν τα κείμενα που αποτελούν πηγή εισόδου των αλγορίθμων τους, χάνοντας έτσι πολύτιμο χρόνο, αλλά και πολλές φορές πολύτιμες πληροφορίες, καθώς υπάρχει αλλοίωση του περιεχομένου των κειμένων από τους αυτόματους μεταφραστές, που χρησιμοποιούνται. Επομένως η ανάπτυξη σωστών λεξικών σε περισσότερες γλώσσες είναι ένας τομέας που πρέπει να αποτελέσει μελλοντικό θέμα μελέτης.

Ένα ακόμη αρνητικό πρόβλημα που παρουσιάζουν τα συστήματα ανάλυσης συναισθημάτων είναι πως δεν παρουσιάζουν μεγάλη ικανότητα στην αναγνώριση κάποιον ειρωνικών πηγών κειμένου ή κάποιων λογοπαιγνίων από τους συγγραφείς τους. Το ίδιο συμβαίνει και με τον σαρκασμό. Πρέπει να δοθεί ανάλογη προσοχή από τους ερευνητές σε αυτό το κομμάτι, καθώς είναι ο λόγος που μένουν σε όχι και τόσο υψηλά ποσοστά επιτυχίας κάποια από τα τελικά ποσοστά των ερευνών που γίνονται. Πρέπει να αναπτυχθούν τεχνικές και συστήματα, τα οποία θα προσπαθούν να αναγνωρίσουν πρότυπα σαν τα προαναφερθέντα προκειμένου να καταφέρουν να κατηγοριοποιήσουν τα κείμενα με ακόμη περισσότερη επιτυχία.

Στα θετικά των συστημάτων ανάλυσης συναισθημάτων βρίσκονται κάποια καλά παραδείγματα ερευνών, τα οποία φαίνεται να έχουν βοηθήσει ιδιαίτερα κάποιες επιχειρήσεις αναφορικά με τους πελάτες τους, καθώς και κάποιους ερευνητές. Είναι σημαντικό να χρησιμοποιούνται τεχνικές crowdsourcing για τη δημιουργία ισορροπημένων λεξικών συναισθημάτων, τα οποία φαίνονται να αποτελούν τη βασική δομή για να πετύχει οποιοδήποτε ερευνητικό μοντέλο.

Τέλος, από πλευράς των τεχνικών που μελετήθηκαν είναι καλό να εφαρμόζονται οι περισσότερες από τις τεχνικές προεπεξεργασίας που προαναφέρθηκαν καθώς φαίνονται να επιδρούν θετικά τις περισσότερες φορές. Οι υβριδικές προσεγγίσεις φαίνεται πως είναι το μέλλον στο συγκεκριμένο τομέα και θα απασχολήσουν τους επιστήμονες τα επόμενα χρόνια.

5. REFERENCES

[1] A. Moreo, M. Romero, J.L. Castro, J.M. Zurita. 2012. *Lexicon-based Comments-oriented News Sentiment Analyzer system*. Dept. of Computer Science and Artificial Intelligence, University of Granada, Spain.

[2] Bing Liu, Lei Zhang. 2011. *A survey of opinion mining and sentiment analysis*. University of Illinois at Chicago

[3] Bing Liu. 2009. *Sentiment Analysis and Subjectivity*. University of Illinois at Chicago.

[4] Caitlin Dreisbach, Theresa A. Koleck, Philip E. 2019. *A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data*. International Journal of Medical Informatics.

[5] Dennys C.A. Mallqui, Ricardo A.S. Fernandes. 2019. *Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques.* Applied Soft Computing Journal 75.

[6] Despoina Chatzakou, Athena Vakali, Konstantinos Kafetsios. 2017. *Detecting variation of emotions in online activities.* Expert Systems With Applications 89 318–332.

[7] Guillermo R. Simari, Eduardo Ferme, Flabio Gutierrez Segura, Jose Antonio Rodriguez Melquiades. 2018. *Advantages in Artificial Intelligence*. 16th Ibero-American Conference on AI.

[8] Jun Feng, Cheng Gong, Xiaodong Li, Raymond Y. K. Lau. 2018. *Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews*. Wireless Communications and Mobile Computing.

[9] Kumar Ravi, Vadlamani Ravi. 2015. *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*. Knowledge-Based Systems.

[10] Nir Friedman, Dan Geiger, Moises Goldszmidt. 1997. *Bayesian Network Classifiers.* Machine Learning, 29, 131–163.

[11] Ronen Feldman. 2013. *Techniques and Applications for Sentiment Analysis.* Communications of the ACM Vol. 56.

[12] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Prateek Vij. 2017. *A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks.* Nanyang Technological University.

[13] Walaa Medhat, Ahmed Hassan, Hoda Korashy. 2014. *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal.