Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής και Τηλεματικής

Καταγραφή επιλογών στη χάραξη πολιτικών με τη χρήση κοινωνικών δικτύων

Διπλωματική Εργασία

του

Ψωμακέλη Ευάγγελου

Α.Μ. 12116

Επιβλέπων : Κωνσταντίνος Τσερπές,
Λέκτορας

Αθήνα, Μάρτιος 2014

Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής και Τηλεματικής

Καταγραφή επιλογών στη χάραξη πολιτικών με τη χρήση κοινωνικών δικτύων

Διπλωματική Εργασία

του

Ψωμακέλη Ευάγγελου

Α.Μ. 12116

Επιβλέπων : Κωνσταντίνος Τσερπές,
Λέκτορας

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28η Μαρτίου 2014.

.....	.....	.....
Κωνσταντίνος Τσερπές, Λέκτορας	Αναγνωστόπουλος Δημοσθένης, Καθηγητής	Βαρβαρίγου Θεοδώρα, Καθηγήτρια

Αθήνα, Μάρτιος 2014

Περιεχόμενα

Περίληψη

Η παρούσα εργασία ασχολείται με τον τομέα του Sentiment Analysis, σε μία προσπάθεια να αναδείξει έναν ή περισσότερους συνδυασμούς μεθόδων επεξεργασίας φυσικού λόγου και αλγορίθμων μηχανικής μάθησης, οι οποίοι θα προσφέρουν δυνατότητες κατηγοριοποίησης διαφόρων πηγών απόψεων για ένα συγκεκριμένο θέμα. Στην παρούσα εργασία θα αναλυθούν οι μέθοδοι επεξεργασίας φυσικού λόγου Bag of Words, N-Grams και N-Gram Graphs, και με τη βοήθεια της ανοιχτής βιβλιοθήκης της Weka, οι αλγόριθμοι μηχανικής μάθησης C4.5, Naïve Bayesian Networks, Support Vector Machines, Logistic Regressionm Multilayer Perceptrons, Best-First Trees και Functional Trees. Σκοπός είναι να ανακαλυφθεί ένας συνδυασμός με αρκετά υψηλή απόδοση, έτσι ώστε να μπορέσει να παραχθεί ένα αξιόπιστο, σε βάθος χρόνου μοντέλο πρόβλεψης της κοινής γνώμης απέναντι σε συγκεκριμένα θέματα. Αυτά τα θέματα στην παρούσα έρευνα θα είναι παράμετροι κάποιας πολιτικής αλλά πρακτικά δεν υπάρχει κάποια σύνδεση μεταξύ του θέματος προς ανάλυση με τον αλγόριθμο που τον αναλύει. Η μόνη τους σύνδεση είναι μέσω των δεδομένων τα οποία θα αναλύει ο αλγόριθμος και θα επιλέγονται με βάση το θέμα που προσπαθούμε να αναλύσουμε.

1 Εισαγωγή

Κατά τη χάραξη μίας νέας πολιτικής, οι υπεύθυνοι για αυτήν πρέπει να προβλέψουν με όσο μεγαλύτερη ακρίβεια μπορούν τις επιπτώσεις της στο κοινωνικό σύνολο. Αν αυτή η διαδικασία δεν καταφέρει να ολοκληρωθεί επιτυχώς, τα αποτελέσματα μίας πολιτικής μπορεί να έχει τελείως διαφορετικά, και πολύ συχνά καταστροφικά, αποτελέσματα σε σχέση με τους στόχους που προσπαθούσε να επιτύχει. Σε κάθε περίπτωση, μία εκ-προτέρων γνώση (apriori) είναι απαραίτητη για την επιλογή της σωστής πολιτικής.

Για να γίνει αυτή η πρόβλεψη όμως, πρέπει να ληφθούν υπόψιν όλες οι παράμετροι που μπορούν να επηρεάσουν το αποτέλεσμα της πολιτικής. Φυσικά αυτό δεν είναι εφικτό με τη σημερινή τεχνολογία και γνώση, καθώς οι παράμετροι σε ένα ρεαλιστικό περιβάλλον είναι πάρα πολλές και συχνά μεταβάλλονται απρόβλεπτα. Για αυτόν τον λόγο οι υπεύθυνοι για τις πολιτικές προσπαθούν να συμπεριλάβουν στην πρόβλεψή τους όσο περισσότερες παραμέτρους μπορούν έτσι ώστε να λάβουν μία ακριβέστερη πρόγνωση των αποτελεσμάτων της κάθε πολιτικής.

Μία από τις σημαντικότερες και μέχρι τώρα αρκετά απρόβλεπτες παραμέτρους αποτελεί η αποδοχή από τον πληθυσμό (public acceptability). Όταν εφαρμόζεται μία νέα πολιτική επηρεάζει άμεσα ή έμμεσα κάποιο κομμάτι του πληθυσμού, ο οποίος μπορεί να είναι θετικός, αρνητικός ή και ουδέτερος απέναντι στις αλλαγές που προσπαθεί να του επιβάλει η πολιτική. Αυτή η στάση του πληθυσμού έχει μεγάλες επιπτώσεις στα αποτελέσματα της κάθε πολιτικής δεδομένου ότι τις περισσότερες φορές ο πληθυσμός μεταβάλει αρκετές από τις βασικές παραμέτρους που χρησιμοποιούνται για την πρόβλεψη των αποτελεσμάτων της.

Μέχρι και την προηγούμενη δεκαετία ο μόνος τρόπος για να λάβουμε μία εκτίμηση της κοινής γνώμης για κάποιο θέμα ήταν τα ερωτηματολόγια και οι συνεντεύξεις. Με τα ερωτηματολόγια δεν ήταν εφικτό να λάβουμε την τρέχουσα άποψη δεδομένου ότι υπήρχε μία καθυστέρηση μεταξύ της απάντησής τους και της επεξεργασίας των αποτελεσμάτων, ενώ με τις συνεντεύξεις δεν ήταν ποτέ εφικτό να αναλυθεί ένα αρκετά μεγάλο δείγμα. Σήμερα έχουμε το Internet και το τεράστιο πλήθος απόψεων που οι χρήστες ανεβάζουν καθημερινά, δίνοντάς μας ελεύθερα μία εκτίμηση για την τρέχουσα κοινή γνώμη απέναντι σε ένα πλήθος θεμάτων.

Ο τεράστιος αυτός όγκος δεδομένων όμως κάνει την ανάλυσή τους από ανθρώπους πρακτικά αδύνατη. Έτσι αναγκαζόμαστε να καταφύγουμε στη χρήση υπολογιστών και ειδικών αλγορίθμων. Το πρόβλημα είναι ότι οι υπολογιστές δεν μπορούν να αναγνωρίσουν το ανθρώπινο συναίσθημα μέσω του γραπτού ή και προφορικού λόγου και άρα δεν μπορούν να ξέρουν αν η κοινή γνώμη είναι υπέρ ή κατά ενός θέματος. Για να αντιμετωπιστεί αυτή η αδυναμία καταφεύγουμε σε τεχνικές μηχανικής μάθησης όπως θα δούμε παρακάτω.

Ο τομέας του Sentiment Analysis είναι ένα αρκετά ενδιαφέρον θέμα μελέτης με χιλιάδες σχετικές έρευνες τα τελευταία χρόνια. Ο όρος Sentiment Analysis αναφέρεται στην ανάλυση κάποιας πηγής πληροφοριών έτσι ώστε να μπορέσουμε να εξάγουμε μία αντικειμενική εκτίμηση του συναισθήματος που εκφράζει. Ο μεγαλύτερος όγκος της έρευνας έχει επικεντρωθεί στην ανάλυση κειμένου αλλά υπάρχουν και παραδείγματα ερευνητών που προσπαθούν να αναλύσουν άλλες πηγές πληροφοριών, όπως ήχο και εικόνα.

Αυτός ο τόσο δημοφιλής τομέας της τεχνητής νοημοσύνης αποτελεί ακριβώς το εργαλείο που χρειαζόμαστε για να ανακαλύψουμε την κοινή γνώμη για κάποιο θέμα. Μπορεί να εφαρμοστεί σε εκατομμύρια κείμενα και να παράγει ποσοστά που θα μας λένε κατά πόσο η άποψη αυτών των κειμένων, κατά μέσο όρο, είναι υπέρ ή κατά ενός θέματος. Ο μόνος περιορισμός στην επιλογή θέματος είναι να μπορούμε να βρούμε σχετικά δεδομένα που εκφέρουν κάποια άποψη.

Ένας από τους πιο κοινούς όρους σε αυτές τις έρευνες είναι η πολικότητα (polarity). Η πολικότητα σε αυτήν την έρευνα θα είναι συνώνυμος όρος με την πολικότητα συναισθήματος (sentiment polarity) και θα αποτελεί την αντιστοίχηση της συναισθηματικής τοποθέτησης της πηγής ως αρνητική, ουδέτερη ή θετική προς τα δεδομένα που περιγράφει. Έτσι, μία θετική πολικότητα ενός κειμένου σημαίνει ότι ο συγγραφέας του κειμένου έχει θετική άποψη ή θετικά συναισθήματα προς το θέμα για το οποίο γράφει, ενώ μία αρνητική σημαίνει αντίστοιχα ότι έχει αρνητικά συναισθήματα ή αρνητική άποψη για το θέμα.

Οι εφαρμογές αυτής της έρευνας είναι επίσης αρκετά πολυάριθμες και εκτείνονται από ερευνητικούς έως και εμπορικούς τομείς. Αυτή η χρησιμότητα μίας τέτοιας έρευνας έρχεται κυρίως λόγω του τεράστιου όγκου δεδομένων που παράγονται στα κοινωνικά δίκτυα σήμερα. Αυτός ο όγκος κάνει την επεξεργασία των δεδομένων από ανθρώπους αδύνατη και άρα απαιτεί έναν αυτοματοποιημένο τρόπο επεξεργασίας τους. Έτσι στατιστικές μελέτες, προβλέψεις κοινωνικών καταστάσεων, δημοτικότητα προϊόντων και πολλά άλλα μπορούν να αντιμετωπιστούν με sentiment analysis τεχνικές.

Σε κάθε περίπτωση η έρευνα δεν έχει καταφέρει να φτάσει σε βαθμό απόλυτης επιτυχίας, αφήνοντας πάντα περιθώρια για εξέλιξη είτε σε ακρίβεια είτε σε απόδοση. Η απόδοση σε αυτόν τον τομέα γίνεται ένας ιδιαίτερα σημαντικός παράγοντας, κυρίως λόγω των αλγορίθμων που είναι απαραίτητη για αυτούς τους σκοπούς. Στη σύντομη επισκόπηση της βιβλιογραφίας που ακολουθεί θα δούμε κάποιους από αυτούς τους αλγορίθμους καθώς και τους λόγους που η απόδοση είναι ένα τόσο σημαντικό ζήτημα.

2 Βιβλιογραφική Επισκόπηση

2.1 Βασικές αρχές

Όπως σημειώθηκε και παραπάνω, το τέλειο εργαλείο για την εξερεύνηση της κοινής γνώμης μας προσφέρεται από το πεδίο της τεχνητής νοημοσύνης και συγκεκριμένα τον τομέα του Sentiment Analysis. Αυτός ο τόσο δημοφιλής στις μέρες μας τομέας, μας προσφέρει τα εργαλεία έτσι ώστε να μπορούμε να αναλύουμε την άποψη που εκφράζουν εκατομμύρια πηγές δεδομένων για ένα ή περισσότερα θέματα. Το μόνο που μας περιορίζει στην επιλογή θέματος είναι η δυνατότητα να βρούμε σχετικά δεδομένα που εκφέρουν κάποια άποψη. Αν μπορούμε να βρούμε τέτοια δεδομένα, τότε μπορούμε να ορίσουμε το θέμα μας σε επίπεδο παραμέτρων μίας πολιτικής και άρα να συσχετίσουμε κάθε παράμετρο με μία κοινή γνώμη, είτε θετική, είτε αρνητική είτε και ουδέτερη.

Με αυτόν τον τρόπο θα μπορέσουμε να παράγουμε ένα μαθηματικό μοντέλο το οποίο θα είναι σε θέση να λαμβάνει τις παραμέτρους που συνιστούν κάποια πολιτική και να μας αποκαλύπτει τι θα σκέπτονταν ο πληθυσμός για αυτήν την πολιτική. Φυσικά για να γίνει κάτι τέτοιο και να είναι αποτελεσματικό σε βάθος χρόνου, χρειαζόμαστε ένα επίπεδο ακρίβειας στους αλγορίθμους του Sentiment Analysis που δεν παρέχεται από τις σημερινές συνήθεις πρακτικές που περιγράφουμε παρακάτω.

Η συνήθης διαδικασία που προκύπτει από τη βιβλιογραφία σε τέτοιες έρευνες είναι η επιλογή πηγής δεδομένων και στη συνέχεια η επεξεργασία των δεδομένων με κάποια μέθοδο επεξεργασίας φυσικού λόγου (NLP) έτσι ώστε να εξαχθούν κάποια αντικειμενικά κριτήρια κατηγοριοποίησης κάθε κειμένου ανάλογα με την πολικότητά του. Τέλος, αυτά τα κριτήρια καθώς και οι τιμές τους εισάγονται σε έναν αλγόριθμο μηχανικής μάθησης ο οποίος έπειτα μπορεί να αντιστοιχίσει όλες αυτές τις τιμές σε ένα αντικειμενικό μέτρο πολικότητας.

Αυτό το μέτρο συνήθως έχει δύο τιμές -αρνητική ή θετική- [8]μερικές φορές έχει τρεις τιμές -αρνητική, θετική ή ουδέτερη- [4] και μερικές φορές έχει κάποια αριθμητική τιμή που συμβολίζει και το κατά πόσο θετική ή αρνητική είναι η πολικότητα. Σε κάθε περίπτωση μπορούμε να κρίνουμε την άποψη του δημιουργού των δεδομένων που αναλύσαμε για το θέμα με το οποίο ασχολείται απλά κοιτώντας μία τιμή. Αυτό μας επιτρέπει να χρησιμοποιούμε τα αποτελέσματα της ανάλυσης σε αυτοματοποιημένα προγράμματα για κάθε πιθανό σκοπό.

2.2 Κοινή Γνώμη

Αν και διαισθητικά, αρκετοί πιστεύουμε ότι η κοινή γνώμη είναι ένας ευμετάβλητος παράγοντας, οι σχετικές έρευνες αποδεικνύουν ότι στην πραγματικότητα είναι κάτι αρκετά σταθερό. Οι James N. Druckman et. al. (2012) [6] μας δείχνουν ότι η ατομική άποψη του κάθε πολίτη είναι αρκετά ευμετάβλητη σε βάθος χρόνου, αλλά η κοινή γνώμη ενός συνόλου πολιτών είναι σχετικά σταθερή. Στην έρευνά τους εξετάζουν αυτό το χάσμα σταθερότητας και τους λόγους για τους οποίους δημιουργείται.

Σύμφωνα με αυτήν την έρευνα λοιπόν, η αστάθεια στις απόψεις ενός συγκεκριμένου πολίτη (micro-level) μπορεί να οφείλεται σε αδυναμία του ερωτηματολογίου να καλύψει ακριβώς τις απόψεις του, σε αδυναμία του ερευνητή να μην επηρεάσει τον πολίτη τη στιγμή που απαντάει στις ερωτήσεις ή και σε αδυναμία του ερωτώμενου να κατανοήσει τις ερωτήσεις ή τις απαντήσεις τους. Αυτά τα λάθη αποδυναμώνονται όταν περάσουμε σε ανώτερο επίπεδο (macro-level) και αρχίσουμε να εξετάζουμε αθροίσματα και μέσους όρους ενός συνόλου απαντήσεων.

Εκτός από αυτό όμως υπάρχει και ένας δεύτερος παράγοντας που μπορεί να επηρεάσει τα αποτελέσματα μίας τέτοιας έρευνας. Η επιλογή θέματος για τα ερωτηματολόγια φαίνεται να έχει αρκετά μεγάλη σημασία. Αν το θέμα αφορά άμεσα τους πολίτες και προβάλλεται από τα μέσα μαζικής ενημέρωσης, τότε αυτοί θα έχουν σχηματίσει κάποια ισχυρή άποψη κατά ή υπέρ του. Σπάνια κάποιος μένει ουδέτερος απέναντι σε ένα τόσο προβεβλημένο θέμα που τον αφορά άμεσα. Σε αυτές τις περιπτώσεις όσο και να προσπαθήσει να επηρεάσει τον ερωτώμενο ο ερευνητής ή όσο κατευθυντική και να είναι η ερώτηση, ο πολίτης θα κρατήσει την άποψή του και μάλιστα μπορεί ακόμα και να την ενδυναμώσει προκειμένου να αντισταθεί σε αυτήν την προσπάθεια υποβολής.

Αν τώρα το θέμα δεν είναι τόσο προβεβλημένο στο κοινό και δεν αφορά άμεσα τους πολίτες, είναι αρκετά πιθανό να μην έχουν σχηματίσει μία ισχυρή άποψη. Έτσι είναι αρκετά πιο ευάλωτοι στη διατύπωση των ερωτήσεων και στο ύφος του ερευνητή. Ακόμα και αν η διατύπωση των ερωτήσεων τους προσφέρει αντικειμενικές ενδείξεις υπέρ ή κατά ενός θέματος, αυτοί με ισχυρή άποψη απλά θα επιμείνουν στην άποψή τους. Αυτό προφανώς κάνει όσους δεν έχουν κάποια ισχυρή άποψη πιο αντικειμενικούς κατά τη λήψη αποφάσεων.

Στις έρευνες που εξετάζουν την άποψη σε ατομικό επίπεδο (micro-level) λοιπόν, τα θέματα που επιλέγονται είναι συνήθως λιγότερο προβαλλόμενα [6], ενώ σε έρευνες μεγαλύτερης κλίμακας (macro-level) τα θέματα είναι πιο προβαλλόμενα και άμεσα [6]. Έτσι, το χάσμα που υπήρχε ήδη από τα τυχαία, ατομικά σφάλματα ενδυναμώνεται από τη στοχευμένη επιλογή θέματος.

2.3 Πηγές δεδομένων

Το πρώτο βήμα σε κάθε έρευνα αυτού του είδους είναι ο καθορισμός πηγών για τα δεδομένα που θα αναλυθούν. Αυτές οι πηγές μπορεί να είναι τα κοινωνικά δίκτυα [4], κάποια blogs [8] ή ακόμα και το YouTube [7]. Στην ουσία μπορούμε να χρησιμοποιήσουμε κάθε συλλογή απόψεων για κάποιο αντικείμενο, αναλύοντάς την και εξάγοντας συμπεράσματα για την πολικότητά του ως προς το συγκεκριμένο θέμα. Σε αυτόν τον τομέα πρέπει να δείξουμε ιδιαίτερη προσοχή κυρίως σε δύο τομείς: την επιλογή των δεδομένων και τους νομικούς περιορισμούς που επιβάλλονται σε αυτά.

Τα δεδομένα σε κάθε περίπτωση είναι πάρα πολλά και συχνά, και αν δεν τα επιλέξουμε σωστά τότε θα έχουμε αρκετά προβλήματα με το μέγεθός τους. Δεν είναι λίγες οι περιπτώσεις που τα δεδομένα ήταν τόσο μεγάλου μεγέθους που οι χρόνοι απόκρισης της ανάλυσης την έκαναν ακατάλληλη για πρακτική χρήση. Έτσι είναι απαραίτητη η επικέντρωση σε ένα μικρό αριθμό πηγών καθώς και η δημιουργία κάποιου μηχανισμού που θα αναγνωρίζει και θα απομονώνει μόνο τα δεδομένα που μας είναι χρήσιμα για τη συγκεκριμένη έρευνα.

Αυτός ο μηχανισμός μπορεί να είναι από ένα απλό εργαλείο αναζήτησης με keywords, μέχρι ένα πολύπλοκο εργαλείο τεχνητής νοημοσύνης που με χρήση semantics ανιχνεύει τα δεδομένα για συγκεκριμένες έννοιες αντί για λέξεις. Φυσικά η ποιότητα των δεδομένων είναι καλύτερη όσο πιο έξυπνος είναι ο μηχανισμός επιλογής τους. Κάποιες φορές όμως είναι απαραίτητο να συλλέγουμε νέα δεδομένα ανά τακτά χρονικά διαστήματα, οπότε ένας περίπλοκος και αργός μηχανισμός συλλογής τους δεν θα ήταν πρακτικός. Δεν θα ήταν λογικό να περιμένουμε μία μέρα για να αναλύσουμε τη σημερινή κοινή γνώμη καθώς δεν θα ήταν πλέον η σημερινή κοινή γνώμη.

Τα νομικά προβλήματα αφορούν κυρίως τα APIs [2] που προσφέρουν κάποια κοινωνικά δίκτυα και blogs. Αυτοί οι περιορισμοί μας απαγορεύουν να χρησιμοποιήσουμε τα δεδομένα που συλλέγουμε όπως και όποτε θέλουμε, θέτοντάς μας συγκεκριμένους κανόνες. Κάποια κοινωνικά δίκτυα απαιτούν την άμεση ενημέρωση των δεδομένων που έχουμε συλλέξει αν ενημερωθεί η πηγή, ενώ κάποια άλλα απαιτούν τη διαγραφή των δεδομένων μετά την πάροδο ενός χρονικού διαστήματος.

2.4 Natural Language Processing

Αφού συλλέξουμε τα δεδομένα πρέπει να δημιουργηθεί κάποιος μηχανισμός που θα μετατρέπει όλο αυτό το κείμενο φυσικού λόγου σε μορφή που μπορεί να επεξεργαστεί από τους αλγορίθμους μηχανικής μάθησης. Φυσικά αυτό το αντικείμενο έχει μελετηθεί εδώ και αρκετά

χρόνια από πολυάριθμες έρευνες και έχουν κατασκευαστεί κάποιες τυποποιημένες μεθόδους, τις δημοφιλέστερες από τις οποίες θα δούμε παρακάτω.

Η πιο κοινώς χρησιμοποιούμενη από αυτές είναι η Bag of Words [8] (τσουβάλι λέξεων) η οποία αποτελεί τον απλό διαχωρισμό του κειμένου σε λέξεις. Σε αυτήν τη μέθοδο δεν έχει καμία σημασία η σειρά των λέξεων ή οι συσχετισμοί μεταξύ τους, είτε αυτοί είναι νοηματικοί είτε γραμματικοί. Στις περισσότερες γλώσσες υλοποιείται απλά βρίσκοντας τα κενά μεταξύ των λέξεων και τα σημεία στίξεως αλλά υπάρχουν και γλώσσες (όπως τα κινέζικα) που ο διαχωρισμός των λέξεων είναι πολύ πιο δύσκολος λόγο της έλλειψης κενών [4].

Αυτή η μέθοδος συνήθως ακολουθείται και από ένα λεξικό που περιέχει τις ήδη βαθμολογημένες με έναν βαθμό πολικότητας λέξεις. Έτσι μπορούμε να αντιστοιχήσουμε την κάθε λέξη με το αντίστοιχο βάρος της. Αν αυτές οι αριθμητικές τιμές συνδυαστούν κατάλληλα, θα μπορούμε να παράγουμε έναν αριθμό που μας δείχνει αν η συνολική πολικότητα του κειμένου είναι θετική, αρνητική η και ουδέτερη. Φυσικά αυτές οι μέθοδοι είναι αρκετά απλουστευμένες καθώς δεν μπορούν να λάβουν υπόψιν την πολικότητα πλαισίου (contextual polarity) [11].

Ως πολικότητα πλαισίου θεωρείται η νέα πολικότητα κάθε λέξης αν λάβουμε υπόψιν το πλαίσιο στο οποίο βρίσκεται [11][4]. Αρκετές φορές μία λέξη έχει ανεξάρτητη πολικότητα από το πλαίσιο μέσα στο οποίο τη συναντούμε, αλλά ακόμα περισσότερες φορές το πλαίσιο μπορεί να μεταβάλει ριζικά την πολικότητα της λέξης. Για παράδειγμα η λέξη «προσβολή» αποτελεί μία αρνητική λέξη, αν όμως η πρόταση στην οποία βρίσκεται είναι η «Αποτελεί προσβολή το να μην αγοράσεις αμέσως το PS4» τότε η λέξη αποκτά θετική πολικότητα ως προς το θέμα, δηλαδή το PS4.

Μία παρόμοια τεχνική είναι ο διαχωρισμός πάλι κάθε πρότασης, αλλά αυτήν τη φορά όχι σε λέξεις αλλά σε N-gramms [4][2]. Αυτά αποτελούν αυτόματα δημιουργημένες λέξεις Ν γραμμάτων οι οποίες για τον άνθρωπο δεν σημαίνουν κάτι αλλά ο υπολογιστής μπορεί να τις αντικαταστήσει με κάποια τιμή πολικότητας. Συνήθως ως Ν χρησιμοποιείται ο αριθμός 3 ή 4 [4][2], παράγοντας έτσι “λέξεις” μεγέθους τριών ή τεσσάρων γραμμάτων αντίστοιχα. Πάλι όμως αυτές οι λέξεις σχηματίζουν ένα τσουβάλι λέξεων χωρίς να λαμβάνουμε υπόψιν τους συσχετισμούς μεταξύ τους.

Η αντιστοίχηση κάθε λέξης από αυτές με την τιμή πολικότητάς της αυτήν τη φορά δεν μπορεί να γίνει με κάποιο λεξικό καθώς δεν υπάρχει ένα τέτοιο λεξικό. Κάθε κείμενο παράγει τα δικά του N-gramms τα οποία μπορεί να είναι μοναδικά για το κάθε θέμα και κάθε πηγή

δεδομένων. Οι μόνοι τρόποι να αντιστοιχήσουμε τα N-gramms με τιμές πολικότητας είναι με ανθρώπινη παρέμβαση, με κάποιο αλγόριθμο μηχανικής μάθησης ή με συνδυασμό των δύο όπως είναι η συνήθης πρακτική.

Αν τώρα πάμε σε ακόμα υψηλότερο επίπεδο μπορούμε να δούμε και μεθόδους που χωρίζουν το κείμενο σε ολόκληρες προτάσεις [10][2] και πραγματοποιούν την ανάλυση σε επίπεδο προτάσεων. Φυσικά αυτή η μέθοδος, αν και πολύ οικονομική από απαιτήσεις χρόνου και επεξεργαστικής ισχύος, δεν είναι αρκετά λεπτομερής για να δώσει εμπορικά αξιοποιήσιμη ακρίβεια. Έτσι, αυτή η μέθοδος έχει μείνει σε ερευνητικό επίπεδο.

2.5 Αλγόριθμοι Μηχανικής Μάθησης

Οι αλγόριθμοι μηχανικής μάθησης είναι απαραίτητοι στη λύση του συγκεκριμένου προβλήματος καθώς μας βοηθούν να βαθμολογήσουμε αντικειμενικά κάθε πηγή δεδομένων με μία τιμή πολικότητας. Φυσικά αυτή η βαθμολόγηση μπορεί να συμβεί σε περισσότερα από ένα επίπεδα κάποιων μεθόδων, όπως στην περίπτωση των N-gramms που θα δούμε παρακάτω. Οι πιο κοινώς χρησιμοποιούμενοι αλγόριθμοι στη βιβλιογραφία είναι τα Naïve Bayesian Networks, ο C4.5, τα Hidden Markov Models και τα Support Vector Machines

Τα Naïve Bayesian Networks λειτουργούν δημιουργώντας δένδρα πιθανών καταστάσεων [4][5], σχετίζοντας έτσι με κάποια πιθανότητα τις προς ανάλυση μεταβλητές με κάποια από τις καταστάσεις της μεταβλητής στόχου μας. Στη συγκεκριμένη περίπτωση αυτό μπορεί να σημαίνει ότι σχετίζει συγκεκριμένα διαστήματα τιμών πολικότητας με την τιμή «Positive» και ένα άλλο με την τιμή «Negative». Ο συγκεκριμένος αλγόριθμος λειτουργεί με βάση τις πιθανότητες και άρα μπορούμε να δούμε και το περιθώριο λάθους που αφήνει.

Ο C4.5 είναι βασισμένος στον ID3 [4][5][9] και δημιουργεί και αυτός δέντρα πιθανών κατηγοριοποιήσεων, αλλά αυτός ο αλγόριθμος δεν βασίζεται σε πιθανότητες. Ο C4.5 μετράει τα σύνολα των δεδομένων μας και λειτουργεί βασισμένος στην τιμή του κέρδους πληροφορίας (Information Gain), το οποίο με τη σειρά του είναι βασισμένο στην αρχή της εντροπίας. Έτσι το δέντρο μας θα κατασκευαστεί όσο μικρότερο γίνεται, μεγιστοποιώντας την πιθανότητα σωστής κατηγοριοποίησης του κειμένου μας σε θετικό, αρνητικό ή και ουδέτερο.

Τα Support Vector Machines (SVMs) λειτουργούν βρίσκοντας κάποια μαθηματική συνάρτηση που μπορεί να αναπαραστήσει τις αντιστοιχήσεις που του δίνουμε κατά την εκπαίδευσή του [12][4][3]. Συγκεκριμένα, κατά την εκπαίδευση λαμβάνει ένα σύνολο τιμών και των αντίστοιχων κατηγοριοποιήσεών τους και προσπαθεί να βρει τη συνάρτηση, η γραφική

παράσταση της οποίας θα έχει τη μικρότερη απόσταση από κάθε σημείο των δεδομένων εκπαίδευσης που του δώσαμε.

Τα Hidden Markov Models (HMMs) λειτουργούν με παρόμοιο τρόπο προσπαθώντας να συσχετίσουν τις τιμές με την κατηγοριοποίηση που τους δώσαμε [7][1]. Αποτελούν έναν ενδιάμεσο αλγόριθμο μεταξύ των SVMs και των Naïve Bayesian Networks αφού προσπαθεί να φτιάξει, χρησιμοποιώντας μαθηματικά μοντέλα, κάποιες αλυσίδες Markov (Markov Chains), τις οποίες μετά κατηγοριοποιεί σύμφωνα με τα δεδομένα εκπαίδευσης που το δώσαμε, σχηματίζοντας ένα δέντρο παρόμοιο με αυτό των Naïve Bayesian Networks.

Ένα νεότερο είδος δέντρων κατηγοριοποίησης είναι τα Best-First Trees (BFTrees) τα οποία λειτουργούν όπως τα C4.5 δέντρα αλλά με ακόμα περισσότερες δυνατότητες κατά τη δημιουργία του δέντρου. Αυτά τα δέντρα χωρίζουν τα δεδομένα σε κλαδιά με βάση τον "καλύτερο" διαχωρισμό [13]. Ως καλύτερος ορίζεται ο διαχωρισμός που θα ικανοποιεί σε μεγαλύτερο βαθμό κάποια συνθήκη που ορίζεται αντικειμενικά κατά την εφαρμογή του. Με αυτόν τον τρόπο μπορούμε να παραμετροποιήσουμε τον αλγόριθμο έτσι ώστε να ταιριάζει καλύτερα στο πρόβλημα που αντιμετωπίζουμε κάθε φορά.

Άλλη μία μέθοδος που αξίζει να αναφερθεί είναι η λογιστική παλινδρόμηση (Logistic Regression). Αυτές οι μέθοδοι προσπαθούν, όπως και τα SVMs, να βρουν κάποια μαθηματική συνάρτηση η οποία θα περιγράφει με όσο μικρότερες αποκλίσεις γίνεται τα δεδομένα στον πολυδιάστατο χώρο. Σε αντίθεση όμως με τα SVMs, η παλινδρόμηση χρησιμοποιεί διάφορες λογιστικές εξισώσεις [14] για τον υπολογισμό αυτής της συνάρτησης, διαφοροποιώντας την αρκετά από των SVMs. Αυτή η διαφορά βέβαια δεν είναι πάντα θετική, καθώς η απόδοση κάθε αλγορίθμου εξαρτάται πάντα από το πρόβλημα που αντιμετωπίζει.

2.6 Πειραματικοί Συνδυασμοί

Στις σύγχρονες έρευνες δεν αρκούμαστε απλά στον συνδυασμό των παραπάνω μεθόδων και για να βελτιώσουμε τα αποτελέσματά τους προσθέτουμε κάποια επιπλέον βήματα και βελτιώσεις. Έτσι οι Godbole, Srinivasaiah και Skiena [8] χρησιμοποίησαν τον κλασικότερο συνδυασμό σε αυτόν τον τομέα, έκαναν ανάλυση χρησιμοποιώντας την Bag of Words μέθοδο και ένα λεξικό δίνοντας όμως βάση στις σχέσεις μεταξύ των λέξεων.

Αφού διαχώρισαν τις φράσεις σε λέξεις, έφτιαξαν νοηματικά μονοπάτια με τη χρήση του WordNet. Έτσι μπόρεσαν να μετρήσουν και τις πολικότητες των κειμένων χωρίς να λαμβάνουν υπόψιν το πλαίσιο αλλά και τις νέες πολικότητες με έναν πολύ απλό μηχανισμό αναγνώρισης

πλαισίου. Φυσικά δεν έπιανε τα σύνθετα νοηματικά πλαίσια αλλά μπορούσε να αναγνωρίσει την άρνηση -και άρα αντιστροφή πολικότητας- μίας λέξης.

Οι Aisopos, Papadakis, Tserpes και Varvarigou [4] χρησιμοποίησαν μία διαφορετική τεχνική τροποποιώντας τα N-gramms. Κατάφεραν να αναπαραστήσουν κάθε πρόταση με ένα γράφο από N-gramms, λαμβάνοντας έτσι υπόψιν τη συσχέτιση μεταξύ τους. Στην παρούσα έρευνα χρησιμοποιήθηκαν τα 2-gramms, 3-gramms και 4-gramms, ενώ ως αλγόριθμοι μηχανικής μάθησης χρησιμοποιηθήκαν οι Naïve Bayesian, C4.5 και Support Vector Machines.

Χρησιμοποιώντας τους γράφους που παράχθηκαν από τα N-gramms των δεδομένων εκπαίδευσης δημιουργήθηκε ένας merged γράφος για κάθε κατηγορία πολικότητας. Υπήρχε ένας γράφος για τις θετικές πολικότητες, ένας για τις αρνητικές και ένας για τις ουδέτερες. Κάθε νέος γράφος συγκρίνονταν με κάθε έναν από αυτούς τους γράφους, υπολογίζοντας τις αποστάσεις τους από αυτούς τους τρεις γράφους.

Ως αποστάσεις ορίστηκαν τρεις διαφορετικές μετρικές, η Containment Similarity, η οποία μετράει κατά πόσο ο γράφος που ελέγχεται περιέχεται στον merged γράφο, η Value Similarity, η οποία λαμβάνει υπόψιν και τα βάρη των σχέσεων μεταξύ των N-gramms και η Normalized Value Similarity που είναι ίδια με την Value Similarity με τη διαφορά ότι τώρα η μετρική ανεξαρτητοποιείται από τα μεγέθη των γράφων. Έτσι, για κάθε γράφο παράγεται ένα διάνυσμα αποστάσεων με εννέα τιμές, δηλαδή τρεις μετρικές για κάθε έναν από τους merged γράφους.

Αυτά τα διανύσματα στη συνέχεια τροφοδοτούνται στους αλγορίθμους μηχανικής μάθησης οι οποίοι τα κατηγοριοποιούν σε θετική, αρνητική ή ουδέτερη πολικότητα. Τα αποτελέσματά τους ήταν αρκετά καλύτερα από τις μεθόδους με απλά N-gramms καθώς οι συσχετίσεις μεταξύ τους αποδεικνύονται πειραματικά ότι έχουν αρκετά μεγάλη επιρροή στα αποτελέσματα. Τα βέλτιστα αποτελέσματα μάλιστα παρατηρήθηκαν από τον συνδυασμό 4-gramm graphs και του C4.5 αλγορίθμου.

Οι Morency, Mihalcea και Doshi [7] μας παρουσιάζουν ένα τρόπο να συλλέξουμε sentiment πληροφορίες από πολυμεσικά δεδομένα. Στη συγκεκριμένη έρευνα κατάφεραν να αναλύσουν βίντεο από το YouTube τα οποία συνέλεξαν με κριτήρια να έχουν ένα συγκεκριμένο θέμα και να αποτελούν τον μονόλογο ενός ανθρώπου που κοιτάει την κάμερα. Η διαδικασία τους αποτελούνταν από τρεις φάσεις, ανάλυση κειμένου, ανάλυση ήχου και ανάλυση εικόνας.

Αρχικά οι ερευνητές παρήγαγαν ένα transcript των λεγόμενων του ομιλητή, στο οποίο εφάρμοσαν τις κλασσικές sentiment analysis μεθόδους που είδαμε και παραπάνω. Έτσι πήραν

μία πρώτη τιμή πολικότητας καθαρά από τα λεγόμενά του η οποία όμως δεν αποτελούσε το τελικό αποτέλεσμα αλλά μόνο ένα ενδιάμεσο αποτέλεσμα το οποίο αργότερα θα συνδυαζόταν με τις υπόλοιπες τιμές.

Στη συνέχεια απομόνωσαν τον ήχο του video και ανέλυσαν τις συχνότητες. Είχαν ήδη ένα σύνολο από στατιστικά στοιχεία τα οποία τους έλεγαν ότι συγκεκριμένα ηχητικά φαινόμενα μπορούσαν να μεταφραστούν σε τιμές πολικότητας. Έτσι, για παράδειγμα, οι μεγάλες παύσεις μπορούσαν να αντιστοιχηθούν με αρνητικές πολικότητες, ενώ τα high pitches μπορούσαν να αντιστοιχηθούν με θετικές πολικότητες. Με αυτόν τον τρόπο άλλη μία τιμή πολικότητας παράχθηκε για το video.

Στην τρίτη φάση της ανάλυσης οι ερευνητές ανέλυσαν την εικόνα του video με παρόμοιο τρόπο με τον ήχο. Είχαν, δηλαδή, ένα σύνολο από παρατηρήσεις που μπορούσαν να αντιστοιχήσουν με συγκεκριμένες πολικότητες, όπως το να κοιτάει ο ομιλητής το δωμάτιο αντί για την κάμερα θεωρούνταν ως αρνητική πολικότητα, ενώ το να χαμογελάει θεωρούνταν ως θετική. Με το σύνολο αυτών των τιμών για τη διάρκεια του video δημιούργησαν την τρίτη τιμή.

Τελικά συνδύασαν τις τρεις τιμές παράγοντας ένα ενιαίο αποτέλεσμα για την πολικότητα του video. Τα αποτελέσματά τους ήταν αρκετά ικανοποιητικά δεδομένου ότι ακόμα και αν κάποιος από τους αλγορίθμους ανάλυσης έπεφτε έξω στην ανάλυσή του, οι άλλοι δύο θα τον κάλυπταν διορθώνοντας το γενικό αποτέλεσμα. Αν και αυτό το πείραμα ήταν με πολύ συγκεκριμένα videos, τα αποτελέσματα είναι αρκετά ενθαρρυντικά.

2.7 Συμπεράσματα

Αν και η μελέτη είναι αρκετά εκτενής, δεν έχει δημιουργηθεί ακόμα ένας αλγόριθμος με απόλυτη επιτυχία και πιθανότατα κάτι τέτοιο δεν είναι ακόμα εφικτό. Παρ’ όλα αυτά οι προσπάθειες συνεχίζουν και κάθε νέα τεχνική χτίζει πάνω στην προηγούμενη έτσι ώστε να βελτιώνεται όλο και περισσότερο η απόδοση αυτών των αλγορίθμων. Ταυτόχρονα όμως εξερευνούνται και νέα πεδία, νέοι αλγόριθμοι και νέες τεχνικές, όπως τα πειράματα με multimodal analysis, που πάνε ακόμα πιο μπροστά τον τομέα του Sentiment Analysis. Οι ερευνητές δεν έχουν σκοπό να σταματήσουν αυτήν την αναζήτηση για ένα σύστημα με σχεδόν ανθρώπινες δυνατότητες στον τομέα του sentiment analysis. Οι έρευνες προβλέπεται να συνεχίσουν μέχρι να επιτύχουμε αυτό το 100% επιτυχίας και ακόμα και τότε πάντα θα μπορούμε να βελτιώσουμε τον χρόνο απόκρισης του συστήματος, μετατρέποντάς το σε real time.

3 Πειράματα και Αποτελέσματα

3.1 Πειραματικά δεδομένα

3.1.1 Περιγραφή

Για την παρούσα έρευνα χρησιμοποιήθηκε μία συλλογή από tweets, τα οποία συλλέχθηκαν από διάφορες πηγές και σε διάφορες μορφές. Λόγω αυτής της ποικιλίας μορφών και κωδικοποιήσεων κρίθηκε απαραίτητο να μεταβληθούν σε ένα ενιαίο XML format, το οποίο σχεδιάστηκε ειδικά για τα πλαίσια της παρούσας έρευνας. Αυτό το XML περιελάμβανε τα πεδία description, polarity, prepolarity, date, id, user, retweets και favorites. Το description περιελάμβανε το κείμενο του tweet έτσι όπως αναρτήθηκε στο Twitter. Το date ήταν η ημερομηνία έκδοσης του tweet, το id ήταν ο αναγνωριστικός κωδικός του όπως μας δόθηκε από το twitter api, ο user ήταν ο χρήστης που το δημοσίευσε, το retweets έδειχνε τον αριθμό αναδημοσιεύσεων και το favorites έδειχνε πόσοι χρήστες δήλωσαν ότι αυτό το tweet είναι αγαπημένο τους.

Το polarity ήταν η πολικότητα που του αντιστοιχούσε ο αλγόριθμος και μπορούσε να ήταν μία θετική, αρνητική ή μηδενική ακέραια τιμή ή ακόμα και κενό αν δεν είχε τρέξει ακόμα κάποιος αλγόριθμος πάνω του. Αντίστοιχα το prepolarity ήταν η πολικότητα που είχε το tweet αν αυτό ήταν ήδη βαθμολογημένο και ακολουθούσε τις πιθανές τιμές του polarity. Τα tweets που συλλέχτηκαν για δοκιμαστικούς λόγους ήταν ήδη βαθμολογημένα από άλλους ερευνητές και έτσι μπορούσαν να χρησιμοποιηθούν άμεσα για την εκπαίδευση και δοκιμή των αλγορίθμων μηχανικής μάθησης. Αντίθετα τα tweets που συλλέγονται σε πραγματικό χρόνο από το Twitter δεν είναι βαθμολογημένα και άρα απαιτείται ανθρώπινη παρέμβαση για να βαθμολογηθεί η απόδοση των αλγορίθμων.

3.1.2 Καθαρισμός

Για να μπορέσουμε να επεξεργαστούμε τα δεδομένα αυτά έπρεπε πρώτα να καθαριστούν ώστε να μπορούν οι αλγόριθμοι να τα χρησιμοποιήσουν πιο εύκολα. Έτσι, δημιουργήθηκε ένας parser γραμμένος σε java, ο οποίος θα δημιουργούσε τα εικονικά tweets στη μνήμη του υπολογιστή διαβάζοντας και καθαρίζοντας τα διαθέσιμα tweets. Ο καθαρισμός χωρίστηκε σε δύο φάσεις: τον καθαρισμό του κειμένου και την αφαίρεση διπλοτύπων.

Ο καθαρισμός κειμένου ήταν μία αρκετά απλή διαδικασία. Ξεκίνησε με την μετατροπή του κειμένου σε lower case χαρακτήρες έτσι ώστε να ενισχυθεί η κατηγοριοποίηση και να διαχωρίσουμε το κείμενο από τα keywords που θα χρησιμοποιήσουμε παρακάτω. Στη συνέχεια όλες οι διευθύνσεις για διαδικτυακούς τόπους αντικαταστάθηκαν με το keyword “URL” γιατί

δεν μας ενδιέφερε στην παρούσα ανάλυση ποια σελίδα δείχνει το συγκεκριμένο tweet, αλλά μας ενδιέφερε το γεγονός ότι δείχνει σε κάποια σελίδα. Για τον ίδιο λόγο οι αναφορές σε χρήστες αντικαταστάθηκαν από το keyword "USER". Τα hashtags μας ήταν τελείως άχρηστα στην ανάλυσή μας και άρα το σύμβολο "#" διαγράφηκε τελείως από τα tweets αφήνοντας πίσω το hashtag ως μία κανονική λέξη. Αποφασίστηκε να μην αλλοιώσουμε τα stopwords και τα σημεία στίξης δεδομένου ότι η επαναλαμβανόμενη χρήση τους ή ακόμα και η έλλειψη χρήσης τους μπορεί να σχετίζονται με κάποια συγκεκριμένα συναισθήματα.

Ο καθαρισμός διπλοτύπων χωρίστηκε με τη σειρά του σε άλλα δύο επίπεδα. Σε πρώτο επίπεδο ένας java αλγόριθμος συνέκρινε τα IDs όλων των tweets μεταξύ τους και διέγραφε από το σύνολο όλα τα διπλότυπα που έβρισκε. Κάποια tweets όμως ήταν απλά αντιγραφές κάποιων άλλων και άρα είχαν το ίδιο κείμενο αλλά με διαφορετικό ID. Έτσι σε δεύτερο επίπεδο ο αλγόριθμος συνέκρινε τα κείμενα των tweets μεταξύ τους, βρίσκοντας ένα βαθμό ομοιότητας. Αν κάποιο tweet κρίνονταν ως πολύ όμοιο με κάποιο άλλο, τότε διαγραφόταν, κρατώντας στο σύνολο δεδομένων μόνο το πρώτο tweet. Όπως είναι εμφανές, αυτή η διαδικασία χρειαζόταν λίγο παραπάνω χρόνο από την απλή σύγκριση των IDs.

3.2 Επεξεργασία φυσικού λόγου

Προκειμένου τα καθαρισμένα και προεπεξεργασμένα πλέον δεδομένα να μπορούν να τροφοδοτηθούν στους αλγορίθμους μηχανικής μάθησης, έπρεπε να περάσουν από ένα ακόμα επίπεδο επεξεργασίας. Η επεξεργασία φυσικού λόγου έχει ως σκοπό να μετατρέψει το κείμενο σε κάτι που ο υπολογιστής μπορεί να καταλάβει και να αναλύσει. Για αυτόν τον σκοπό έχουν αναπτυχθεί πάρα πολλές τεχνικές, κάθε μία από τις οποίες έχει τις δυνάμεις και τις αδυναμίες της. Αρκετές από αυτές απορρίφτηκαν αμέσως καθώς δεν μπορούσαν να καλύψουν τις ανάγκες της παρούσας έρευνας αλλά τρεις από αυτές εξερευνήθηκαν ώστε να συγκρίνουμε τα αποτελέσματά τους. Αυτές ήταν οι Bag of Words, N-grams και N-gram Graphs τις οποίες θα παρουσιάσουμε αναλυτικότερα παρακάτω.

3.2.1 Bag of Words

Ίσως η πιο γνωστή και ευρέως χρησιμοποιούμενη τεχνική, η Bag of Words δουλεύει χωρίζοντας κάθε πρόταση σε ένα σύνολο λέξεων. Αυτές οι λέξεις συγκεντρώνονται σε ένα εικονικό τσουβάλι σχηματίζοντας το λεγόμενο Bag of Words. Ο αλγόριθμος αυτός δεν μπορεί να λάβει υπόψιν του τις σχέσεις μεταξύ των λέξεων ή ακόμα και τη σειρά που εμφανίστηκαν. Λόγω αυτής της αδυναμίας θεωρείται ένας αρκετά αδύναμος αλγόριθμος που δεν μπορεί να προβλέψει την πολικότητα πλαισίου, μία έννοια που παρουσιάσαμε παραπάνω και περιπλέκει αρκετά τη διαδικασία της ανάλυσης.

Συνήθως αυτές οι μέθοδοι χρησιμοποιούν ένα λεξικό το οποίο μπορεί να αντιστοιχίσει κάθε λέξη σε αυτό το τσουβάλι λέξεων με μία αριθμητική τιμή. Αυτή η τιμή αποτελεί την πολικότητα της συγκεκριμένης λέξης. Αν τώρα αντικαταστήσουμε κάθε λέξη με την πολικότητά της και βγάλουμε έναν μέσο όρο, θα λάβουμε τη συνολική πολικότητα της πρότασης. Γενικεύοντας, αν βγάλουμε έναν μέσο όρο αυτών θα λάβουμε την πολικότητα του κειμένου, έναν αριθμό που θα μας λέει αν το κείμενο εκφράζει μία αρνητική, μία θετική ή μία ουδέτερη άποψη απέναντι στο θέμα του.

3.2.2 N-grams

Μία λίγο πιο σύγχρονη και ισχυρή τεχνική είναι τα N-grams. Αυτή η μέθοδος σχηματίζει πάλι τσουβάλια λέξεων αλλά αυτήν τη φορά οι λέξεις δεν είναι κατανοητές από τον άνθρωπο. Κάθε λέξει αποτελείται από N χαρακτήρες (συνήθως 3, 4 ή 5). Στην περίπτωσή μας χρησιμοποιήσαμε και τους τρεις αυτούς αριθμούς σε τρεις διαφορετικές οικογένειες πειραμάτων έτσι ώστε να μελετήσουμε και την επίδραση της αύξησης ή μείωσης του αριθμού N σε κάθε διαφορετικό αλγόριθμο μηχανικής μάθησης. Φυσικά όσο μεγαλύτερο το N τόσο δυσκολότερη η δουλειά του αλγορίθμου, δεδομένου ότι δημιουργούνται μεγαλύτερες και περισσότερες λέξεις για επεξεργασία.

Αφού το νέο αυτό τσουβάλι ψευδολέξεων δημιουργηθεί, πρέπει να παραχθούν κάποιες τιμές ώστε να μπορούμε να τροφοδοτήσουμε τα δεδομένα μας στους αλγορίθμους μηχανικής μάθησης. Λόγω του γεγονότος ότι κάθε κείμενο παράγει το δικό του μοναδικό σύνολο ψευδολέξεων, δεν υπάρχει κάποιο λεξικό ώστε να αντικαταστήσουμε τις λέξεις αυτές με μία αριθμητική τιμή. Έτσι πρέπει αυτές οι τιμές να παραχθούν εκείνη τη στιγμή. Αν και ο καλύτερος τρόπος θα ήταν να βαθμολογείται κάθε N-gram από κάποιον άνθρωπο, κάτι τέτοιο δεν είναι εφικτό λόγω του τεράστιου πλήθους τους. Έτσι καταφεύγουμε σε ένα επιπλέον επίπεδο μηχανικής μάθησης.

Αυτό το επιπλέον επίπεδο θα εκπαιδεύεται με διάφορα τσουβάλια από N-grams τα οποία θα έχουν παραχθεί από ήδη βαθμολογημένα ως προς το sentiment τους κείμενα. Έτσι ανάλογα με τν συχνότητα που παρουσιάζεται κάθε N-gram στα θετικά, αρνητικά ή ουδέτερα τσουβάλια θα μεταβάλλεται και η πολικότητά του. Τελικά, κάθε N-gram θα συσχετιστεί με μία θετική, αρνητική ή μηδενική αριθμητική τιμή, η οποία θα αναπαριστά την θετική, αρνητική ή ουδέτερη αντίστοιχα πολικότητά του. Στη συνέχεια θα μπορούμε απλά να αντικαταστήσουμε το κάθε N-gram με την τιμή του και να βγάλουμε μέσους όρους όπως ακριβώς και με την κλασσική Bag of Words μέθοδο.

3.2.3 N-gram Graphs

Ως ένας άμεσος απόγονος των N-grams, τα N-gram graphs χρησιμοποιούν ακριβώς τις ίδιες μεθόδους με τον πρόγονό τους αλλά με τη διαφορά ότι αυτά λαμβάνουν υπόψιν και τις σχέσεις μεταξύ γειτονικών N-grams, ξεφεύγοντας έτσι από τις κλασσικές bag of words μεθόδους. Όπως και με τα κλασσικά N-grams, επιλέξαμε να πειραματιστούμε με 3-grams, 4-grams και 5-grams graphs, παρατηρώντας εκπληκτικά αποτελέσματα όπως θα δούμε στα παρακάτω κεφάλαια.

Σε αντίθεση με τα N-grams, τώρα δεν χρησιμοποιούμε ένα τσουβάλι ασύνδετων N-grams αλλά ένα συνδεδεμένο γράφο για την αναπαράσταση κάθε κειμένου. Έτσι η αντικατάσταση κάποιου N-gram με μία τιμή πολικότητας δεν θα μπορούσε να μας δώσει ένα αντικειμενικό μέτρο της τιμής πολικότητας του συνολικού γράφου. Για αυτόν τον λόγο έπρεπε να καταλήξουμε σε κάποια πιο πολύπλοκα μέτρα σύγκρισης. Έτσι, ακλουθώντας τη δουλειά των Fotis Aisopos et. al. (2012) [4] δημιουργήσαμε τρεις merged γράφους, ένα για τα θετικά, ένα για τα αρνητικά και ένα για τα ουδέτερα κείμενα.

Κάθε νέος γράφος που εισάγονταν προς κατηγοριοποίηση έπρεπε να συγκριθεί και με τους τρεις αυτούς γράφους. Η σύγκριση γινόταν με τρεις διαφορετικές μετρικές αποστάσεων, μία που απλά έδειχνε το κατά πόσο ο νέος γράφος περιέχονταν στον merged, μία που λάμβανε υπόψη και τα βάρη μεταξύ των σχέσεων των N-grams και μία που αφαιρούσε την εξάρτηση με τα σχετικά μεγέθη των δύο συγκρινόμενων γράφων. Έτσι, κάθε νέος γράφος μπορούσε να αντιστοιχηθεί με εννέα αριθμητικές τιμές, τρεις για κάθε merged γράφο.

Αυτές οι τιμές στη συνέχεια τροφοδοτούνταν σε κάθε έναν από τους αλγορίθμους μηχανικής μάθησης έτσι ώστε να κατηγοριοποιηθούν σε θετική, αρνητική ή ουδέτερη πολικότητα. Οι αλγόριθμοι αυτοί είναι απαραίτητοι σε αυτό το στάδιο λόγω του ότι δεν μπορούμε να γνωρίζουμε εκ των προτέρων τα όρια που πρέπει να έχουν οι αποστάσεις με τον κάθε merged γράφο έτσι ώστε να κατηγοριοποιήσουμε αυτόματα έναν νέο γράφο, απλά βλέποντας τις τιμές των εννιά αποστάσεων.

3.3 Πειράματα

Στα πλαίσια της παρούσας έρευνας εκτελέστηκαν πάνω από 250 πειράματα με σκοπό να διακρίνουμε ποιος συνδυασμός τεχνικών επεξεργασίας φυσικού λόγου και αλγορίθμων μηχανικής μάθησης θα μας έδινε την υψηλότερη δυνατή ακρίβεια. Σε αυτήν τη φάση αποκλείσαμε ένα μεγάλο σύνολο αλγορίθμων λόγω της ακαταλληλότητάς τους για την κατηγοριοποίηση που απαιτούνταν. Τα πειράματα χωρίστηκαν σε 3 μεγάλες οικογένειες

ανάλογα με τη μέθοδο επεξεργασίας φυσικού λόγου που χρησιμοποιούσαν. Αυτές ήταν οι Bag of Words, N-Grams και N-Gram Graphs.

3.3.1 Bag of Words

Για την Bag of Words τα πράγματα ήταν αρκετά απλά. Αρχικά χωρίσαμε κάθε κείμενο σε λέξεις χρησιμοποιώντας ένα parser γραμμένο σε Java. Ως διαχωριστικό λέξεων πήραμε το κενό (whitespace). Στη συνέχεια αφαιρέσαμε τα σημεία στίξης για να αντιστοιχηθούν πιο εύκολα οι λέξεις με τις καταχωρήσεις του λεξικού που θα δούμε παρακάτω. Ολοκληρώνοντας αυτήν τη διαδικασία είχαμε αντιστοιχήσει κάθε πρόταση σε ένα "Τσουβάλι λέξεων" για τις οποίες όμως δεν έχουμε κρατήσει τη σειρά με την οποία εμφανίστηκαν ή τις μεταξύ τους σχέσεις, μόνο το πλήθος εμφανίσεων.

Στη συνέχεια χρησιμοποιήσαμε το SentiWordNet 3.0 για να αντιστοιχήσουμε κάθε λέξη σε ένα αριθμητικό μέγεθος. Το SentiWordNet είναι ένα λεξικό το οποίο περιέχει αριθμητικά βάρη του συναισθήματος που απεικονίζει κάθε λέξη. Έτσι αντί να αναλύσουμε τις λέξεις, αναλύουμε πλέον την αριθμητική αξία του συναισθήματος που εκφράζει. Κάθε λέξη έχει ένα θετικό και ένα αρνητικό βάρος το οποίο μπορεί να κυμαίνεται από 0 έως 1 και δείχνει την πιθανότητα η λέξη να είναι αρνητική η θετική αντίστοιχα. Η πιθανότητα η λέξη να είναι ουδέτερη μπορεί να βρεθεί με τον απλό τύπο 1 - (αρνητική τιμή + θετική τιμή).

Στα πειράματά μας συγκρίναμε αυτές τις τρεις τιμές και αντιστοιχήσαμε κάθε λέξη με έναν αριθμό, χρησιμοποιώντας την εξής διαδικασία. Αν η αρνητική τιμή της λέξης υπερτερούσε τότε χρησιμοποιούσαμε αυτήν αλλά με αρνητικό πρόσημο. Αντίστοιχα, αν η θετική τιμή υπερτερούσε τότε χρησιμοποιούσαμε αυτή με θετικό πρόσημο. Αν υπερτερούσε η ουδέτερη τιμή τότε απλά θέταμε την αξία της λέξης στο 0. Στη συνέχεια απλά βγάζαμε έναν μέσο όρο των τιμών για όλη την πρόταση. Αν αυτός ο μέσος όρος ήταν μεγαλύτερος του μηδέν, η πρόταση εξέφραζε θετικό συναίσθημα, αν ήταν μικρότερος του μηδέν τότε εξέφραζε αρνητικό ενώ αν ήταν μηδέν τότε ήταν μία ουδέτερη πρόταση.

Βέβαια, όπως θα δούμε και στα αποτελέσματα παρακάτω, αυτή η εκδοχή ήταν πολύ απλοποιημένη με αποτέλεσμα να μην ξεπερνάει κατά πολύ το 33% επιτυχίας που είναι το όριο τυχαιότητας. Έτσι στη συνέχεια εκπαιδεύσαμε τους αλγορίθμους μηχανικής μάθησης που θα δούμε και παρακάτω, έτσι ώστε να ανακαλύψουν τα πραγματικά όρια για αυτόν τον μέσο όρο, δηλαδή από ποια τιμή και κάτω θα μπορούμε να θεωρούμε με ασφάλεια ότι μία πρόταση είναι αρνητική και από ποια τιμή και πάνω θετική. Φυσικά οι ενδιάμεσες τιμές θα ήταν οι ουδέτερες προτάσεις.

3.3.2 N-Grams

Για τα πειράματά μας επιλέξαμε να χρησιμοποιήσουμε τους αριθμούς 3,4 και 5 για μέγεθος ψευδολέξεων. Έτσι είχαμε τρεις υποοικογένειες πειραμάτων, τα 3-Grams, 4-Grams και 5-Grams. Αυτό μας έδωσε τη δυνατότητα να ελέγξουμε τον τρόπου που επηρεάζεται κάθε αλγόριθμος μηχανικής μάθησης με τη διαφοροποίηση στο πλήθος και το μέγεθος των δεδομένων που αναλύει. Όπως θα δούμε παρακάτω, όσο αυξάνεται το μέγεθος των N-Grams τείνουμε προς καλύτερα αποτελέσματα μέχρι που φτάνουμε σε κάποιο όριο, μετά το οποίο κάθε αύξηση του N οδηγεί σε μείωση της απόδοσης.

Για να μετρήσουμε την πολικότητα κάθε ψευδολέξης δημιουργήσαμε ένα parser πάλι σε Java, στον οποίο τροφοδοτήσαμε τα σύνολα θετικών, αρνητικών και ουδέτερων tweets που είχαμε ως σύνολο εκπαίδευσης. Αυτός ανέλυσε τα κείμενα, τα διέσπασε σε N-Grams, με το κατάλληλο κάθε φορά μέγεθος λέξης, και στη συνέχεια δημιούργησε ένα θετικό, ένα ουδέτερο και ένα αρνητικό τσουβάλι ψευδολέξεων. Αυτό το τσουβάλι περιείχε κάθε N-Gram που βρήκε στο αντίστοιχο σύνολο μαζί με τον αριθμό εμφανίσεών του. Έτσι ξέραμε όχι απλά αν ένα N-Gram έχει εμφανιστεί σε κάποιο θετικό, ουδέτερο ή αρνητικό tweet αλλά και πόσο συχνά εμφανίζεται στα θετικά, ουδέτερα ή αρνητικά tweets αντίστοιχα.

Στη συνέχεια απλά αναζητούσαμε κάθε νέο N-Gram που ερχόταν για βαθμολόγηση σε αυτά τα τρία τσουβάλια. Αν η ψευδολέξη υπήρχε πράγματι τότε απλά κρατούσαμε τον αριθμό εμφανίσεών της. Αν δεν υπήρχε τότε κρατούσαμε ως αριθμό εμφανίσεων το 0. Έτσι προέκυπταν τρεις τιμές, μία θετική, μία ουδέτερη και μία αρνητική για κάθε N-Gram. Στη συνέχεια δημιουργήσαμε ένα συνολικό τσουβάλι ψευδολέξεων το οποίο περιείχε όλα τα N-Grams που συναντήσαμε. Για κάθε εμφάνιση του N-Gram στο θετικό τσουβάλι, το βάρος του στο συνολικό τσουβάλι αυξανόταν κατά ένα, ενώ για κάθε εμφάνισή του στο αρνητικό μειωνόταν κατά ένα. Οι εμφανίσεις του στο ουδέτερο δεν επηρέαζαν το βάρος.

Με αυτόν τον τρόπο κάθε νέο N-Gram μπορούσε να αντικατασταθεί με έναν ακέραιο αριθμό που έδειχνε την πολικότητά του. Αυτή η δυνατότητα έκανε εφικτή την επιτυχημένη ανάλυσή τους από τους αλγορίθμους μηχανικής μάθησης που θα δούμε παρακάτω. Κάθε αλγόριθμος εκπαιδευόταν με αυτές τις αριθμητικές αναπαραστάσεις της πολικότητας κάθε N-Gram με αποτέλεσμα να μπορεί να προβλέψει την κατηγορία ενός νέου N-Gram ανάλογα με την τιμή πολικότητας που του έχουμε αντιστοιχήσει χρησιμοποιώντας την παραπάνω διαδικασία.

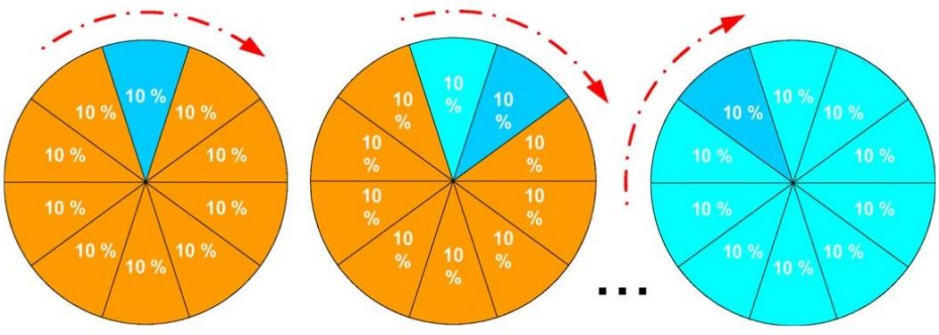
Για αυτήν την ανάλυση χρησιμοποιηθήκαν επτά αλγόριθμοι μηχανικής μάθησης, όλοι υλοποιημένοι στην ανοιχτή Java βιβλιοθήκη της Weka. Αυτοί οι αλγόριθμοι έχουν παρουσιαστεί αναλυτικότερα σε ανώτερο κεφάλαιο και είναι οι εξής: Naïve Bayesian Networks, C4.5, Support

Vector Machines, Logistic Regression, Multilayer Perceptrons, Best-First Trees και Functional Trees. Αυτοί οι επτά αλγόριθμοι επιλέχθηκαν λόγω της ειδικότητάς τους σε τόσο απαιτητικές λειτουργίες κατηγοριοποίησης και των συμβατών με τα ερευνητικά μας ερωτήματα δυνατοτήτων που προσφέρουν.

Η βιβλιοθήκη της Weka έχει καταφέρει να ομογενοποιήσει τους αλγορίθμους αυτούς έτσι ώστε να δέχονται τα δεδομένα στην ίδια μορφή και να βγάζουν τα αποτελέσματα σε παρόμοιες μορφές. Αυτό διευκόλυνε εξαιρετικά την πειραματική διαδικασία, δίνοντάς μας τη δυνατότητα να δοκιμάσουμε και τους επτά αλγορίθμους στα ίδια δεδομένα και να συγκρίνουμε τα αποτελέσματά τους, τα οποία θα δούμε σε επόμενο κεφάλαιο.

Για να δούμε αυτά τα αποτελέσματα χρησιμοποιήσαμε πάλι τη βιβλιοθήκη της Weka και συγκεκριμένα το crossValidateModel της. Αυτό το μοντέλο μας επιτρέπει να τροφοδοτήσουμε τους αλγόριθμους με ένα σύνολο δεδομένων εκπαίδευσης από το οποίο τα εννέα δέκατα θα χρησιμοποιηθούν για εκπαίδευση και το ένα για δοκιμή. Αυτό το ένα επιλέγεται τυχαία κάθε φορά που τρέχουν τα πειράματα αλλά μένει ίδιο και για τους επτά αλγορίθμους. Αυτή η διαδικασία επαναλαμβάνεται δέκα φορές έτσι ώστε κάθε υποσύνολο έχει την ευκαιρία να γίνει το κομμάτι δοκιμής. Στη συνέχεια βγάζουμε έναν μέσο όρο του ποσοστού επιτυχίας του κάθε αλγορίθμου και τα συγκρίνουμε.

Η διαδικασία που περιγράψαμε στην προηγούμενη παράγραφο είναι γνωστή ως 10-fold cross validation και χρησιμοποιείται πολύ συχνά στη βιβλιογραφία όταν θέλουμε να δοκιμάσουμε πειραματικά την απόδοση κάποιου αλγορίθμου κατηγοριοποίησης. Το crossValidateModel είναι απλά μία Java υλοποίηση αυτής της μεθόδου, που μπορεί να λάβει ορίσματα ώστε να παράγει ένα σταθερό, semi-randomized ή και randomized σύνολο δοκιμής. Στην παρούσα έρευνα επιλέξαμε το semi-randomized, τυχαιοποιόντας μόνο το πρώτο υποσύνολο που επιλέγεται για δοκιμή και στη συνέχεια ανατρέχοντας σειριακά όλα τα υπόλοιπα.

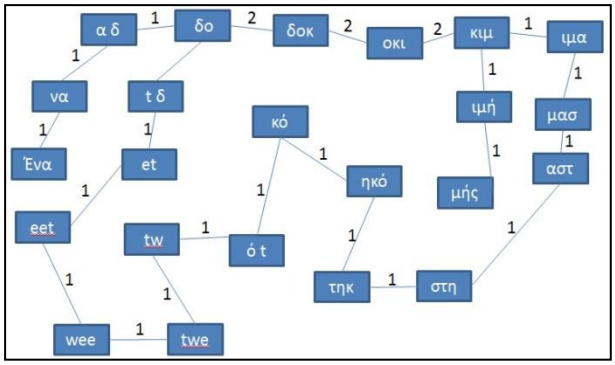


Εικόνα 1: Σχηματική αναπαράσταση της 10-fold cross validation πηγή: Delen et. al. (2004)

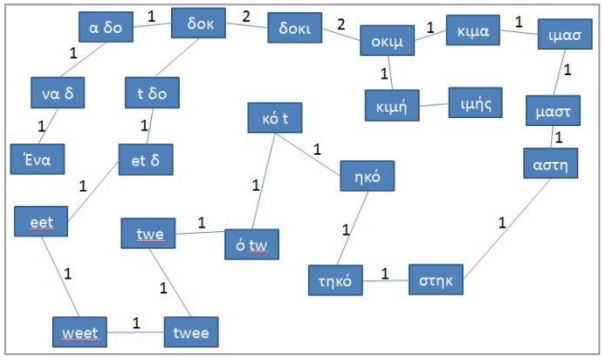
3.3.3 N-Gram Graphs

Τα N-Gram Graphs ήταν αρκετά πιο πολύπλοκα και οδηγούσαν σε ιδιαιτέρως χρονοβόρα πειράματα. Όπως και με τα απλά N-Grams, έτσι και εδώ τα πειράματα χωρίστηκαν σε τρεις υποοικογένειες με 3, 4 ή 5 χαρακτήρες σε κάθε N-Gram. Αυτήν τη φορά όμως δεν αναπαριστούσαμε κάθε tweet με ένα τσουβάλι λέξεων, αλλά με έναν γράφο ο οποίος μας έδειχνε και τις σχέσεις μεταξύ γειτονικών N-Grams. Όπως θα δούμε σε παρακάτω κεφάλαιο, τα αποτελέσματα ήταν αρκετά εκπληκτικά και ταυτόχρονα πολύ καλύτερα από των απλών N-Grams.

Για να καταλάβουμε καλύτερα τη διαδικασία με την οποία δημιουργούνται τα N-Gram Graphs αρκεί να κοιτάξουμε τους παρακάτω γράφους (Εικόνες 2.α και 2.β) της φράσης "Ένα δοκιμαστικό tweet δοκιμής". Στην αριστερή εικόνα (2.α) βλέπουμε τον 3-Gram γράφο της πρότασης ενώ στη δεξιά (2.β) βλέπουμε τον 3-Gram γράφο. Τα βάρη που είναι σημειωμένα στις ακμές δείχνουν το πλήθος εμφανίσεων αυτής της σχέσης. Όπως φαίνεται, αν ακολουθήσουμε τον γράφο προσπερνώντας Ν-1 κόμβους τότε σχηματίζεται η πρόταση. Δηλαδή για τον 3-Gram γράφο αν προσπερνάμε 2 κόμβους κάθε φορά ή για τον 4-Gram γράφο 3 κόμβους, τότε μπορούμε να ανακτήσουμε την αρχική πρόταση.



Εικόνα 1.α



Εικόνα 2.β

Κάθε tweet λοιπόν διασπάστηκε σε N-Grams του προκαθορισμένου κάθε φορά μήκους, τα οποία χρησιμοποιήθηκαν για τη δημιουργία ενός γράφου. Κάθε κόμβος του γράφου είναι μία ψευδολέξη ενώ κάθε ακμή του δείχνει μία σχέση γειτνίασης μεταξύ των δύο αυτών ψευδολέξεων. Αν μέσα στο κείμενο εμφανιστούν οι ίδιες ψευδολέξεις με την ίδια σχέση γειτνίασης, τότε το βάρος της ακμής που συμβολίζει αυτήν τη σχέση αυξάνεται κατά ένα.

Αμέσως μετά την αναπαράσταση κάθε κειμένου με έναν γράφο παρουσιάστηκε η ανάγκη να βαθμολογήσουμε αριθμητικά την πολικότητα κάθε τέτοιου γράφου. Για αυτόν το σκοπό ακολουθήσαμε τη διαδικασία που αναλύεται από τους Fotis Aisopos et. al. (2012) [4] η οποία περιγράφεται στη συνέχεια. Χρησιμοποιώντας τον τύπο της update functionality παράγεται ένας συνολικός γράφος για κάθε μία από τις τρεις κατηγορίες πολικότητας που εξετάζουμε. Αυτός ο

συνολικός γράφος G_u περιγράφεται από τον μαθηματικό τύπο σύμφωνα με τον οποίο αν G_u είναι ο συνολικός γράφος, E^u είναι το σύνολο των ακμών του G_u, V^u είναι το σύνολο κόμβων του G_u και W^{u_i} το σύνολο των βαρών κάθε ακμής είναι ο εξής: $G_u = (E^u, V^u, W^{u_i})$. Το E^u τώρα μπορούμε να το υπολογίσουμε ως εξής: $E^u = E^{GTp} \cup E^{Gti}$, όπου το E^{GTp} είναι το σύνολο ακμών του συνολικού γράφου μέχρι στιγμής και E^{Gti} είναι το σύνολο ακμών του γράφου που θέλουμε να συγχωνέψουμε, ο οποίος είναι ο γράφος νούμερο i που συγχωνεύουμε στον συνολικό. Αντίστοιχα για να βρούμε το Vu χρησιμοποιούμε το $Vu = V^{GTp} \cup V^{Gti}$, ενώ ο τύπος για το βάρος είναι λίγο πιο πολύπλοκος. Έτσι για το βάρος της ακμής e έχουμε $W^{u_i}(e) = W^{GTp}(e)+(W^{Gti}(e)-W^{GTp}(e)) \times 1/i$, εξασφαλίζοντας έτσι ότι το συγχωνευμένο βάρος θα πλησιάζει το μέσο βάρος της συγκεκριμένης ακμής για όλους τους συγχωνευμένους γράφους.

Με αυτόν τον τρόπο έχουμε τρεις συνολικούς γράφους, έναν για τα θετικά tweets, έναν για τα αρνητικά και ένα για τα ουδέτερα. Κάθε νέος γράφος θα συγκρίνεται με αυτούς τους τρεις έτσι ώστε να κρίνουμε σε ποια από τις τρεις αυτές κατηγορίες ανήκει. Για να δούμε όμως κατά πόσο ένας γράφος μοιάζει με έναν από τους τρεις συνολικούς, πρέπει να ορίσουμε κάποιες συγκεκριμένες μετρικές απόστασης. Αυτές οι μετρικές ορίζονται επίσης από τους Fotis Aisopos et. al. (2012) [4] ως Containment Similarity (CS), Value Similarity (VS) και Normalized Value Similarity (NVS), οι οποίες θα αναλυθούν στη συνέχεια, παρουσιάζοντας και τη χρησιμότητά τους αλλά και τους μαθηματικούς τους τύπους.

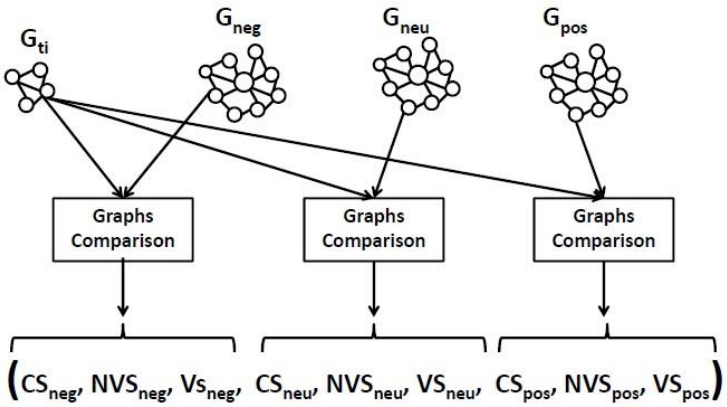
Η Containment Similarity ελέγχει κατά πόσο ο γράφος που δοκιμάζουμε περιέχεται στον συνολικό γράφο. Αν λοιπόν θεωρήσουμε Gu τον συνολικό γράφο και Gi τον εξεταζόμενο γράφο τότε η CS περιγράφετε από τον τύπο $\mathrm{CS}(\mathrm{Gu}, \mathrm{Gi}) = \sum_{\mathrm{e} \in \mathrm{Gi}} (\mu(\mathrm{e}, \mathrm{Gu}) / \min(|\mathrm{Gu}|, |\mathrm{Gi}|))$, όπου e είναι μία ακμή, μ(e,Gu) είναι ίσο με ένα αν η ακμή υπάρχει στο Gu και 0 αν δεν υπάρχει και |G| είναι ο αριθμός ακμών του γράφου G.

Η Value Similarity εξετάζει πάλι τον αριθμό κοινών ακμών αλλά αυτήν τη φορά λαμβάνει υπόψιν και το βάρος αυτών των ακμών. Έτσι, αν παρομοίως θεωρήσουμε Gu τον συνολικό γράφο και Gi τον εξεταζόμενο γράφο, τότε η VS περιγράφεται από τον τύπο $VS(Gu, Gi) = \frac{\sum_{e \in Gi} \frac{min(W(e,Gu), W(e,Gi))}{max(W(e,Gu), W(e,Gi))}}{max(|Gu|, |Gi|)}$, όπου e είναι μία ακμή, W(e,G) είναι το βάρος της ακμής e στον γράφο G και |G| είναι ο αριθμός ακμών του γράφου G. Το μέγεθος αυτής της απόστασης είναι κοντά στο 1 για γράφους που είναι σχεδόν πανομοιότυποι.

Η Normalized Value Similarity είναι ίδια με την Value Similarity μόνο που αυτήν τη φορά η απόσταση ανεξαρτητοποιείται από τα μεγέθη των δύο γράφων. Αυτή η ανεξαρτητοποίηση μας

παρέχει μία μετρική η οποία μπορεί να χρησιμοποιηθεί σε ακόμα περισσότερα προβλήματα χωρίς να επηρεάζεται από στοιχεία που θα μπορούσαν να προσθέτουν θόρυβο στα αποτελέσματα της μέτρησης των αποστάσεων. Έτσι για να βρούμε την NVS αρκεί ο εξής τύπος: $NVS(\text{Gu}, \text{Gi}) = \frac{\text{VS(Gu,Gi)}}{\text{SS(Gu,Gi)}}$, όπου SS(Gu,Gi) = min(|Gu|,|Gi|) / max(|Gu|,|Gi|).

Έτσι κάθε νέος γράφος μπορεί πλέον να αναπαρασταθεί από ένα σύνολο εννέα αριθμητικών μεγεθών. Όπως φαίνεται στην Εικόνα 3, κάθε νέος γράφος (G_{ti}) συγκρίνεται και με τους τρεις συνολικούς γράφους (G_{neg}, G_{neu} και G_{pos}) και κάθε σύγκριση παράγει τρείς διαφορετικές τιμές αποστάσεων. Αυτές οι εννέα τιμές αποτελούν το διάνυσμα που αντιστοιχεί στον συγκεκριμένο γράφο και θα χρησιμοποιηθεί για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης καθώς και για τη διαδικασία της κατηγοριοποίησης.



Εικόνα 3: Η διαδικασία σύγκρισης. πηγή: Fotis Aisopos et. al. (2012)

Για λόγους πρακτικότητας οι συνολικοί γράφοι έπρεπε να "κλαδευτούν", μία διαδικασία πιο γνωστή ως pruning. Αν δεν εκτελούσαμε αυτήν τη διαδικασία τα αποτελέσματα θα ήταν σαφώς βελτιωμένα αλλά δεν θα τα βλέπαμε πρακτικά ποτέ καθώς οι αλγόριθμοι θα έτρεχαν για πολύ μεγάλα και άρα μη πρακτικά χρονικά διαστήματα. Έτσι, σε κάθε πείραμα ορίζαμε ένα κατώφλι το οποίο χρησιμοποιούνταν στο κούρεμα του γράφου. Αν μία ακμή είχε βάρος μικρότερο από το κατώφλι τότε διαγραφόταν από τον συνολικό γράφο και άρα δεν λαμβάνονταν υπόψη στον υπολογισμό των διανυσμάτων. Στην παρούσα έρευνα χρησιμοποιήσαμε δύο κατώφλια, το 0.01 και το 0.001 τα οποία επιλέχθηκαν ώστε να δείξουν πώς επηρεάζει ο βαθμός λεπτομέρειας του γράφου το τελικό αποτέλεσμα.

Εκπαιδεύοντας τους επτά αλγορίθμους, που είδαμε και παραπάνω, με αυτά τα διανύσματα τους δώσαμε τη δυνατότητα να διακρίνουν τα όρια μέσω των οποίων ένα νέο διάνυσμα μπορούσε να κατηγοριοποιηθεί ως θετικό, ουδέτερο ή αρνητικό. Κάθε μία από τις τιμές του διανύσματος συγκρίνονταν με τις τιμές που ο αλγόριθμος ήδη γνώριζε από την εκπαίδευσή του και μας δίνονταν ως αποτέλεσμα τρεις πιθανότητες: η πιθανότητα αυτός ο γράφος να είναι θετικής πολικότητας, η πιθανότητα να είναι αρνητικής και η πιθανότητα να είναι ουδέτερης. Φυσικά το άθροισμα των τριών αυτών έβγαινε πάντα 1, δείχνοντας ότι ένας γράφος θα έμπαινε σε τουλάχιστον μία από αυτές τις κατηγορίες.

3.3.4 Combinational

Τα συνδυαστικά (Combinational) πειράματα αποτελούν την καινοτομία της παρούσας έρευνας και θα παρουσιαστούν λεπτομερώς στη συνέχεια. Η βασική ιδέα αυτών των πειραμάτων είναι να χρησιμοποιηθεί μία μορφή συμψηφισμού των αποτελεσμάτων και των επτά αλγορίθμων έτσι ώστε να καταλήξουν σε μία κοινή απόφαση όσον αφορά τη διαδικασία της κατηγοριοποίησης. Κατά την αναζήτησή μας στη διαθέσιμη βιβλιογραφία δεν συναντήσαμε κάποια σχετική προσπάθεια.

Σε αυτά τα πειράματα χρησιμοποιήσαμε και τη μέθοδο των N-Grams αλλά και την μέθοδο των N-Gram Graphs για την επεξεργασία του φυσικού λόγου. Ως μήκος ψευδολέξεων χρησιμοποιήσαμε πάλι τους αριθμούς 3,4 και 5 έτσι ώστε να υπάρχει κάποια ομοιομορφία με τα προηγούμενα πειράματα για λόγους σύγκρισης της απόδοσής τους. Στην ουσία η συνδυαστική απόφαση πήρε τον ρόλο των αλγορίθμων μηχανικής μάθησης στα προηγούμενα πειράματα.

Όπως και στα υπόλοιπα, έτσι και σε αυτά τα πειράματα χρησιμοποιήσαμε την 10-fold cross validation τεχνική για να δοκιμάσουμε την απόδοση. Αυτήν τη φορά όμως η βιβλιοθήκη της Weka δεν μπορούσε να καλύψει τις ανάγκες μας με τις έτοιμες μεθόδους της για την validation διαδικασία. Αν και η Weka παρέχει τη δυνατότητα για crossvalidation, η μέθοδός της έχει σχεδιαστεί για να εφαρμόζεται σε μόνο έναν αλγόριθμο μηχανικής μάθησης ενώ εμείς χρειαζόμασταν επτά αλγορίθμους ταυτόχρονα. Παρ' όλα αυτά μπορούσαμε ακόμα να χρησιμοποιήσουμε τις μεθόδους της βιβλιοθήκης για να διασπάσουμε το σύνολο δεδομένων σε δέκα semi-randomized υποσύνολα με ασφάλεια.

Έτσι, κρίθηκε απαραίτητο να δημιουργηθεί μία εξειδικευμένη λύση σε Java, η οποία θα διασπούσε τα διαθέσιμα δεδομένα σε δέκα ίσα κομμάτια με τη βοήθεια της Weka, θα εκπαίδευε τους επτά αλγορίθμους χρησιμοποιώντας τα εννέα από αυτά και θα τους δοκίμαζε με το δέκατο. Για κάθε ένα tweet που δοκιμάζεται από τους επτά αλγορίθμους θα αποφασίζαμε την κατηγορία στην οποία ανήκει με τον αλγόριθμο που θα παρουσιάσουμε παρακάτω. Στη συνέχεια θα συνέκρινε την αποφασισμένη κατηγορία με την πραγματική κατηγορία του tweet, την οποία γνωρίζουμε αφού το tweet είναι μέρος του συνόλου εκπαίδευσης, και αν ήταν ίδια θα το θεωρούσε επιτυχημένο. Τέλος θα έβγαζε ένα ποσοστό επιτυχίας.

Αυτός ο αλγόριθμος υλοποιήθηκε και έτρεχε 10 φορές για κάθε πείραμα με κυλιόμενα σύνολα δεδομένων, έτσι ώστε κάθε υποσύνολο να έχει την ευκαιρία να γίνει το σύνολο δοκιμής. Στη συνέχεια τα ποσοστά επιτυχίας από κάθε τρέξιμο του αλγορίθμου αθροίστηκαν και διαιρέθηκαν διά του δέκα έτσι ώστε να παραχθεί ένα μέσο ποσοστό επιτυχίας. Αυτό το ποσοστό

ήταν και το μέτρο σύγκρισης της επιτυχίας των συνδυαστικών πειραμάτων σε σχέση με τα προηγούμενα πειράματα που τρέξαμε.

Ο αλγόριθμος που αποφάσιζε σε ποια από τις τρεις κατηγορίες άνηκε κάθε tweet έπρεπε να αναπτυχθεί επίσης από την αρχή. Αν και αρχικά δοκιμάστηκαν απλές μέθοδοι, όπως η μέθοδος της πλειοψηφίας ή του μέσου όρου, στη συνέχεια για την ανάπτυξή του χρησιμοποιήθηκαν διάφοροι τύποι αποστάσεων έτσι ώστε να αποφασίσουμε κατά πόσο η άποψη του κάθε αλγορίθμου πρέπει να συμβάλει στην τελική απόφαση.

3.3.4.1 Πλειοψηφία

Κατά τα πειράματα που χρησιμοποιούσαν πλειοψηφία για τη λήψη αποφάσεων το μόνο που είχαμε να κάνουμε ήταν να αθροίσουμε τις αποφάσεις των επτά αλγορίθμων, χωρισμένες κατά κατηγορία. Έτσι η κατηγορία με τους περισσότερους "ψήφους" ήταν και η κατηγορία στην οποία άνηκε το συγκεκριμένο tweet. Συμπερασματικά, αυτή είναι μία αρκετά απλοποιημένη μέθοδος, δεδομένου ότι δεν λαμβάνει υπόψιν την ισχύ κάθε αλγορίθμου και το πόσο "σίγουρος" είναι για την ψήφο του. Γενικά ένας αλγόριθμος θεωρούμε ότι είναι σίγουρος για την ψήφο του αν η ισχυρότερη πιθανότητα από τις τρεις που παράγει είναι αρκετά μεγαλύτερη από τις άλλες δύο.

3.3.4.2 Μέσος Όρος

Σε αυτήν τη περίπτωση έχουμε μία μικρή βελτίωση στη λήψη αποφάσεων καθώς τώρα λαμβάνεται υπόψιν σε κάποιο βαθμό και η βεβαιότητα του κάθε αλγορίθμου. Κατά τη μέθοδο αυτή αθροίζουμε όλες τις πιθανότητες που παρήγαγαν οι επτά αλγόριθμοι για κάθε μία από τις τρεις κατηγορίες ξεχωριστά και στη συνέχεια διαιρούμε κάθε άθροισμα διά επτά, παράγοντας έτσι έναν μέσο όρο της πιθανότητας να ανήκει το εξεταζόμενο tweet σε κάθε μία από τις κατηγορίες. Η μεγαλύτερη πιθανότητα από αυτές τις τρεις μας δείχνει την κατηγορία στην οποία ανήκει το tweet και μάλιστα μπορούμε πλέον να δούμε και τον βαθμό βεβαιότητας της απόφασης.

3.3.4.3 Αποστάσεις

Σε αυτά τα πειράματα χρησιμοποιήσαμε πάλι την τεχνική των μέσων όρων για τον υπολογισμό των τριών πιθανοτήτων αλλά αυτήν τη φορά για τη λήψη της τελικής απόφασης πλησιάσαμε περισσότερο στη λογική των clusters. Για κάθε ένα από τα σύνολα θετικών, ουδέτερων ή αρνητικών γράφων δημιουργήθηκε ένα cluster πιθανοτήτων στον τρισδιάστατο χώρο, όπου κάθε διάσταση συμβόλιζε την πιθανότητα το tweet να είναι αρνητικό, θετικό ή ουδέτερο. Στη συνέχεια υπολογίστηκε το centroid κάθε ενός από αυτά τα clusters. Το centroid είναι το κεντρικό σημείο του cluster, το οποίο δεν είναι απαραίτητο να είναι ένα από τα πραγματικά σημεία του.

Αν τώρα αναπαραστήσουμε τον μέσο όρο που παράγεται σύμφωνα με την παραπάνω παράγραφο ως ένα σημείο σε αυτό το τρισδιάστατο χώρο, μπορούμε να υπολογίσουμε την απόστασή του από κάθε ένα από τα centroids. Όπως είναι φυσικό, το centroid που είναι πιο κοντά στο σημείο θα κρίνει το cluster στο οποίο ανήκει και άρα την κατηγορία του tweet. Για τον υπολογισμό αυτής της απόστασης χρησιμοποιήθηκαν οι αποστάσεις που βλέπουμε παρακάτω.

3.3.4.3.1 Ορθοδρομική απόσταση

Πρόκειται για τον ορισμό της απόστασης σε ένα σφαιρικό σύστημα. Χρησιμοποιείται κυρίως για να μετρήσει αποστάσεις πάνω σε πλανήτες ή σφαιρικούς χώρους και χωρίζεται σε τρία είδη, την απόσταση ημιτόνου, συνημίτονου και εφαπτομένης. Εμείς στα πειράματά μας χρησιμοποιήσαμε και τα τρία ήδη της σε ανεξάρτητα πειράματα για λόγους σύγκρισης των αποδόσεών τους.

Οι μαθηματικοί τύποι για τον υπολογισμό τους είναι οι εξής:

Ημιτόνου: $D_{ortho(cos)} = \arccos(n1 \cdot n2)$

Συνημίτονου: $D_{ortho(sin)} = \arcsin(|n1 \times n2|)$

Εφαπτομένης: $D_{ortho(tan)} = \arctan((|n1 \times n2|) / (n1 \cdot n2))$

όπου $n1 \cdot n2$ το εσωτερικό γινόμενο των δύο διανυσμάτων και $n1 \times n2$ το εξωτερικό τους γινόμενο.

3.3.4.3.2 Ευκλείδεια απόσταση

Η ευκλείδεια απόσταση αποτελεί τον κλασσικό τύπο της απόστασης που χρησιμοποιούμε αρκετά συχνά. Μετράει την απόσταση μεταξύ δύο σημείων σε έναν χώρο ορθοκανονικών αξόνων, στους οποίους αν x_i και x_j είναι τα δύο σημεία και k οι διαστάσεις του χώρου, τότε συμβολίζεται με τον γνωστό τύπο που ακολουθεί:

3.3.4.3.3 Manhattan

Η απόσταση Manhattan είναι θεί3 θ

$$D_{Manhattan}(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$

3.3.4.3.4 Chebychev

Η απόσταση Chebychev είναι πιο γνωστή ως chess distance ή απόσταση σκακιέρας. Αφού υπολογίσει τις αποστάσεις κάθε διάστασης των σημείων ξεχωριστά μας δίνει ως απόσταση τη μεγαλύτερη από αυτές, θεωρώντας ότι η μεγαλύτερη απόσταση είναι η σημαντικότερη. Ο τύπος υπολογισμού της είναι ο εξής:

$$D_{Chebychev}(x_i, x_j) = \max_k |x_{ik} - x_{jk}|$$

3.3.4.3.5 Ομοιότητα Συνημίτονου

Η ομοιότητα συνημίτονου, ή Cosine similarity όπως είναι πιο γνωστή, αποτελεί μία μετρική που μας δείχνει πόσο διαφορετικό είναι ένα διάνυσμα από ένα άλλο. Βασικό στοιχείο της είναι το συνημίτονο και χρησιμοποιείται ευρέως σε εφαρμογές ανάλυσης δεδομένων και κατηγοριοποίησης. Δεδομένου ότι αποτελεί μέτρο ομοιότητας, μας δείχνει πόσο κοντά είναι το εξεταζόμενο σημείο σε κάποιο cluster, έχοντας μία αντιστρόφως ανάλογη σχέση με την απόσταση. Ο μαθηματικός της τύπος είναι ο εξής:

$$S_{Cosine}(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\|x_i\| \cdot \|x_j\|} = \sum_{k=1}^{d} \frac{x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{d} x_{ik}^2} \cdot \sqrt{\sum_{k=1}^{d} x_{jk}^2}}$$

3.4 Αποτελέσματα

Το σύνολο δεδομένων που τελικά συγκεντρώσαμε για την εκπαίδευση των αλγορίθμων αποτελούνταν τελικά από 4451 tweets. Από αυτά τα 1203 ήταν θετικά βαθμολογημένα, τα 1313 ήταν ουδέτερα, ενώ τα 1935 ήταν αρνητικά. Η βαθμολόγησή τους έγινε από διάφορους ερευνητές, χωρίς κάποιο αυτοματοποιημένο σύστημα. Για να εξασφαλιστεί η αντικειμενικότητα σε αυτήν τη βαθμολόγηση, κάθε tweet είχε βαθμολογηθεί από 3 έως και 5 ερευνητές, χωρίς να γνωρίζει ο καθένας τη βαθμολόγηση που έδινε ο άλλος. Έτσι μπορούσαμε απλά να ακολουθήσουμε την πλειοψηφία των βαθμολογιών για να βρούμε σε ποια κατηγορία άνηκε κάθε tweet.

3.4.1 Bag of Words

Δυστυχώς αυτές οι τεχνικές δεν κατάφεραν να φτάσουν σε ποσοστά επιτυχίας συγκρίσιμα με των υπολοίπων τεχνικών. Στα πρώτα πειράματα που επιχειρήσαμε χρησιμοποιήθηκε η απλουστευμένη έκδοση της μεθόδου, κατά την οποία το κείμενο χωρίζεται σε λέξεις και κάθε λέξη αντιστοιχίζεται με την αξία της ενώ στη συνέχεια αθροίζοντας αυτές τις αξίες οδηγούμαστε

στην αξία του συνολικού κειμένου. Τα ποσοστά επιτυχίας αυτής της μεθόδου δεν κατάφεραν να ξεπεράσουν το 40%, μένοντας πολύ κοντά στο όριο τυχαιότητας, δηλαδή το 33.33%.

Για να βελτιωθεί αυτό το ποσοστό χρησιμοποιήσαμε τρεις από τους αλγορίθμους μηχανικής μάθησης που μας ήταν διαθέσιμοι μέσω της βιβλιοθήκης της Weka. Τα αποτελέσματα και πάλι δεν ήταν τόσο αποτελεσματικά όσο των υπόλοιπων μεθόδων επεξεργασίας φυσικού λόγου αλλά τώρα καταφέραμε να ξεπεράσουμε το φράγμα του 40%. Κάθε ένας από τους αλγορίθμους εκπαιδευόταν στα αθροίσματα των αξιών κάθε λέξης έτσι ώστε να παράγει κάποια όρια. Αυτά τα όρια θα βοηθούσαν τον αλγόριθμο να διακρίνει αν ένα νέο κείμενο είναι θετικής, αρνητικής ή ουδέτερης πολικότητας.

Πιο συγκεκριμένα, χρησιμοποιήσαμε τους αλγορίθμους C4.5, Naïve Bayesian Networks και Support Vector Machines, οι οποίοι μας έδωσαν ένα μέσο ποσοστό επιτυχίας του 44.84%, 45.07% και 43.70% αντίστοιχα. Αυτά τα ποσοστά υπολογίστηκαν χρησιμοποιώντας τον 10-fold crossvalidator που παρέχεται από τη βιβλιοθήκη της Weka. Η βελτίωση είναι εμφανής αλλά και πάλι είναι κάτω από το 50% που διαισθητικά θεωρείται ένα αξιόπιστο όριο.

3.4.2 3-Grams

Με λίγο καλύτερα αποτελέσματα από την Bag of Words, αυτή η μέθοδος πλησιάζει ακόμα περισσότερο το 50% αλλά δυστυχώς δεν καταφέρνει να το φτάσει. Σε αυτά τα πειράματα χρησιμοποιήθηκαν κι οι επτά αλγόριθμοι δίνοντας κατά μέσο όρο αποτελέσματα κοντά στο 46%. Συγκεκριμένα χρησιμοποιήσαμε τους αλγορίθμους C4.5 με 46.29% επιτυχία, Naïve Bayesian Networks με 46.65%, Support Vector Machines με 45.85%, Logistic Regression με 46.80%, Multilayer Perceptrons με 46.68%, Best-First Trees με 46.67% και Functional Trees με 46.75%.

Σε δεύτερη φάση εξετάσαμε αν οι αλγόριθμοι μηχανικής μάθησης ενισχύονται από ένα πιο ισορροπημένο σύνολο εκπαίδευσης. Έτσι μειώσαμε το σύνολο στα 3609 tweets, από τα οποία 1203 tweets είναι σε κάθε μία από τις τρεις κατηγορίες. Με αυτό το μειωμένο σύνολο τα αποτελέσματα πραγματικά βελτιώθηκαν, ξεπερνώντας το 50% και φτάνοντας σε αποδόσεις γύρω από το 52%. Συγκεκριμένα έχουμε τους αλγορίθμους C4.5 με 51.64% επιτυχία, Naïve Bayesian Networks με 52.02%, Support Vector Machines με 51.38%, Logistic Regression με 52.19%, Multilayer Perceptrons με 52.16%, Best-First Trees με 52.05% και Functional Trees με 52.10%.

3.4.3 4-Grams

Με την αύξηση του μεγέθους των λέξεων έχουμε και μία αρκετά σημαντική αύξηση της απόδοσης των αλγορίθμων μηχανικής μάθησης. Στην περίπτωση των 4-Grams καταφέραμε να κατηγοριοποιήσουμε τα tweets με επιτυχία που ξεπερνούσε το 50%. Συγκεκριμένα, εξετάσαμε τους αλγορίθμους C4.5 με 54.25% επιτυχία, Naïve Bayesian Networks με 54.30%, Support Vector Machines με 54.19%, Logistic Regression με 54.36%, Multilayer Perceptrons με 54.23%, Best-First Trees με 54.19% και Functional Trees με 54.22%.

Και σε αυτήν τη φάση δοκιμάσαμε να ισοσταθμήσουμε τον καταμερισμό των tweets μεταξύ των τριών κατηγοριών καταλήγοντας σε ένα μικρότερο αλλά πιο αποδοτικό σύνολο εκπαίδευσης. Ακόμα μία φορά αυτή η μείωση βοήθησε στη διαδικασία της κατηγοριοποίησης, αυξάνοντας τα ποσοστά επιτυχίας. Συγκεκριμένα, έχουμε τους αλγορίθμους C4.5 με 64.69% επιτυχία, Naïve Bayesian Networks με 65.01%, Support Vector Machines με 64.64%, Logistic Regression με 65.21%, Multilayer Perceptrons με 65.15%, Best-First Trees με 65.10% και Functional Trees με 65.20%.

3.4.4 5-Grams

Αυξάνοντας το μήκος των ψευδολέξεων κατά ένα ακόμα γράμμα καταφέραμε να υψώσουμε τα ποσοστά επιτυχίας σχεδόν στο διπλάσιο του ποσοστού τυχαίας επιλογής. Με ποσοστά να φτάνουν περίπου στο 64% η ανάλυση με τη χρήση αυτής της μεθόδου επεξεργασίας φυσικού λόγου αποτελεί μία γρήγορη, απλή και αρκετά αξιόπιστη λύση για εφαρμογές που δεν απαιτούν πολύ ακρίβεια στην κατηγοριοποίηση. Συγκεκριμένα, έχουμε τους αλγορίθμους C4.5 με 64.09% επιτυχία, Naïve Bayesian Networks με 64.47%, Support Vector Machines με 63.68%, Logistic Regression με 64.74%, Multilayer Perceptrons με 64.72%, Best-First Trees με 64.65% και Functional Trees με 64.78%.

Εξισορροπώντας την κατάτμηση του συνόλου σε αυτήν την οικογένεια πειραμάτων καταφέραμε να φτάσουμε σε ποσοστά επιτυχίας αξιοποιήσιμα και από τις λίγο πιο απαιτητικές εμπορικές εφαρμογές. Τα ποσοστά επιτυχίας περιστρέφονταν πλέον γύρω από το 75.5% σε μία εκπληκτική αύξηση κατά περίπου 10%. Συγκεκριμένα, χρησιμοποιηθήκαν οι αλγόριθμοι C4.5 με 75.55% επιτυχία, Naïve Bayesian Networks με 75.75%, Support Vector Machines με 75.10%, Logistic Regression με 75.88%, Multilayer Perceptrons με 75.84%, Best-First Trees με 75.78% και Functional Trees με 75.85%.

3.4.5 3-Gram graphs

Αυτή η οικογένεια πειραμάτων επηρεάζεται όχι μόνο από το σύνολο δεδομένων και το μήκος των N-Grams αλλά και από το κατώφλι που ορίζουμε για το κλάδεμα του γράφου. Εμείς

επιλέξαμε δύο τιμές κατωφλιού απλά για να εξετάσουμε τις επιπτώσεις της αύξησης ή της μείωσής του στην απόδοση αλλά και στον χρόνο εκτέλεσης της ανάλυσης. Με τον δεκαπλασιασμό του κατωφλιού, από 0.001 σε 0.01, παρατηρήσαμε μείωση του χρόνου εκτέλεσης κατά 70% ή ακόμα και περισσότερο σε κάποιες περιπτώσεις αλλά το ποσοστό επιτυχίας των αλγορίθμων μειώθηκε κατά περίπου 30%.

Συγκεκριμένα, για κατώφλι ίσο με 0.01 έχουμε τους αλγορίθμους C4.5 με 60.11% επιτυχία, Naïve Bayesian Networks με 56.08%, Support Vector Machines με 59.80%, Logistic Regression με 66.46%, Multilayer Perceptrons με 66.14%, Best-First Trees με 58.48% και Functional Trees με 65.64%. Αυξάνοντας την ακρίβεια του δέντρου ορίζοντας το κατώφλι στο 0.001 έχουμε τους C4.5 με 92.28% επιτυχία, Naïve Bayesian Networks με 88.95%, Support Vector Machines με 90.67%, Logistic Regression με 94.10%, Multilayer Perceptrons με 93.99%, Best-First Trees με 92.34% και Functional Trees με 94.03%.

Για λόγους πληρότητας εξετάσαμε και αυτήν την οικογένεια πειραμάτων με το μικρότερο, εξισορροπημένο σύνολο αλλά εδώ τα αποτελέσματα έδειχναν κάτι τελείως διαφορετικό. Αντίθετα με τη βελτίωση που παρατηρήσαμε στα απλά N-Grams, εδώ η απόδοση έμεινε σχετικά σταθερή και μάλιστα υπήρχαν ακόμα και μειωμένα ποσοστά επιτυχίας σε σχέση με το μεγαλύτερο σύνολο δεδομένων. Συγκεκριμένα, για κατώφλι 0.01 είχαμε τους αλγορίθμους C4.5 με 59.60% επιτυχία, Naïve Bayesian Networks με 55.05%, Support Vector Machines με 66.72%, Logistic Regression με 66.73%, Multilayer Perceptrons με 66.58%, Best-First Trees με 59.31% και Functional Trees με 66.08%.

3.4.6 4-Gram graphs

Πάλι σε αντίθεση με τα κλασικά N-Grams, η αύξηση του μήκους των ψευδολέξεων δεν οδήγησε σε υψηλότερα ποσοστά επιτυχίας. Αντίθετα σε αρκετούς από τους αλγορίθμους μείωσε τα ποσοστά επιτυχίας κατά λίγες μονάδες. Στις περισσότερες των περιπτώσεων όμως τα μέσα ποσοστά έμειναν σταθερά. Ακόμα και όταν ρίχναμε το κατώφλι στο 0.001, τα ποσοστά είχαν κάποια βελτίωση σε σχέση με τα 3-Gram Graphs αλλά ήταν μία σχεδόν αμελητέα βελτίωση.

Συγκεκριμένα, με κατώφλι 0.01 έχουμε τους αλγορίθμους C4.5 με 58.86% επιτυχία, Naïve Bayesian Networks με 55.71%, Support Vector Machines με 57.96%, Logistic Regression με 64.31%, Multilayer Perceptrons με 64.10%, Best-First Trees με 58.49% και Functional Trees με 63.04%, ενώ με το κατώφλι στο 0.001 έχουμε τους C4.5 με 93.01% επιτυχία, Naïve Bayesian Networks με 88.72%, Support Vector Machines με 91.63%, Logistic Regression με 94.18%, Multilayer Perceptrons με 94.53%, Best-First Trees με 93.01% και Functional Trees με 94.20%.

Περνώντας στο εξισορροπημένο σύνολο δεδομένων παρατηρούμε και πάλι μία πτώση των ποσοστών επιτυχίας, αν και δεν είναι πολύ μεγάλη. Συγκεκριμένα, για κατώφλι 0.01 παρατηρούμε τους αλγορίθμους C4.5 με 56.92% επιτυχία, Naïve Bayesian Networks με 52.73%, Support Vector Machines με 63.09%, Logistic Regression με 65.33%, Multilayer Perceptrons με 64.81%, Best-First Trees με 56.82% και Functional Trees με 64.05%.

3.4.7 5-Gram graphs

Αυξάνοντας λίγο ακόμα το μήκος των ψευδολέξεων στα 5 γράμματα παρατηρούμε ότι τα ποσοστά επιτυχίας συνεχίζουν να μειώνονται και μάλιστα πλέον μειώνονται και για τις δύο τιμές κατωφλιού. Αυτό μας δείχνει ότι δεν έχει κάποιο νόημα να συνεχίσουμε να αυξάνουμε το μήκος των ψευδολέξεων καθώς με κάθε αύξηση θα οδηγούμαστε σε ακόμα χαμηλότερα ποσοστά επιτυχίας.

Συγκεκριμένα, για κατώφλι ίσο με 0.01 έχουμε τους αλγορίθμους C4.5 με 56.50% επιτυχία, Naïve Bayesian Networks με 55.04%, Support Vector Machines με 54.73%, Logistic Regression με 61.46%, Multilayer Perceptrons με 60.92%, Best-First Trees με 56.63% και Functional Trees με 60.25%. Ρίχνοντας την τιμή του κατωφλιού στο 0.001, αυξάνοντας έτσι την ακρίβεια και το μέγεθος των γράφων, έχουμε τους αλγορίθμους C4.5 με 92.21% επιτυχία, Naïve Bayesian Networks με 89.28%, Support Vector Machines με 90.28%, Logistic Regression με 93.66%, Multilayer Perceptrons με 93.86%, Best-First Trees με 92.33% και Functional Trees με 93.85%.

Αν τώρα χρησιμοποιήσουμε το ισομοιρασμένο σύνολο δεδομένων, παρατηρούμε κάποια ακόμα μικρότερα ποσοστά επιτυχίας, δείχνοντας για ακόμα μία φορά τη διαφορά που υπάρχει μεταξύ των απλών N-Grams και των γράφων. Έτσι, με κατώφλι 0.01 έχουμε τους αλγορίθμους C4.5 με 56.32% επιτυχία, Naïve Bayesian Networks με 49.09%, Support Vector Machines με 56.71%, Logistic Regression με 61.85%, Multilayer Perceptrons με 61.07%, Best-First Trees με 55.71% και Functional Trees με 59.90%.

3.4.8 Combinational experiments

Συνδυάζοντας τα αποτελέσματα των επτά αλγορίθμων επιχειρήσαμε να βελτιώσουμε τα ποσοστά επιτυχίας στην οικογένεια των N-Grams αλλά και στην οικογένεια των N-Gram Graphs. Αν και στις περισσότερες των περιπτώσεων τα αποτελέσματα δεν ήταν τόσο ενθαρρυντικά, υπήρχαν κάποιες οικογένειες πειραμάτων στις οποίες παρατηρήθηκε μία βελτίωση αρκετών μονάδων.

Η μεγαλύτερη βελτίωση που παρατηρήθηκε είναι στην οικογένεια των 5-Grams με το εξισορροπημένο σύνολο δεδομένων, όπου είχαμε τις τεχνικές Μέσου όρου με ποσοστό επιτυχίας

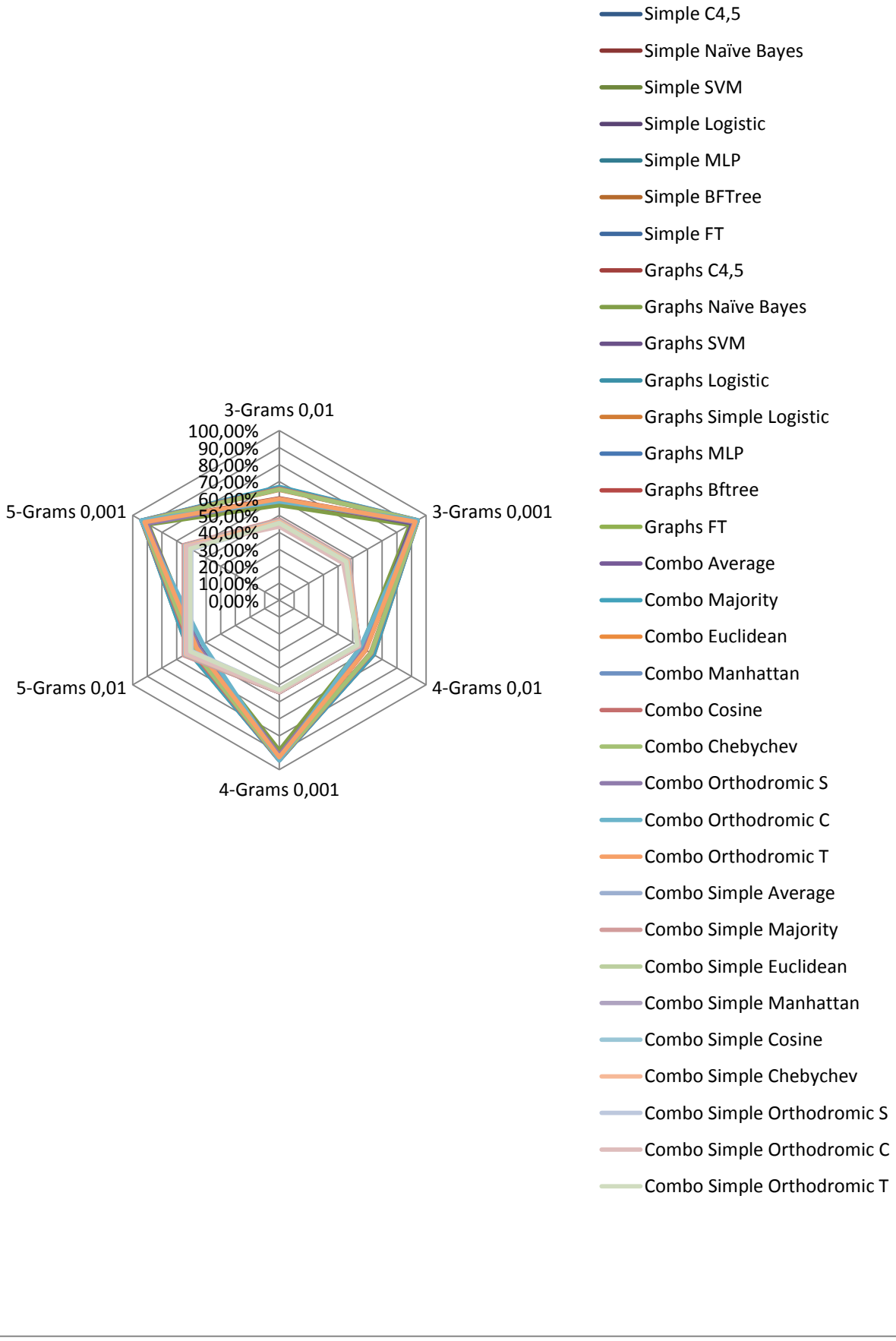
82.49%, Πλειοψηφίας με ποσοστό 83.15%, Ευκλείδειας απόστασης με ποσοστό 82.49%, απόστασης Manhattan με ποσοστό επίσης 82.49%, ομοιότητας συνημίτονου με ποσοστό 82.80%, απόστασης Chebychev με ποσοστό 82.49%, Ορθοδρομική απόσταση ημιτόνου με ποσοστό 78.54%, Ορθοδρομική απόσταση συνημίτονου με ποσοστό 82.14% και Ορθοδρομική απόσταση εφαπτομένης με ποσοστό 80.99%.

Αν συλλογιστούμε ότι το ανώτερο ποσοστό επιτυχίας που επιτύχαμε με τους παραδοσιακούς αλγορίθμους μηχανικής μάθησης στην ίδια οικογένεια πειραμάτων ήταν το 75.88% με τη λογιστική παλινδρόμηση, τότε έχουμε μία βελτίωση περίπου της τάξης του 7%. Αν και αυτή η βελτίωση φαίνεται μικρή, στην ουσία είναι απλά μία επιβεβαίωση ότι ο συνδυασμός των κλασσικών αλγορίθμων μηχανικής μάθησης για τη λήψη αποφάσεων κατηγοριοποίησης είναι μία τεχνική που πραγματικά μπορεί να βελτιώσει την αξιοπιστία τους.

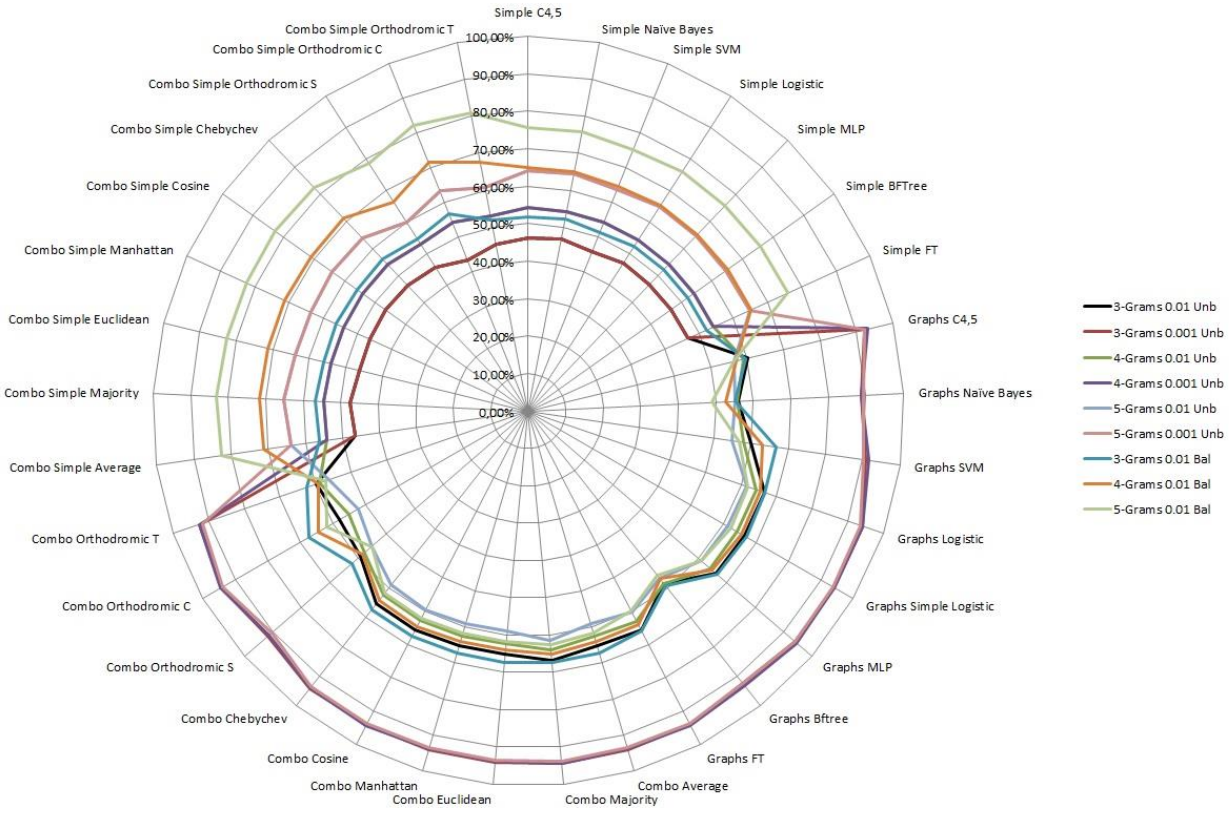
Βέβαια, λόγω του μεγάλου αριθμού παραμετροποιήσημων στοιχείων που εμπλέκονται σε μία τέτοια ανάλυση, απαιτείται αρκετή περαιτέρω έρευνα προκειμένου να αυξηθεί ακόμα περισσότερο η αξιοπιστία της κατηγοριοποίησης. Κάθε αλλαγή παραμέτρων μπορεί να βελτιώσει την απόδοση ενός από τους αλγορίθμους που χρησιμοποιούνται, μειώνοντας όμως την απόδοση των υπολοίπων, μειώνοντας έτσι και την απόδοση των συνδυαστικών αποφάσεων. Για αυτόν το λόγο απαιτείται προσεκτική μελέτη για κάθε νέα αλλαγή που εισάγεται στο σύστημα.

3.4.9 Διαγράμματα

Ο συνολικός όγκος των πειραμάτων φαίνεται καλύτερα στα παρακάτω γραφήματα. Το πρώτο γράφημα μας δείχνει ένα ξεκάθαρο διαχωρισμό μεταξύ των οικογενειών των N-Grams (εσωτερικός κύκλος) και των οικογενειών των N-Gram graphs (εξωτερικός κύκλος). Βλέπουμε ότι τα πειράματα με γράφους είναι σε κάθε περίπτωση καλύτερα από αυτά με κλασικά N-Grams εκτός ίσως από τις περιπτώσεις που χρησιμοποιήσαμε 5-Grams και ως κατώφλι την τιμή 0.01. Επίσης, γίνεται εμφανής η επίδραση του κατωφλιού στην απόδοση των αλγορίθμων που χρησιμοποιούν γράφους.


3-Grams 0,01
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
3-Grams 0,001
4-Grams 0,01
4-Grams 0,001
5-Grams 0,01
5-Grams 0,001
Simple C4,5
Simple Naïve Bayes
Simple SVM
Simple Logistic
Simple MLP
Simple BFTree
Simple FT
Graphs C4,5
Graphs Naïve Bayes
Graphs SVM
Graphs Logistic
Graphs Simple Logistic
Graphs MLP
Graphs Bftree
Graphs FT
Combo Average
Combo Majority
Combo Euclidean
Combo Manhattan
Combo Cosine
Combo Chebychev
Combo Orthodromic S
Combo Orthodromic C
Combo Orthodromic T
Combo Simple Average
Combo Simple Majority
Combo Simple Euclidean
Combo Simple Manhattan
Combo Simple Cosine
Combo Simple Chebychev
Combo Simple Orthodromic S
Combo Simple Orthodromic C
Combo Simple Orthodromic T

Στο δεύτερο γράφημα μπορούμε να δούμε λίγο πιο ξεκάθαρα τη δυνατότητα παραμετροποίησης των αλγορίθμων που χρησιμοποιούν τους γράφους. Στο πάνω ημισφαίριο του γραφήματος βλέπουμε όλους τους αλγορίθμους που χρησιμοποιούν απλά N-Grams, ενώ στο κάτω αυτούς που χρησιμοποιούν γράφους. Όπως είναι εμφανές, τα ποσοστά στο κάτω ημισφαίριο κυμαίνονται από ίσα, με του πάνω, μέχρι και αρκετά ανώτερά τους, ανάλογα με το κατώφλι που κάθε φορά ορίζουμε. Αυτό μας δίνει τη δυνατότητα να εκτιμούμε την ισορροπία μεταξύ ακρίβειας και χρόνου εκτέλεσης που απαιτείται από την εφαρμογή μας και να θέτουμε κάθε φορά το κατάλληλο κατώφλι.



4 Συμπεράσματα

Αν και ο τομέας του Sentiment Analysis έχει εξερευνηθεί αρκετά τα τελευταία χρόνια, όπως βλέπουμε οι πιθανοί συνδυασμοί τεχνικών και παραμέτρων είναι αρκετοί για να απασχολούν τους ερευνητές για πολλά ακόμα χρόνια. Στην παρούσα έρευνα προσπαθήσαμε να ανακαλύψουμε τον συνδυασμό που καταλήγει στη μεγαλύτερη δυνατή αξιοπιστία έτσι ώστε να μπορέσουμε να δημιουργήσουμε ένα ακριβές μοντέλο πρόβλεψης της άποψης του πληθυσμού για μία πολιτική.

Ίσως το σημαντικότερο στοιχείο σε μία τέτοια έρευνα είναι η επιλογή της μεθόδου επεξεργασίας φυσικού λόγου. Η αναπαράσταση του κειμένου σε μία μορφή επεξεργάσιμη από τους αλγορίθμους μηχανικής μάθησης είναι αυτό που αρκετές φορές θα κρίνει αν ένας αλγόριθμος θα μπορέσει να κατηγοριοποιήσει τα δεδομένα ή όχι. Σκεπτόμενοι ότι οι αλγόριθμοι αυτοί έχουν μαθηματικές βάσεις είναι εμφανές ότι δεν μπορούμε να μεταβάλουμε τη βασική τους λειτουργία έτσι ώστε να ενισχύσουμε δραματικά την απόδοσή τους.

Ο τομέας της επεξεργασίας φυσικού λόγου όμως διαφέρει με την έννοια ότι είναι ένα πιο αυθαίρετο πεδίο. Κάθε ερευνητής έχει την ελευθερία να αναπαραστήσει τα δεδομένα όπως θέλει και αυτή η ελευθερία είναι που κάνει τον τομέα του sentiment analysis τόσο ενεργό. Φυσικά και υπάρχουν τυποποιημένοι μέθοδοι αναπαράστασης, όπως αναφέραμε και παραπάνω, αλλά για να πετύχουμε μία πραγματική βελτίωση της αξιοπιστίας των αποτελεσμάτων μας πρέπει να δημιουργούμε όλο και καλύτερες αναπαραστάσεις, διευκολύνοντας έτσι τους αλγορίθμους μηχανικής μάθησης.

Μέσω των πολυάριθμων πειραμάτων καταλήξαμε στο συμπέρασμα ότι τα N-Gram graphs μπορούν να παραμετροποιηθούν έτσι ώστε να μας παρέχουν αξιοπιστία αρκετά ισχυρή για τις απαιτήσεις ενός τέτοιο μοντέλου πρόβλεψης. Το μόνο ελάττωμα είναι ότι πρέπει να εξισορροπήσουμε τον χρόνο που απαιτείται για την ολοκλήρωση της ανάλυσης, ο οποίος αυξάνεται αρκετά αν προσπαθήσουμε να αυξήσουμε την αξιοπιστία.

Τελικά, όποια τεχνική επεξεργασίας φυσικού λόγου και αλγόριθμο μηχανικής μάθησης και να χρησιμοποιήσουμε, αν θέλουμε να αυξήσουμε την αξιοπιστία των αποτελεσμάτων θα πρέπει να θυσιάσουμε την ταχύτητα απόκρισης των αλγορίθμων. Όσο πιο αξιόπιστη η ανάλυση τόσο πιο πολύ αργεί να ολοκληρωθεί, φτάνοντας σε κάποιες περιπτώσεις σε μη ρεαλιστικούς χρόνους εκτέλεσης.

5 Μελλοντικές κατευθύνσεις

Σε συνέχεια της παρούσας έρευνας θα μπορούσαν να αναλυθούν ακόμα περισσότερες μορφές δεδομένων. Θα μπορούσαν να χρησιμοποιηθούν ως πηγές άλλα κοινωνικά δίκτυα, όπως το facebook, το disquis ή ακόμα και τα YouTube και Instagram. Η ανάλυση της εικόνας και του video δεν είναι τόσο μακριά στο μέλλον όπως μας απέδειξαν οι Louis-Philippe Morency et. al. (2011) στο έργο τους. Με την προσθήκη νέων πηγών και νέων μορφών δεδομένων θα μπορούσαμε να δούμε σε ακόμα μεγαλύτερο εύρος πώς ανταποκρίνονται οι διάφοροι αλγόριθμοι που μελετήσαμε, παρουσιάζοντας έτσι μία ακόμα σφαιρικότερη εικόνα της τρέχουσας τεχνολογίας.

Σε ένα διαφορετικό επίπεδο θα μπορούσαν να βελτιωθούν οι ίδιοι οι αλγόριθμοι, τροποποιώντας κατάλληλα τις παραμέτρους τους έτσι ώστε να οδηγηθούμε σε ακόμα καλύτερα αποτελέσματα. Αυτές οι παράμετροι φυσικά είναι διαφορετικές για κάθε πιθανό πείραμα οπότε χρειάζεται αρκετή μελέτη και πάρα πολλές πειραματικές δοκιμές μέχρι να οδηγηθεί κάποιος σε αυτά τα βέλτιστα αποτελέσματα που επιδιώκει. Ειδικά στον τομέα των συνδυαστικών αποφάσεων, οι συνδυασμοί παραμέτρων είναι υπερβολικά πολλοί. Είναι τόσο πολλοί ώστε θα μπορούσαν να αναπτυχθούν ειδικοί αλγόριθμοι βελτιστοποίησης που να αποφασίζουν τις παραμέτρους ανάλογα με το πρόβλημα που πρέπει να αντιμετωπιστεί.

Τέλος, για να καταφέρουμε να ασχοληθούμε με ακόμα μεγαλύτερα σύνολα δεδομένων και ακόμα περισσότερα πειράματα θα χρειαστούμε τεχνικές που βελτιώνουν τους χρόνους αποκρίσεις αυτών των πειραμάτων σε επίπεδα σχεδόν πραγματικού χρόνου. Αυτό θα μπορούσε να γίνει με τη χρήση κατανεμημένων συστημάτων επεξεργασίας, όπως το Hadoop. Κάθε αλγόριθμος θα μπορούσε να κατανέμει το φόρτο εργασίας του σε ένα σύνολο μηχανημάτων, ρίχνοντας έτσι σε πολύ χαμηλά επίπεδα τον χρόνο που απαιτείται για την επεξεργασία.

6 Βιβλιογραφία

[1] Baum L. E., Petrie T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 1554-1563.

[2] Bo Pang, Lillian Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 1--135.

[3] Cortes C., Vapnik V. (1995). Support Vector Networks. *Machine Learning*, 273-297.

[4] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, Theodora A. Varvarigou (2012). Content vs. context for sentiment analysis: a comparative analysis over microblogs. *HT* (pp. 187-196). ACM.

[5] George H. John, P. Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338-345). San Mateo: Morgan Kaufmann.

[6] James N. Druckman, Thomas J. Leeper (2012). Is Public Opinion Stable? Resolving the Micro/Macro Disconnect in Studies of Public Opinion. *Dædalus, the Journal of the American Academy of Arts & Sciences*, 50-68.

[7] Louis-Philippe Morency, Rada Mihalcea, Payal Doshi (2011). Towards multimodal sentiment analysis: harvesting opinions from the web. *ICMI* (pp. 169-176). ACM.

[8] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena. (2007). Large-Scale Sentiment Analysis for News and Blogs. *International Conference on Weblogs and Social Media ICWSM.*

[9] Quinlan John Ross (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 77-90.

[10] R. McDonald, K. Hannan, T. Neylon, M. Wells, J. Reynar (2007). Structured Models for Fine-to-Coarse Sentiment Analysis.

[11] Theresa Wilson, Janyce Wiebe, Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *hltemnlp2005*, (pp. 347-354). Vancouver, Canada.

[12] Tony Mullen, Nigel Collier (2004). Sentiment Analysis Using Support Vector Machines with Diverse Information Sources., (pp. 412-418). Barcelona, Spain.

[13] Shi Haijian (2007). Best-first decision tree learning. University of Waikato.

[14] Diego Alejandro Salazar, Jorge Iván Vélez, Juan Carlos Salazar (2012). Comparison between SVM and Logistic. Revista Colombiana de Estadística, 223-237.