

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή εργασία

Θέμα:Διερεύνηση τεχνικών εξόρυξης γνώσης από ελληνικά σώματα κειμένων του κοινωνικού ιστού.

Δρουγκάνη Ανθή Α.Μ. Π2014104

Επιβλέπουσα: Κερμανίδου Κάτια

Τριμελής Επιτροπή: Κάτια Κερμανίδου , Φοίβος
Μυλωνάς , Θεμιστοκλής Έξαρχος

Περίληψη

Η πτυχιακή εργασία με θέμα « Διερεύνηση τεχνικών εξόρυξης γνώσης από ελληνικά σώματα κειμένων του κοινωνικού ιστού. » περιγράφει το πως οι τεχνικές εξόρυξης γνώσης εφαρμόζονται στον κοινωνικό ιστό και συγκεκριμένα σε αναρτήσεις που έχουν γραφεί στην ελληνική γλώσσα. Ο κοινωνικός ιστός αποτελεί ένα σύγχρονο μέσο επικοινωνίας και έκφρασης συναισθημάτων καθώς και προτιμήσεων. Χάρη στις τεχνικές εξόρυξης γνώσης μπορούμε να αντλήσουμε ένα σύνολο πληροφοριών που αφορούν τόσο τον κοινωνικό όσο και τον εμπορικό τομέα μέσα από τις δημοσιεύσεις των χρηστών στα λεγόμενα social media. Η παρούσα εργασία αποτελεί βιβλιογραφική έρευνα βασιζόμενη σε υπάρχουσες εργασίες. Κάθε μία από αυτές τις εργασίες περιγράφεται σε ξεχωριστό κεφάλαιο από τις υπόλοιπες. Αρχικά θα γίνει μία εισαγωγική παρουσίαση του θέματος, όπου αναλύονται βασικές έννοιες και μεθοδολογίες μηχανικής μάθησης που εφαρμόζονται στις έρευνες που επιλέχθηκαν. Στη συνέχεια παρουσιάζονται οι μελέτες αυτές ενώ αργότερα συγκεντρώνονται όλα τα πειράματα που πραγματοποιήθηκαν σε κάθε μία ,ώστε να μπορέσουμε να αντλήσουμε τα συμπεράσματά μας μέσα από μία αναλυτικότερη περιγραφή των αποτελεσμάτων τους με στόχο τη διευκόλυνση στη σύγκρισή τους. Τέλος παρουσιάζονται τα τελικά συμπεράσματα της εργασίας.

Λέξεις-Κλειδιά

Μηχανική Μάθηση , Ανάλυση Συναισθήματος, Εξόρυξη γνώσης
,Ελληνικός κοινωνικός ιστός

Key-words

Machine learning, Sentiment Analysis, Data Mining , Greek
social web

Περιεχόμενα

1. Εισαγωγή

1.1.Κοινωνικά Δίκτυα - online social networks (OSNs)

1.1.1 Τι είναι τα κοινωνικά Δίκτυα (OSNs)

1.2 Μηχανική μάθηση

1.2.1 Επιβλεπόμενη

1.2.2 Μη Επιβλεπόμενη

1.2.3Βαθιά μάθηση

1.3 Τι είναι η εξόρυξη γνώσης από σώματα κειμένων

1.4 Ανάλυση συναισθήματος και εξόρυξη γνώμης(sentiment analysis and opinion mining)

2 Ανάλυση εργασιών

2.1 Sentiment Analysis for Reputation Management: Mining the Greek Web

2.2Offensive Language Identification in Greek

2.3 Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features

2.4Authorship Attribution and Gender Identification in Greek Blogs

2.5 Nowcasting the Stance of Social Media Users in a Sudden Vote: The Case of the Greek Referendum

2.6Sentiment Analysis of Greek Tweets and Hashtags using Sentiment Lexicon

2.7A Deep Learning Approach to Aspect-Based Sentiment Prediction

2.8A Classifier to Distinguish Between Cypriot Greek and Standard Modern Greek

2.9Argument Extraction from News, Blogs, and the Social Web

2.10 A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets

2.11 Political sentiment analysis of tweets before and after the Greek elections of May 2012

3. Αποτελέσματα πειραμάτων εργασιών

4. Συμπεράσματα

Βιβλιογραφικές Αναφορές

1.Εισαγωγή

Στην εποχή που ζούμε ο κοινωνικός ιστός αποτελεί αναπόσπαστο κομμάτι της καθημερινότητας και της ζωής μας. Στον κοινωνικό ιστό(social media) συνεχώς αναρτώνται καινούργια δεδομένα και πληροφορίες που αφορούν την καθημερινότητα ,την κουλτούρα αλλά και τον εμπορικό τομέα .Συνεπώς είναι επιτακτική η ανάγκη ενός μέσου ,το οποίο εντοπίζει ,αναγνωρίζει, προβλέπει και κατηγοριοποιεί το τι ακριβώς έχει ειπωθεί στις αναρτήσεις των χρηστών. Αυτή την ανάγκη ικανοποιούν οι τεχνικές εξόρυξη γνώσης(data mining techniques) από σώματα κειμένων του κοινωνικού ιστού. Χάρη στη μηχανική μάθηση είναι εφικτή η εκπαίδευση αλγορίθμων και δεδομένων ,ώστε να εξαγάγουν γνώσεις και νέα δεδομένα. Αυτή η γνώση έχει μεγάλη σημασία ,καθώς διευκολύνει τόσο τη λήψη αποφάσεων σε απλά ζητήματά ενώ παράλληλα συνδράμει στην ανάπτυξη του επιχειρησιακού, εμπορικού αλλά και πολιτικού τομέα, καθώς προσφέρονται πληροφορίες που αφορούν τις πεποιθήσεις, τις απόψεις ,τις προτιμήσεις αλλά και τις καταναλωτικές συνήθειες των χρηστών. Συνεπώς δε θα είναι πλέον απαραίτητες οι δημοσκοπήσεις ,καθώς αντλούνται απλόχερα δεδομένα ,τα οποία εφόσον θα υποστούν τις απαραίτητες τροποποιήσεις ώστε να μπορούν να είναι κατανοητά από τους υπολογιστικούς αλγορίθμους ,είναι ικανά να εξαγάγουν ορθά και κατατοπιστικά πορίσματα που αφορούν όλους τους προαναφερθέντες κλάδους.

Η παρούσα εργασία αποτελεί βιβλιογραφική έρευνα. Σκοπός της είναι η ανάλυση τεχνικών εξόρυξης γνώσης που αφορούν σώματα κειμένου το ελληνικού κοινωνικού ιστού. Έχουν πραγματοποιηθεί αρκετές μελέτες που αφορούν τον κλάδο της μηχανικής μάθησης ,της ανάλυσης συναισθήματος και γενικότερα της εξόρυξης γνώσης. Ωστόσο η πλειοψηφία αυτών των ερευνών και συνεπώς των πειραμάτων που πραγματοποιούνται αφορούν κυρίως κείμενα ή αναρτήσεις στα κοινωνικά δίκτυα γραμμένα στην αγγλική γλώσσα. Για τις ανάγκες της συγκεκριμένης βιβλιογραφικής έρευνας συλλέχθηκαν μελέτες που αφορούν τον ελληνικό κοινωνικό ιστό. Αυτές οι μελέτες αποτελούν βασικό κομμάτι της εργασίας ,καθώς αφορούν υπαρκτά πειράματα που

υλοποιήθηκαν και συνδράμουν στη βαθύτερη κατανόηση των τεχνικών εξόρυξης γνώσης που περιγράφεται στο καθένα.

Με την ολοκλήρωση της εισαγωγής η εργασία διεκπεραιώνεται ως εξής: Αρχικά αναλύονται βασικοί όροι και έννοιες που αφορούν τις μεθοδολογίες και τα κοινωνικά δίκτυα που πραγματεύονται οι μελέτες που επιλέχθηκαν, προετοιμάζοντας και προσδίδοντας μία σφαιρική εικόνα στον αναγνώστη και ιδίως σε άτομα που δε σχετίζονται με το αντικείμενο. Στη συνέχεια περιγράφονται οι μελέτες που αφορούν εξόρυξη γνώσης από τον ελληνικό κοινωνικό ιστό, ενώ στο αμέσως επόμενο κεφάλαιο αναλύονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν στις μελέτες αυτές. Τέλος αναλύονται τα τελικά συμπεράσματα που εξήχθησαν από την εκπόνηση της εργασίας.

1.1. Κοινωνικά Δίκτυα - online social networks (OSNs)

1.1.1 Τι είναι τα κοινωνικά Δίκτυα (OSNs)

Ένα online κοινωνικό δίκτυο(OSN) είναι μία κοινωνική δομή που αποτελείται από άτομα ή οργανισμούς, που μπορούν να κληθούν ως κόμβοι-nodes.[α] Μεταξύ των κόμβων δημιουργούνται σύνδεσμοι ,οι οποίοι αποτελούν τους διαφορετικούς τύπους σχέσεων και αλληλεξάρτησης μεταξύ των κόμβων. Τέτοιοι σύνδεσμοι μπορεί να είναι η φιλία ή τα κοινά ενδιαφέροντα, ενώ αυτή η αλληλεπίδραση υφίσταται στα πλαίσια ηλεκτρονικής επικοινωνίας και κυρίως μέσω του Internet. Συμπεραίνουμε λοιπόν ότι ένα κοινωνικό δίκτυο βασίζεται σε δύο παραμέτρους του κόμβους και τους συνδέσμους. Το περιεχόμενο των κόμβων αφορά την κοινή πηγή ενδιαφέροντος ενώ οι σύνδεσμοι ανταλλάζουν τις πληροφορίες μεταξύ των κόμβων.

Τα κοινωνικά δίκτυα που χρησιμοποιούνται στις μελέτες που περιγράφονται στην εργασία είναι τα εξής:



Facebook: Το Facebook είναι ένας ιστοχώρος κοινωνικής δικτύωσης, όπου οι χρήστες μπορούν να επικοινωνούν μεταξύ τους, να αλληλεπιδρούν μέσω του like και σχολίων σε αναρτήσεις των επαφών τους. Δημιουργήθηκε το 2004 από τον Μαρκ Ζάκερμπεργκ ως μέλος του Πανεπιστημίου Χάρβαρντ. Ενώ αρχικά απευθυνόταν μόνο σε φοιτητές του Χάρβαρντ, το 2006 έγινε προσβάσιμο από όλο τον πλανήτη. Πλέον διαθέτει 2.6 δισεκατομμύρια ενεργούς χρήστες παγκοσμίως.



Twitter: Το Twitter, το οποίο είναι μία πλατφόρμα microblogging, αποτελεί έναν ιστοχώρο κοινωνικής δικτύωσης, όπου οι εγγεγραμμένοι χρήστες του μπορούν να στέλνουν και να διαβάζουν σύντομα μηνύματα. Αυτά τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν. Τα μηνύματα του Twitter ονομάζονται tweets και το κάθε ένα μπορεί να αποτελείται το πολύ από 280 χαρακτήρες. Δημιουργήθηκε το 2006 από τον Τζακ Ντόρσεϊ και σήμερα έχει περίπου 330 εκατομμύρια ενεργούς χρήστες, αποτελώντας έναν από τους πιο δημοφιλείς ιστοτόπους.



Instagram: Το Instagram είναι μία δωρεάν εφαρμογή κοινωνικής δικτύωσης που οι χρήστες μπορούν να μοιράζονται φωτογραφίες και βίντεο με τους ακολούθους τους (followers). Επιπλέον μπορούν να σχολιάζουν και να δηλώνουν ότι τους αρέσουν οι συγκεκριμένες δημοσιεύσεις. Πλέον υπάρχει και η δυνατότητα προσωπικής συνομιλίας (chat) με τους υπόλοιπους χρήστες. Δημιουργήθηκε από δύο απόφοιτους του Πανεπιστημίου του Στάνφορντ, τους Κέβιν Σίστρομ και Μάικ Κρίγκερ και ξεκίνησε τον Οκτώβριο του 2010. Μόλις δύο μήνες αργότερα, ο αριθμός των εγγεγραμμένων χρηστών έφτασε το

1.000.000.Το 2012 αγοράστηκε από το Facebook.Σήμερα μετράει περίπου 1 δισεκατομμύριο ενεργούς χρήστες.



Blogs: Τα Blogs αποτελούν μία μορφή ιστοχώρου, το περιεχόμενο του οποίου ποικίλει ανάλογα με το πως θέλει να το αξιοποιήσει ο κάτοχος του.Κάθε κοινός χρήστης του διαδικτύου μπορεί να ξεκινήσει το δικό του blog και να σχολιάσει ή να παρουσιάσει οποιοδήποτε θέμα θελήσει. Η ονομασία του αρχικά ήταν weblog αλλά καθιερώθηκε η ονομασία blog,καθώς ο συντάκτης αναφερόταν ως blogger.Το weblog ήταν επινόηση του Jorn Barger το 1997.



Tripadvisor: Το Tripadvisor είναι μία ταξιδιωτική εταιρία που παρουσιάζει αξιολόγησης ταξιδιωτικών προορισμών ,ξενοδοχείων και εστιατορίων. Επιπλέον διαθέτει διαδραστικά ταξιδιωτικά φόρουμ. Ιδρύθηκε από τον Στίβεν Κάουφερ το 2000.

1.2 Μηχανική μάθηση:

Το Διαδίκτυο διαθέτει έναν τεράστιο όγκο δεδομένων ενώ παράλληλα συνεχίζει να διευρύνεται διαρκώς. Λαμβάνοντας υπόψιν αυτό το γεγονός ,συμπεραίνουμε λοιπόν ότι η εύρεση συγκεκριμένων πληροφοριών και ο συγχρωτισμός του αποτελούν μία διαδικασία που δεν μπορεί να πραγματοποιηθεί με χειρονακτικό τρόπο. Συνεπώς κρίνεται απαραίτητη η ύπαρξη μίας μορφής τεχνητής νοημοσύνης η οποία θα μπορεί να εκπαιδεύεται και να συγκρατεί τη γνώση που έλαβε. Μια τέτοιας μορφής τεχνητή νοημοσύνη είναι η μηχανική μάθηση(machine learning).Η μηχανική μάθηση δε συσχετίζεται μόνο με βάσεις δεδομένων ,αλλά έχει την ικανότητα να βρίσκει λύσεις σε ζητήματα όπως η αναγνώριση εικόνας ή ήχου. Ουσιαστικά συμβάλει στην κατασκευή αλγορίθμων που μπορούν να μάθουν από τα δεδομένα και να τα χρησιμοποιήσουν για μελλοντικές προβλέψεις.

1.2.1 Επιβλεπόμενη: Η επιβλεπόμενη μηχανική μάθηση είναι η κατασκευή αλγορίθμων που μπορούν να παράγουν γενικά μοτίβα και υποθέσεις κάνοντας χρήση εξωτερικά εισαγόμενων επισημειωμένων δεδομένων για την πρόβλεψη της έκβασης μελλοντικών δεδομένων. Οι αλγόριθμοι κατηγοριοποίησης της επιβλεπόμενης μηχανικής μάθησης έχουν ως στόχο την κατηγοριοποίηση των δεδομένων που έλαβαν από προηγούμενες πληροφορίες[β].

Ονομάζεται επιβλεπόμενη, διότι η διαδικασία εκπαίδευσης ενός αλγορίθμου μπορεί να θεωρηθεί ως ένας δάσκαλος που επιβλέπει τη διαδικασία μάθησης. Ο αλγόριθμος επαναλαμβάνει τις προβλέψεις των δεδομένων ενώ αυτά διορθώνονται από τον επιβλέποντα. Η διαδικασία τερματίζει όταν ο αλγόριθμος φτάσει σε ένα αποδεκτό επίπεδο απόδοσης[δ]. Στην επιβλεπόμενη μάθηση τα προβλήματα χωρίζονται σε δύο κατηγορίες: Σε προβλήματα ταξινόμησης (Classification) και παλινδρόμησης (Regression).

Classification: Ένα πρόβλημα θεωρείται πρόβλημα κατηγοριοποίησης, όταν η μεταβλητή της εξόδου του είναι μία κατηγορία.

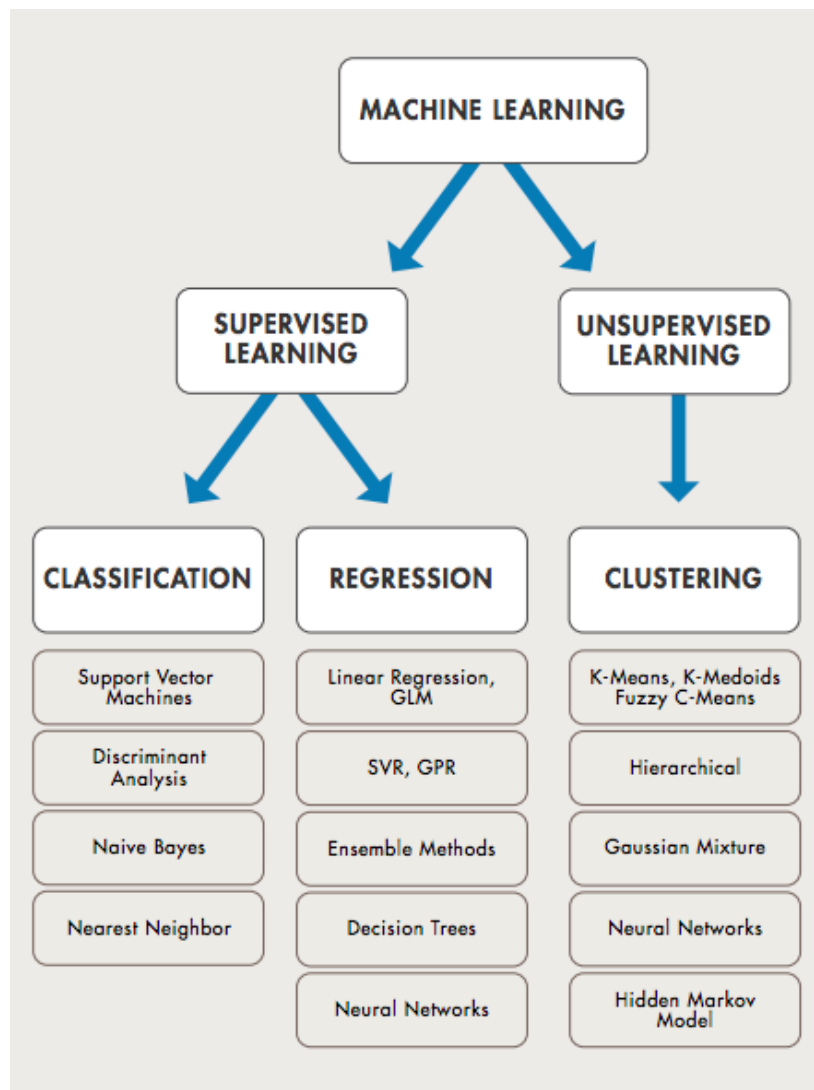
Regression: Ένα πρόβλημα θεωρείται πρόβλημα παλινδρόμησης όταν η μεταβλητή της εξόδου του είναι μία πραγματική τιμή.

1.2.2 Μη Επιβλεπόμενη: Ονομάζεται μη επιβλεπόμενη μάθηση, γιατί σε αντίθεση με την επιβλεπόμενη, δεν υπάρχουν σωστές απαντήσεις και δεν υπάρχει κάποιος να επιτηρεί. Οι αλγόριθμοι λειτουργούν μόνοι τους ανακαλύπτοντας τα μη-επισημειωμένα δεδομένα. Στόχος της μη επιβλεπόμενης μάθησης είναι η μοντελοποίηση της υποκείμενης δομής ή διανομής στα δεδομένα, ώστε να αποκτήσει γνώσεις πάνω σε αυτά[δ].

Μία κατηγορία προβλημάτων μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση (Clustering).

Clustering : Η συσταδοποίηση είναι η διαδικασία όπου τα αντικείμενα-διανύσματα , εντάσσονται σε ομάδες. Τα αντικείμενα που ανήκουν στην ίδια ομάδα είναι όμοια μεταξύ τους, ενώ αν ανήκουν σε διαφορετική είναι ανόμοια.

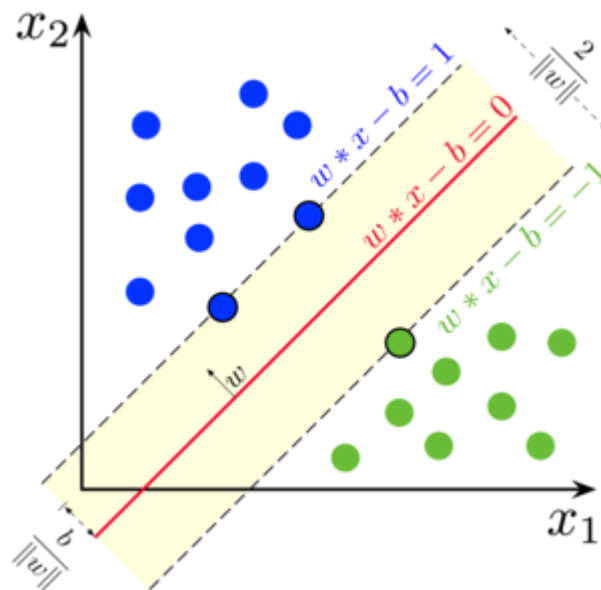
Ημι-επιβλεπόμενη : Η ημι-επιβλεπόμενη μηχανική μάθηση αποτελεί συνδυασμό της επιβλεπόμενης και της μη επιβλεπόμενης. Οι αλγόριθμοι εκπαιδεύονται τόσο από επισημειωμένα όσο και από μη επισημειωμένα δεδομένα.



Εικόνα1:Πηγή quantumcomputingtech.blogspot.com

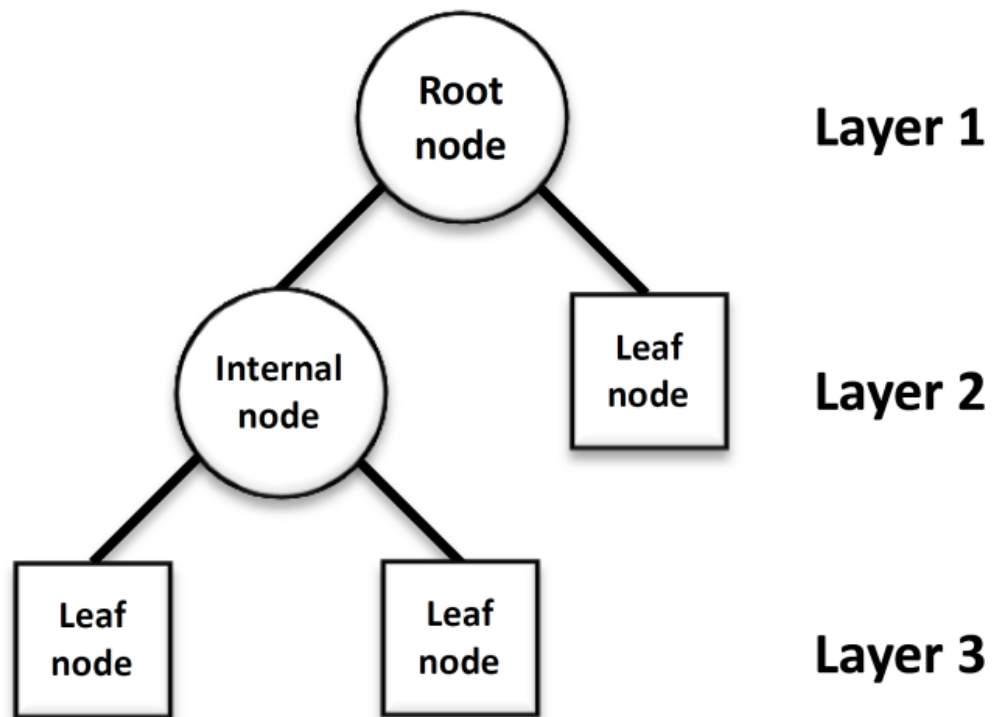
Τα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται είναι τα εξής:

- Support Vector Machines(SVMs): Είναι μοντέλα επιβλεπόμενης μάθησης , τα οποία μέσω αλγορίθμων αναλύουν δεδομένα μέσω διαδικασιών ταξινόμησης(classification) και ανάλυσης παλινδρόμησης(data regression). Τα SVMs είναι ιδανικά για την επίλυση προβλημάτων δυαδικής ταξινόμησης μέσω ενός υπερπλάνου αν και υπάρχουν αρκετές πρόσθετες μέθοδοι που τα χρήζουν κατάλληλα και για προβλήματα πολλαπλών τάξεων.
 - Sequential Minimal Optimization (SMO): Είναι μια βελτιστοποιημένη έκδοση του Support Vector Machines (SVMs).



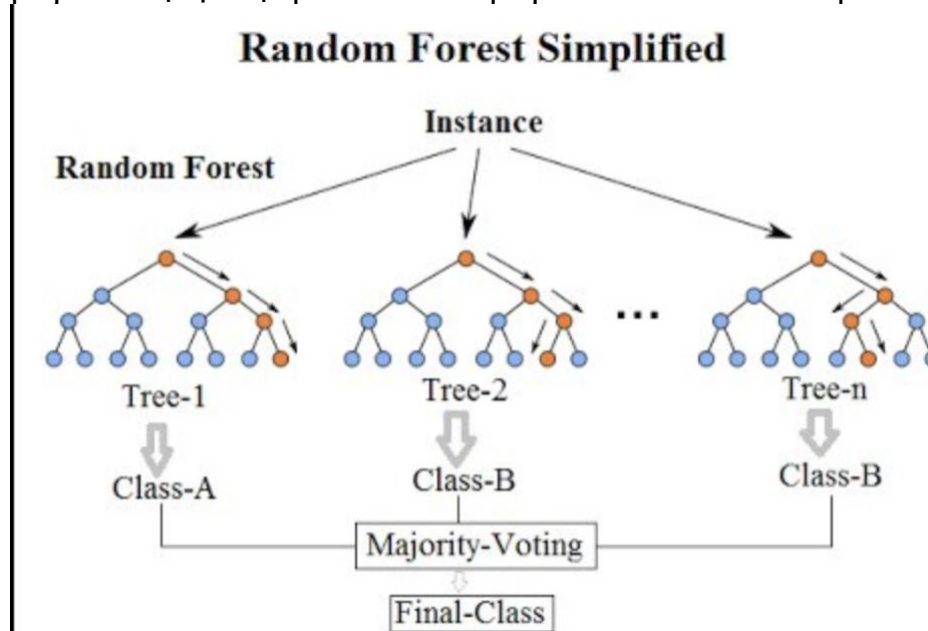
Εικόνα SVMs Πηγή: Wikipedia

- Decision Trees(DTs):Τα δένδρα αποφάσεων αποτελούν ένα εργαλείο πρόβλεψης και ταξινόμησης. Είναι μοντέλο επιβλεπόμενης μάθησης. Έχουν τη μορφή διαγράμματος ροής, όπου κάθε κόμβος αναπαριστά ένα χαρακτηριστικό προς επιλογή ,ενώ τα κλαδιά αναπαριστούν το αποτέλεσμα της κάθε επιλογής.



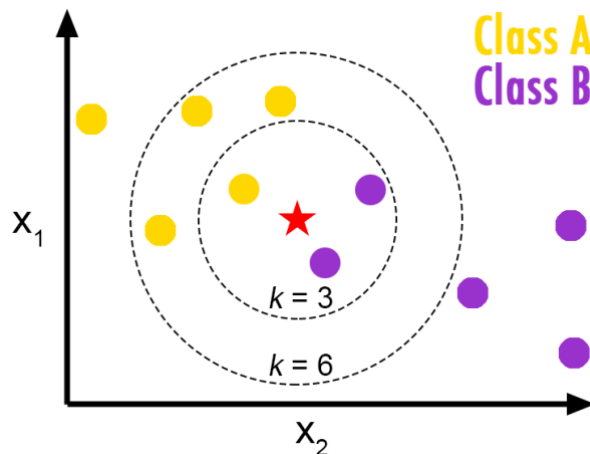
Decision Tree Πηγή: www.researchgate.net

-Random Forest(RF):Είναι ένας αλγόριθμος ταξινόμησης ή παλινδρόμησης ,που επιλέγει τυχαίες τιμές και δεδομένα, δημιουργώντας πολλά δέντρα αποφάσεων. Προσφέρει μεγαλύτερη ακρίβεια σε ένα μεγάλο σύνολο δεδομένων.



Πηγή: medium.com

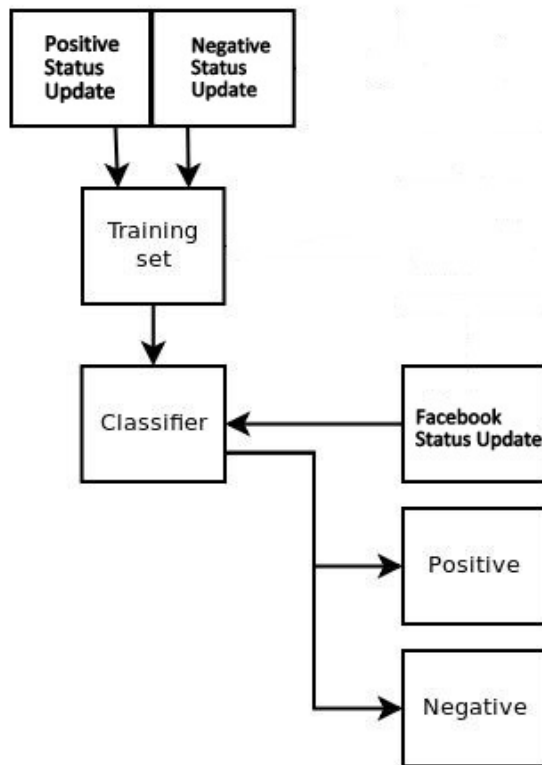
- K-Nearest Neighbors(KNNs):Είναι ένας αλγόριθμος επιβλεπόμενης μάθησης ,που μπορεί να χρησιμοποιηθεί για διεργασίες ταξινόμησης και παλινδρόμησης. Ένα αντικείμενο ταξινομείται λαμβάνοντας υπόψη τους γείτονές του.



K-NN Πηγή: medium.com

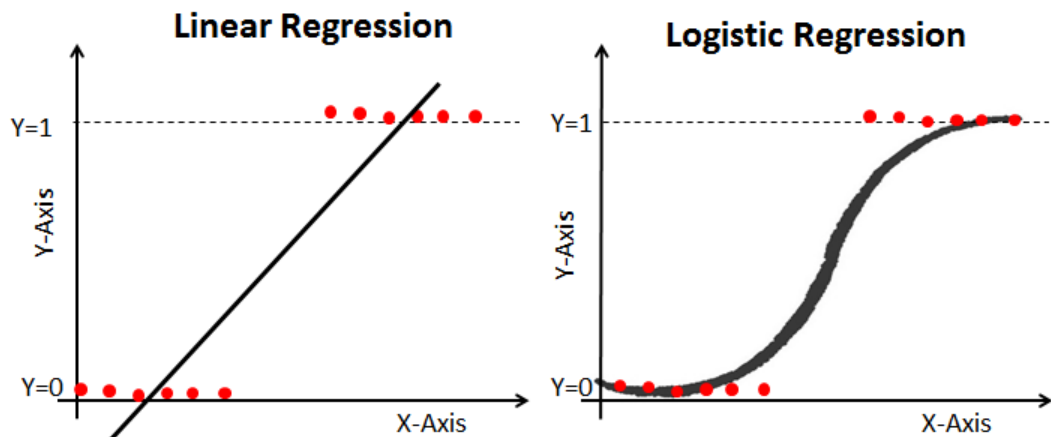
- Naïve Bayes(NB):Οι ταξινομητές Naïve Bayes είναι μία ομάδα αλγορίθμων ταξινόμησης που βασίζονται στο θεώρημα του Bayes. Ο απλός ταξινομητής Bayes είναι μια μέθοδος μάθησης που στηρίζεται σε κατανομές πιθανότητας, με την προϋπόθεση πως τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους. Μοντελοποιεί ένα έγγραφο με βάση την παρουσία ή την απουσία λέξεων.

Θεώρημα Bayes : $P(A|B)=P(B|A)P(A)/P(B)$



Εικόνα Naïve Bayes Πηγή: www.researchgate.net

- Multinomial Naïve Bayes: Αποτελεί μία έκδοση του Naïve Bayes ειδικά σχεδιασμένη για έγγραφα κειμένου, η οποία βασίζεται Στη συχνότητα εμφάνισης των λέξεων και προσαρμόζει τους υπολογισμούς ανάλογα με αυτή.
- Bernoulli Naïve Bayes: Αυτός ο ταξινομητής μοιάζει αρκετά με το Multinomial Naïve Bayes, με τη διαφορά ότι δε βασίζεται στη συχνότητα εμφάνισης των λέξεων ,αλλά ταξινομεί με βάση τις τιμές ναι ή όχι.
- Gaussian Naïve Bayes: Χρησιμοποιείται με την παράθεση ότι οι προβλέψεις έχουν μη ευδιάκριτες ,συνεχείς τιμές .
- Logistic Regression: Είναι μία προγνωστική ανάλυση που χρησιμοποιείται για την περιγραφή δεδομένων και για να εξηγήσει τη σχέση μεταξύ εξαρτώμενης δυαδικής μεταβλητής και άλλων ανεξάρτητων μεταβλητών.
- Linear regression: Χρησιμοποιείται για την πρόβλεψη της σχέσεως μεταξύ δύο μεταβλητών ή παραγόντων.(εξαρτημένες μεταβλητές).



Πηγή: medium.com

1.2.3 Βαθιά μάθηση :

Η βαθιά μάθηση είναι μία κατηγορία μηχανικής μάθησης, η οποία μας στέλνει ένα βήμα πιο κοντά στον προορισμό της μηχανικής μάθησης : Την τεχνητή νοημοσύνη. Η βαθιά μάθηση αφορά την εκμάθηση πολλαπλών επιπέδων αναπαράστασης και την κατανόηση δεδομένων , όπως εικόνες , ήχος, κείμενο. Ουσιαστικά χρησιμοποιεί τεχνητά νευρωνικά δίκτυα. Τα επίπεδα στα εκπαιδευμένα στατικά μοντέλα ,αντιστοιχούν σε ξεχωριστά επίπεδα εννοιών. Οι έννοιες υψηλότερου επιπέδου ορίζονται από αυτές ενός χαμηλότερου.Επιπλέον τα περισσότερα μοντέλα βασίζονται στη μη επιβλεπόμενη μάθηση.[ε]

1.3.Τι είναι η εξόρυξη γνώσης από σώματα κειμένων:

Εξόρυξη γνώσης(Data Mining) είναι η διαδικασία εξαγωγής χρήσιμης πληροφορίας μέσω αλγόριθμων κατηγοριοποίησης και ομαδοποίησης. Στόχος της συγκεκριμένης διαδικασίας είναι η βοήθεια του ανθρώπου στη λήψη αποφάσεων εξάγωντας κάποιο πρωτότυπο μέσω αυτόματων ή μη αυτόματων αναλύσεων μεγάλου όγκου δομημένων δεδομένων. Τα δεδομένα αυτά αφού εκπαιδευτούν μπορούν να χρησιμοποιηθούν μέσω της μηχανικής μάθησης για την εξαγωγή ορθών συμπερασμάτων.

Η εξόρυξη γνώσης από κείμενα(Text Mining) είναι η εύρεση χρήσιμων πληροφοριών από δεδομένα που είναι σε μορφή κειμένου. Η διαδικασία εξόρυξης γνώσης από κείμενα είναι η ίδια με αυτή του Data Mining, με βασική διαφορά ότι εκτός από δομημένα το Text Mining μπορεί να διαχειριστεί και αδόμητα ή ημι-δομημένα δεδομένα.

1.4 Ανάλυση συναισθήματος και εξόρυξη γνώμης(sentiment analysis and opinion mining)

Η ανάλυση συναισθήματος (SA) ή αλλιώς εξόρυξη γνώμης, είναι ουσιαστικά η επεξεργασία και η ανάλυση της φυσικής γλώσσας καθώς και η άντληση γνώσης που αφορά απόψεις, συναισθήματα, συμπεριφορές και στάσεις που εκφράζονται σε κείμενα. Η ανάλυση συναισθήματος είναι ιδιαίτερα χρηστική καθώς μπορεί να αντλήσει συμπεράσματα για τις πολιτικές πεποιθήσεις των ατόμων, την ευφυΐα αγοράς, για το πόσο ικανοποιημένοι είναι οι πελάτες κ.α.

2 Ανάλυση εργασιών

Οι εργασίες που αφορούν τις τεχνικές εξόρυξης γνώσης από σώματα κειμένων στον κοινωνικό ιστό αφθονούν, ωστόσο ελάχιστες είναι αυτές που επικεντρώνονται σε κείμενα γραμμένα στην ελληνική γλώσσα. Παρακάτω περιγράφονται οι εργασίες που συγκεντρώθηκαν και αφορούν τεχνικές εξόρυξης δεδομένων από σώματα κειμένων του ελληνικού κοινωνικού ιστού. Στις περισσότερες μελέτες που περιγράφονται στην παρούσα εργασία για την εξόρυξη γνώσης ακολουθείται η εξής διαδικασία: Αρχικά πραγματοποιείται η συλλογή δεδομένων από σώματα κειμένου του κοινωνικού ιστού. Έπειτα τα δεδομένα υφίστανται μία προεπεξεργασία αφαιρώντας ή προσθέτοντας χαρακτηριστικές όπου χρειάζεται(π.χ. αφαίρεση σημείων στίξης, τονισμός), ενώ

γίνονται και οι απαραίτητοι μετασχηματισμοί, ώστε να μετατραπούν σε μορφή κατανοητή από τους υπολογιστικούς κώδικες. Στη συνέχεια εφαρμόζονται τεχνικές εξόρυξης γνώσης/δεδομένων καταλήγοντας έτσι στα επιθυμητά αποτελέσματα, τα οποία πλέον μπορούν να ερμηνευτούν και να αξιολογηθούν.

2.1 Sentiment Analysis for Reputation Management: Mining the Greek Web

Η παρούσα εργασία αναφέρεται στην ανάλυση συναισθήματος και την εξόρυξη γνώμης που αφορά φυσικά πρόσωπα, προϊόντα καθώς και εταιρείες του ελληνικού ιστού. Συγκεκριμένα παρουσιάζεται η πρώτη αυτοματοποιημένη εμπορική πλατφόρμα στην Ελλάδα που πραγματοποιεί τις παραπάνω διεργασίες. Το όνομα της πλατφόρμας είναι PaloPro και αποτελεί προϊόν των εταιρειών Intellitech¹ και Palo².

Ουσιαστικά πρόκειται για μία συνδρομητική υπηρεσία, τα δεδομένα της οποίας συλλέγονται σε πραγματικό χρόνο και προέρχονται από μία ποικιλία πηγών του ελληνικού κοινωνικού ιστού(ειδήσεις, δημοσιεύσεις σε blogs,βίντεο, συζητήσεις,forums,καθώς και αναρτήσεις στο Facebook, στο Twitter και στο Youtube) με στόχο την παρακολούθηση και την ανάλυσή τους. Έτσι ο χρήστης μπορεί να έχει εικόνα για τις αρνητικές , τις θετικές ή τις ουδέτερες απόψεις πάνω στο αντικείμενο που τον αφορά.

Το PaloPro τροφοδοτείται από το προϊόν OpinionBuster , το οποίο μέσω αλγορίθμων μηχανικής μάθησης επεξεργάζεται τη φυσική γλώσσα. Το OpinionBuster είναι το πρώτο προϊόν που προσφέρει NER(Named Entity Recognition)για την ελληνική γλώσσα. Ως NER ορίζεται η διαδικασία εντοπισμού αναφορών οντοτήτων καθώς και η

ταξινόμησή τους σε κατηγορίες. Οι δύο πιο καθιερωμένες τεχνικές που χρησιμοποιούνται στη NER είναι η επιβλεπόμενη μηχανική μάθηση και τα χειρόγραφα σετ κανόνων.

Και οι δύο αυτές τεχνικές απαιτούν χειροκίνητες διαδικασίες για την εκπαίδευση των δεδομένων και την ανάπτυξη κανόνων ενώ βασικό μειονέκτημα αποτελεί και η ανάγκη εγκυρότητας. Μια ανάγκη που είναι δύσκολο έως αδύνατο να καλυφθεί. Συνεπώς εφαρμόζονται ορισμένες μεθοδολογίες για την αποτελεσματικότερη επίλυση του συγκεκριμένου ζητήματος. Έτσι γίνεται εκμετάλλευση μεγάλων εξωτερικών πηγών γνώσης, όπως η Wikipedia. Επιπλέον διατηρείται ένα σύνολο οντοτήτων που βρίσκονται στην ελληνική αγορά(εταιρείες, προϊόντα) ώστε να είναι πιο εύκολη η αποσαφήνισή τους. Επιπροσθέτως χρησιμοποιείται μια γραμματική NER ειδικά σχεδιασμένη για την εξαγωγή ονομασμένων οντοτήτων. Για την αποσαφήνιση των οντοτήτων γίνεται ταξινόμηση των εγγράφων ενώ χρησιμοποιείται μηχανική μάθηση βασισμένη σε Conditional Random Fields και στην ανάλυση κοινής αναφοράς στοιχείων βασιζόμενη σε κανόνες. Τέλος εφαρμόζεται ένα σύνολο κανόνων φιλτραρίσματος που συνδυάζουν τις πληροφορίες από όλους τους παραπάνω παράγοντες ώστε να αποφασιστεί η τελική τους ταξινόμηση.

Βασικό στοιχείο της παρούσας εργασίας είναι η ανάλυση συναισθήματος και η εξόρυξη γνώμης. Η ανάλυση συναισθήματος είναι ένας δημοφιλής τομέας έρευνας ,όπου συνεχώς παρουσιάζονται νέες προσεγγίσεις. Αυτές οι προσεγγίσεις εκμεταλλεύονται πηγές γνώσης λέξεων που αφορούν συναισθήματα και προσπαθούν να εντοπίσουν το συναίσθημα που εκφράζεται σε μία λέξη λαμβάνοντας υπόψη τα συμφραζόμενα.

Η ανάλυση συναισθήματος έχει επικεντρωθεί κυρίως σε κείμενα γραμμένα στην αγγλική γλώσσα. Συνεπώς είναι δύσκολο να έχει την ίδια απόδοση όταν χρειαστεί να ανταποκριθεί σε ανάλογες διεργασίες που αφορούν κείμενα διαφορετικών γλωσσών, όπως τα ελληνικά. Για αυτό το λόγο χρησιμοποιούνται δύο βασικοί μέθοδοι προσέγγισης κειμένων γραμμένα σε διαφορετικές γλώσσες. Η στατιστική και η λεξική. Η στατιστική χρειάζεται εκπαιδευμένα δεδομένα για τις διάφορες γλώσσες, τα οποία είναι συνήθως αραιά, ενώ η λεξική απαιτεί λεξικούς πόρους σε διάφορες γλώσσες, οι οποίοι δεν είναι πάντα διαθέσιμοι. Προκειμένου να μετριάσει το πρόβλημα τις έλλειψης πόρων καταβάλλεται προσπάθεια για την απόκτηση λεξικών και γραμματικών για ανάλυση συναισθημάτων. Αρχικά γίνεται η μετάφραση ενός αγγλικού λεξικού στην επιθυμητή γλώσσα και έπειτα τα αποτελέσματα χρησιμοποιούνται για την εκπαίδευση ενός στατιστικού ταξινομητή. Μία επιπλέον μεθοδολογία που εφαρμόζεται είναι η επιλογή λέξεων-κλειδιών συγκρίνοντας λέξεις των σωμάτων κειμένων. Το μοτίβο της κάθε λέξης-κλειδί παράγει πρότυπα γραμματικής που συμβάλουν στην ταξινόμηση συναισθήματος.

Στο πλαίσιο της αξιολόγησης του συγκεκριμένου συστήματος χρησιμοποιήθηκαν σώματα κειμένου επιλεγμένα με χειροκίνητο τρόπο, τα οποία προέρχονται από τροφοδοσίες RSS των εφημερίδων “Real News” και “Καθημερινή”, και περιλαμβάνουν αναφορές σε όλες τις οντότητες και τους συσχετισμούς τους. Για τις ανάγκες της συγκεκριμένης έρευνας επιλέχθηκε ο τομέας που αφορούσε την πολιτική και από τις δύο εφημερίδες. Έτσι σε χρονικό διάστημα δύο μηνών συλλέχθηκαν 2.300 κείμενα. Η ακρίβεια των αποτελεσμάτων που προέκυψε ήταν αρκετά υψηλή 93% ωστόσο η πολικότητα είχε τιμή 64%. Η όλη διαδικασία οδήγησε στο συμπέρασμα ότι τα περισσότερα εσφαλμένα αποτελέσματα οφείλονται στην

απουσία υποκατηγοριών πλαισίου των εμπλεκόμενων ρημάτων.

Η επόμενη έρευνα αναφέρεται σε μια αυτοματοποιημένη διαδικασία αξιολόγησης δύο οντοτήτων της τηλεφωνικής εταιρείας “Vodafone” και της τράπεζας “Alpha Bank”. Τα δεδομένα πάρθηκαν από τον κοινωνικό ιστό (Αναρτήσεις στο Facebook, tweets, blogs). Τα αποτελέσματα έδωσαν ακρίβεια 80%.

Συμπεραίνουμε λοιπόν πως η ανάλυση συναισθήματος αποτελεί ένα απαραίτητο εργαλείο για την εξόρυξη γνώσης από τον κοινωνικό ιστό και το διαδίκτυο γενικότερα, καθώς κατέχει κεντρικό ρόλο στην παρακολούθηση της πολικότητας των οντοτήτων, καθώς και των απόψεων και των συναισθημάτων που εκφράζονται σε κείμενα που έχουν αναρτηθεί.

2.2 Offensive Language Identification in Greek

Ο κοινωνικός ιστός αποτελεί ένα καθημερινό μέσο επικοινωνίας, συνεπώς είναι αναπόφευκτη η ύπαρξη προσβλητικού ή υβριστικού περιεχομένου. Συνεχώς διερευνώνται τρόποι αντιμετώπισης του συγκεκριμένου φαινομένου ενώ παράλληλα δημιουργούνται συστήματα που ανιχνεύουν τέτοιου τύπου φρασεολογία και την κατηγοριοποιούν ανάλογα με το είδος της (λεκτική επίθεση, λέξεις που εκφράζουν μίσος ή θυμό, καθώς και ρατσιστικά ή σεξιστικά σχόλια κ.α.). Η συγκεκριμένη έρευνα παρουσιάζει το πρώτο ελληνικό σχολιασμένο σύνολο δεδομένων που αναγνωρίζει την προσβλητική γλώσσα μέσα σε σώματα κειμένων του κοινωνικού ιστού και συγκεκριμένα το Twitter. Το σύνολο δεδομένων είναι το Offensive Greek Tweet Dataset (OGTD) εμπνευσμένο από το αγγλικό OLID.

Το OGTD περιλαμβάνει περίπου 4.800 δημοσιεύσεις του Twitter οι οποίες έχουν κατηγοριοποιηθεί ανάλογα με το αν είναι προσβλητικές ή όχι. Τα tweets σχολιάστηκαν από μία ομάδα εθελοντών ακολουθώντας την ίδια μεθοδολογία σχολιασμού με το OLID. Βασικό στοιχείο του

γραφτού λόγου στην ελληνική γλώσσα είναι ο τονισμός των λέξεων. Όμως οι χρήστες δεν τονίζουν σωστά πάντα τις λέξεις που χρησιμοποιούν ενώ άλλες φορές παραλείπουν εντελώς τον τονισμό των λέξεων. Συνεπώς βασικό μέλημα αποτελεί η ομαλοποίηση της κάθε λέξης στην ισοδύναμη μη τονισμένη μορφή της ώστε να υπάρξει μία ομοιομορφία δεδομένων και να αποφευχθεί η ασάφεια που μπορεί να προκύψει από ένα σύνολο τονισμένων, μη τονισμένων και μερικώς τονισμένων tweets.

Για την εκπαίδευση μοντέλων μηχανικής μάθησης στο OGTD πραγματοποιήθηκαν πειράματα, το καθένα εκ των οποίων χρησιμοποιεί ένα διαφορετικό συνδυασμό από μία ομάδα χαρακτηριστικών (TF / IDF, unigrams, bigrams, POS και ετικέτες σχέσης εξάρτησης). Οι λειτουργίες TF/IDF συνήθως χρησιμοποιούνται στην ταξινόμηση κειμένου και ορίζουν πόσο σημαντική είναι μία λέξη για το σώμα κειμένου που βρίσκεται. Για κάθε σημείο σε ένα ελληνικό tweet χρησιμοποιήθηκε ο αγωγός spaCy3 χάρη στον οποίο εξήχθησαν οι ετικέτες σχέσης εξάρτησης και τα POS. Δυστυχώς μέχρι στιγμής δεν υπάρχει ελληνικό λεξικό συναισθήματος που να πληρεί τις απαραίτητες προϋποθέσεις για τη συγκεκριμένη διεργασία.

Για τις ανάγκες της συγκεκριμένης μελέτης χρησιμοποιήθηκαν μοντέλα κλασσικής μηχανικής μάθησης και μοντέλα βαθιάς μάθησης. Για την κλασσική μηχανική μάθηση χρησιμοποιήθηκαν 5 διαφορετικά μοντέλα:

- Ένα SVM με γραμμικό πυρήνα
- Ένα SVM με ακτινικής βάσης πυρήνα (RBF)
- Ένας γραμμικός ταξινομητής με εκμάθηση Stochastic Gradient Descent (SGDC).
- Το θεώρημα Multinomial Naïve Bayes, το οποίο λειτουργεί με μετρήσεις περιστατικών.
- Το θεώρημα Bernoulli Naïve Bayes, που έχει σχεδιαστεί για δυαδικά χαρακτηριστικά

Κάθε κλασσικό μοντέλο συμπεριλήφθηκε υπό την προϋπόθεση ότι θα μπορούσε ως είσοδο να πάρει πίνακες για τοποθέτηση και να έχει εκπαιδευτεί με τις προεπιλεγμένες ρυθμίσεις λόγω του μεγέθους του συνόλου δεδομένων.

Για την βαθιά μάθηση χρησιμοποιήθηκαν 6 διαφορετικά μοντέλα:

- Pooled GRU
- Stacked LSTM with Attention
- LSTM and GRU with Attention
- 2D Convolution with Pooling
- GRU with Capsule
- LSTM with Capsule and Attention
- BERT

Κάθε μοντέλο βαθιάς μάθησης χρησιμοποιήθηκε για τον εντοπισμό της επιθετικότητας σε σώματα κειμένων.

Με βάση τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν στην τρέχουσα μελέτη, τα οποία αφορούν την απόδοση μεμονωμένων ταξινομητών για την αναγνώριση υβριστικής φρασεολογίας προκύπτουν τα εξής συμπεράσματα:

Όσον αφορά τα μοντέλα της κλασσικής μηχανικής μάθησης, τα χαρακτηριστικά TF/IDF unigram, το Linear SVM μπορεί να θεωρηθεί το οριακά καλύτερο μοντέλο που έχει εκπαιδευτεί με OGTD, καθώς τα ποσοστά της μέσης ακρίβειας και ανάκλησης είναι υψηλότερα.

Τα μοντέλα εκπαιδευμένα με χαρακτηριστικά TF/IDF bigram είχαν χειρότερη απόδοση με όλες τις τιμές χαμηλότερες από αυτές του TF/IDF unigram, με εξαίρεση το Multinomial Naïve Bayes που βελτιώθηκε στο macro-F1 για την κατηγορία «Μη προσβλητικό».

Γενικότερο συμπέρασμα είναι πως ακόμη και στην περίπτωση χρήσης POS και ετικετών σχέσης εξάρτησης τα χαρακτηριστικά unigram είχαν πολύ καλύτερα αποτελέσματα από τα bigram.

Στην περίπτωση της βαθιάς μάθησης το μοντέλο « LSTM and GRU with Attention» θεωρείται το καλύτερο μοντέλο που έχει εκπαιδευτεί για το OGTD, καθώς ξεπέρασε σε όλα τα επίπεδα ,όλα τα μοντέλα κλασσικής μηχανικής μάθησης καθώς και τα υπόλοιπα μοντέλα βαθιάς μάθησης.

2.3 Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features

Ο τουρισμός είναι ένας παράγοντας ζωτικής σημασίας για την Ελλάδα. Με την ανάπτυξη του Web 2.0 ο κοινωνικός ιστός έχει εξελιχθεί σε ένα μέσον από το οποίο μπορούμε να αντλήσουμε πληροφορίες με στόχο τη βελτίωση και την καλύτερη οργάνωση κάθε τομέα που αφορά την καθημερινότητα, την εξυπηρέτηση πελατών αλλά και την καλύτερη λειτουργία επιχειρήσεων, όπως οι ξενοδοχειακές μονάδες. Η συγκεκριμένη μελέτη εστιάζει στην εξόρυξη γνώσης από κριτικές ξενοδοχείων της ελληνικής έκδοσης του Tripadvisor.

Στόχος της παρούσας μελέτης είναι η δημιουργία ενός πρωτότυπου για την ταξινόμηση κριτικών ξενοδοχείων ανάλογα με το συναίσθημα που εκφράζεται σε κάθε μία. Τα σώματα κειμένων που συλλέχθηκαν προέρχονται από 1.800 κριτικές ξενοδοχείων στο Tripadvisor γραμμένες στα ελληνικά. Οι μισές κριτικές είναι θετικές ενώ οι υπόλοιπες αρνητικές. Τα σώματα κειμένων που επιλέχθηκαν εξετάστηκαν με χειροκίνητο τρόπο για τυχών γραμματικά ή συντακτικά λάθη και έπειτα εφαρμόστηκαν διορθώσεις όπου ήταν απαραίτητο.

Η εκπαίδευση των δεδομένων πραγματοποιείται μέσω ενός αλγόριθμου δυαδικής ταξινόμησης Support Vector Machines(SVMs). Ο αλγόριθμος αυτός επισημαίνει την πολικότητα(θετική ή αρνητική) σε

κείμενα τα οποία με τη χρήση unigrams αναπαρίστανται ως διανύσματα χαρακτηριστικών. Για την υλοποίηση των πειραμάτων της εργασίας αυτής έγινε χρήση του λογισμικού RapidMiner, το οποίο παρέχει απαραίτητα εργαλεία για τη στατιστική ανάλυση κειμένου. Χρησιμοποιήθηκαν δύο μέθοδοι ταξινόμησης: Η τεχνική TF/IDF του σάκου λέξεων και η τεχνική Term Occurrence (TO). Για να μπορέσει ο αλγόριθμος να διαβάσει τα δεδομένα χρειάστηκαν ορισμένες τροποποιήσεις. Κάθε κείμενο επισημάνθηκε αυτόματα και φιλτράρεται ανάλογα με το μήκος των στοιχείων του. Στοιχεία μικρότερα των 4 και μεγαλύτερα των 25 χαρακτήρων, καθώς και στοιχεία όπως αντωνυμίες, προθέσεις και άρθρα αφαιρέθηκαν. Με την ολοκλήρωση αυτής της διεργασίας το σύνολο των δεδομένων μπορεί να αναπαρασταθεί σε πίνακα.

Η πρώτη μέθοδος στάθμισης που χρησιμοποιήθηκε είναι η προσέγγιση TF-IDF Bag-of-Words, ένα μοντέλο γλώσσας στο οποίο ένα κείμενο αναπαρίσταται ως ένας σάκος λέξεων, όπου κάθε λέξη (unigram) θεωρείται ως ένα χαρακτηριστικό. Στην προκειμένη περίπτωση κάθε έγγραφο αντιπροσωπεύεται ως μια μη ταξινομημένη συλλογή χαρακτηριστικών. Για τη δημιουργία διανύσματος χαρακτηριστικών χρησιμοποιείται το TF-IDF, η πιο κοινή μέθοδος στάθμισης του τομέα ανάκτησης πληροφοριών που μπορεί να αυξήσει σημαντικά την ακρίβεια ταξινόμησης των αποτελεσμάτων ανάλυσης συναισθημάτων. Στη συγκεκριμένη έρευνα συλλέχθηκαν τα πρώτα 1000 χαρακτηριστικά με το υψηλότερο TF-IDF βάρος. Δεδομένου αυτών των χαρακτηριστικών ο αλγόριθμος εκπαιδεύεται ώστε να ορίζει την πολικότητα νέων ντοκουμέντων, δηλαδή να κρίνει αν αποτελούν θετικές ή αρνητικές αξιολογήσεις. Το TF μετράει το πόσο σημαντικό είναι ένα στοιχείο για το έγγραφο στο οποίο βρίσκεται ενώ το IDF εκτιμά το πόσο σπάνια εμφανίζεται ένας όρος στο συγκεκριμένο έγγραφο. Το βάρος TF-IDF είναι υψηλότερο είτε όταν ο όρος εμφανίζεται πολλές φορές σε ένα έγγραφο είτε λιγότερες μέσα σε ένα σύνολο εγγράφων. Ενώ όταν το βάρος είναι χαμηλότερο ισχύει το αντίστροφο.

Η δεύτερη προσέγγιση που χρησιμοποιήθηκε είναι η Term Occurrence (TO), η οποία αποτελεί την πιο απλή προσέγγιση για τον καθορισμό του

συναισθήματος που επικρατεί σε ένα κείμενο. Η πολικότητα του εγγράφου αποτελεί το κριτήριο ταξινόμησης του. Αν οι λέξεις με θετική σημασία είναι περισσότερες τότε χαρακτηρίζεται θετική η αξιολόγηση ,ενώ αν συμβαίνει το αντίθετο χαρακτηρίζεται αρνητική. Αν οι θετικές με τις αρνητικές εκφράσεις είναι ισότιμες τότε η αξιολόγηση χαρακτηρίζεται ουδέτερη.

Στο πρώτο βήμα για να υλοποιηθεί η προσέγγιση (ΤΟ) χρειάστηκε να δημιουργηθεί χειροκίνητα μία ελληνική βιβλιοθήκη με λέξεις που έχουν θετική και λέξεις που έχουν αρνητική σημασία. Το δεύτερο βήμα ήταν να εισαχθούν οι δύο λίστες που δημιουργήθηκαν στον αλγόριθμο και να δοθεί ο ακριβής αριθμός που εμφανίζεται ένας όρος στο έγγραφο. Τέλος ο αλγόριθμος μέτρησε τη συχνότητα εμφάνισης θετικών και αρνητικών όρων στο έγγραφο και τα κατηγοριοποίησε.

Για την επικύρωση των δεδομένων χρησιμοποιήθηκε η διαδικασία tenfold cross validation. Τα δεδομένα χωρίστηκαν σε δέκα πτυχές και η κάθε μία περιλάμβανε 180 αξιολογήσεις ξενοδοχείων. Η απόδοση του ταξινομητή αξιολογήθηκε με βάση την εγκυρότητα, την ανάκληση και την ακρίβεια. Η μέθοδος TF-IDF bag-of-words έχει ποσοστό ακρίβειας 95.78 % ενώ η μέθοδος ΤΟ 71.76 %. Συμπεραίνουμε λοιπόν ότι η πρώτη μέθοδος είναι πιο αποτελεσματική από τη δεύτερη.

2.4 Authorship Attribution and Gender Identification in Greek Blogs

Τα Blogs όντας κομμάτι του κοινωνικού ιστού ,αποτελούν χώρο έκφρασης για άτομα διαφορετικών ηλικιών, φύλου, ιδεολογιών και προτιμήσεων. Συνεπώς είναι σημαντική η δυνατότητα εξατομίκευσής τους βάση αυτών των κριτηρίων. Η συγκεκριμένη μελέτη επικεντρώνεται στην εύρεση του συγγραφέα ενός Blog καθώς και του φύλου του μέσω της εκπαίδευσης αλγορίθμων μηχανικής μάθησης. Τα δεδομένα που χρησιμοποιούνται προέρχονται από σώματα κειμένων ελληνικών Blogs. Γίνεται χρήση δύο τύπων έρευνας. Authorship Attribution και Author profiling. Στην πρώτη περίπτωση σκοπός είναι η εύρεση του συντάκτη ενός Blog μέσα από ένα σύνολο κειμένων n

γνωστών συγγραφέων, υπό την υπόθεση ότι ο συγκεκριμένος συγγραφέας εμπεριέχεται στη λίστα. Ο δεύτερος τύπος έρευνας χρησιμοποιείται σε εφαρμογές που σχετίζονται με την ανάλυση πληροφοριών και την εξόρυξη γνώσης και συναισθήματος σε περιπτώσεις που θέλουμε να αποκτήσουμε γνώση για το φύλο, την ηλικία και την ψυχοσύνθεση του συγγραφέα. Για να είναι εφικτές οι παραπάνω διαδικασίες ήταν απαραίτητη η δημιουργία ενός ελληνικού λεξικού από την αρχή. Τα τελικά δεδομένα που συλλέχθηκαν προέρχονται από 20 ελληνικά Blogs ,τα 10 εκ των οποίων είχαν άντρες συγγραφείς και τα άλλα 10 γυναίκες .Επιλέχθηκαν 50 αναρτήσεις από τον κάθε συγγραφέα και έχουν κοινό θέμα τις προσωπικές υποθέσεις. Τα αποτελέσματα της έρευνας έδειξαν ότι οι γυναίκες αναρτούν μεγαλύτερες δημοσιεύσεις με μικρότερη ποικιλία στο μέγεθος.

Στη συνέχεια γίνεται χρήση ενός συνόλου στατιστικών λειτουργιών καθώς και χαρακτήρων και λέξεων με χαρακτηριστικά n-grams για να παρατηρηθεί ο συσχετισμός της συγγραφής με το φύλο του συγγραφέα. Από αυτά τα n-grams μετρήθηκαν τα 300 που εμφανίζονταν πιο συχνά και η συχνότητά τους ομαλοποιήθηκε σε ένα κείμενο 100 λέξεων. Για την υλοποίηση της συγκεκριμένης μεθοδολογίας έγινε χρήση του αλγόριθμου Sequential Minimal Optimization (SMO) μια βελτιστοποιημένη έκδοση του Support Vector Machines (SVMs),τα οποία αποτελούν μία από τις πιο σύγχρονες μεθόδους ταξινόμησης στη μηχανική μάθηση.Τα SVMs είναι ιδανικά για την επίλυση προβλημάτων δυαδικής ταξινόμησης αν και υπάρχουν αρκετές πρόσθετες μέθοδοι που τα χρήζουν κατάλληλα και για προβλήματα πολλαπλών τάξεων. Έπειτα στα πλαίσια της ταξινόμησης που πραγματοποιήθηκε έγινε χρήση της μεθοδολογίας 10-fold cross-validation, ώστε να αποφευχθούν τυχών διακυμάνσεις στην απόδοση του αλγορίθμου. Τα αποτελέσματα είχαν 85.4% ακρίβεια για απόδοση συγγραφής και 82.6% για την ταυτοποίηση του φύλου. Και τα δύο αποτελέσματα ακρίβειας θεωρήθηκαν αρκετά ικανοποιητικά.

Στη συνέχεια, πραγματοποιήθηκε ένα πείραμα το οποίο εξετάζει το πως ο αριθμός των υποψηφίων επηρεάζει την έκβαση των αποτελεσμάτων της ακρίβειας. Τα δεδομένα χωρίστηκαν σε 4 ομάδες

(2,4,8 και 16 συγγραφέων) και για την κάθε μία εφαρμόστηκαν 10 διαφορετικοί συνδυασμοί συγγραφέων. Τα αποτελέσματα έδειξαν ότι όσο λιγότεροι ήταν οι συγγραφείς τόσο μεγαλύτερο ήταν το ποσοστό ακρίβειας ,με μεγαλύτερη απόδοση όταν οι συγγραφείς δεν ξεπερνούν τους 4.Έπειτα μέσω του αλγόριθμου επιλογής χαρακτηριστικών κειμένων προς ταξινόμηση Information Gain, διερευνήθηκε πως επιδρούν τα στυλομετρικά χαρακτηριστικά σε διαδικασίες όπως η ταυτοποίηση του φύλου. Η διαδικασία αυτή οδήγησε στο συμπέρασμα ότι ένας χαρακτήρας n-gram περιέχει σημαντικές πληροφορίες για τη συγγραφή ενώ μια λέξη n-gram συμβάλει σημαντικά στην ταυτοποίηση του φύλου του συγγραφέα.

2.5 Nowcasting the Stance of Social Media Users in a Sudden Vote: The Case of the Greek Referendum

Η παρούσα μελέτη αφορά την πρόβλεψη της ψήφου ενός χρήστη και των τελικών αποτελεσμάτων μίας ψηφοφορίας, λαμβάνοντας ως δεδομένα το περιεχόμενο των κοινωνικών δικτύων. Πιο συγκεκριμένα εστιάζει στην άντληση αυτών των πληροφοριών σε περίπτωση που η συγκεκριμένη ψηφοφορία είναι απροσδόκητη ,ξαφνική και με περιορισμένο χρονικό περιθώριο. Η έρευνα αφορά ένα πραγματικό γεγονός, το δημοψήφισμα που πραγματοποιήθηκε στην Ελλάδα τον Ιούνιο του 2015.

Για τις ανάγκες της διεργασίας πάρθηκε περιεχόμενο των χρηστών του Twitter, από τα οποία φαινόταν το τι θα ψήφιζε ο κάθε χρήστης και στη συνέχεια συγκρίθηκαν με τα πραγματικά αποτελέσματα του δημοψηφίσματος. Ορίστηκε μία δυαδική εργασία ταξινόμησης ευαίσθητη στο χρόνο, λαμβάνοντας ως δεδομένα την ψήφο των χρηστών (ναι ή όχι).Για ορισμένους λογαριασμούς, όπως αυτούς που ανήκουν σε κομματικές παρατάξεις ήτανε ήδη γνωστή η απάντηση .Συνεπώς χρησιμοποιήθηκαν για μια ημι-επιβλεπόμενη διαδικασία υπό την προϋπόθεση ότι όποιος χρήστης αναμετάδιδε τα Tweets της συγκεκριμένης παράταξης ,είχε κοινή στάση στην ψηφοφορία. Έτσι

χρησιμοποιήθηκαν 267 λογαριασμοί ως seeds ,148 ναι και 119 όχι, που αφορούν την περίοδο πριν την ολοκλήρωση της ψηφοφορίας. Στη συνέχεια καταμετρήθηκαν οι λογαριασμοί που κοινοποιούσαν από άλλους χρήστες tweets που αφορούν την ψήφο τους καθώς και όσοι χρησιμοποιούσαν hashtags που υποστήριζαν κάποιο από τις δύο επιλογές ψήφου. Ο τελικός αριθμός των λογαριασμών που χρησιμοποιήθηκαν είναι ο 2,197 (οχι 77%, ναι 23%).

Για τη διεξαγωγή του πειράματος προτείνεται μία ημι-επιβλεπόμενη μάθηση ενός πυρήνα πολλαπλών επιλίξεων που λαμβάνει δεδομένα από κείμενα με χρονική ευαισθησία και πληροφορίες δικτύου.

Εφαρμόστηκαν οι εξείς λειτουργίες:

- Textual Information(TEXT):Έγιναν ενσωματώσεις λέξεων, εκπαιδεύοντας μία συλλογή tweets που δεν είχαν αναδημοσιευθεί. Στη συνέχεια έγινε η κατάλληλη επεξεργασία ώστε αφαιρεθούν οι περιττοί χαρακτήρες ενώ επιπλέον αφαιρέθηκαν και λέξεις που εμφανίζονταν λιγότερο από 5 φορές. Κάθε λέξη εκπαιδεύτηκε σε διάνυσμα 50 διαστάσεων ,έτσι ώστε κάθε tweet να αναπαριστάται ως μέσος όρος κάθε διάστασης.
- Network Information (NETWORK):Εκπαιδεύτηκαν ενσωματώσεις LINE σε διαφορετικά χρονικά βήματα των γραφημάτων .Ενώ επιλέχθηκαν τα re-tweets έναντι των αναφορών χρηστών , καθώς έχουν μεγαλύτερη πολικότητα. Μέσω της εκπαίδευσης του LINE δημιουργήθηκαν αναπαραστάσεις χρηστών 50 διαστάσεων.

Τα μοντέλα που εφαρμόστηκαν είναι τα εξής:

- Convolution Kernel Models: MCKL και SVM
- Baselines: Logistic Regression (LR) ,a feed-forward neural network (FF) ,Random Forest (RF) ,SVM.

Κάθε μέρα εφαρμοζόταν μία διεργασία 5-fold cross-validation στο σετ εκπαίδευσης για την επιλογή παραμέτρων. Τα αποτελέσματα έδειξαν

ότι όσο πιο κοντά στην ημερομηνία του δημοψηφίσματος εφαρμοζόντουσαν οι παραπάνω διεργασίες ,τόσο πιο έγκυρα ήταν τα αποτελέσματα. Τελικό συμπέρασμα του πειράματος είναι πως το MCKL(Multiple Convolution Kernel Learning)είχε τη μεγαλύτερη ακρίβεια με ποσοστό 20% στο F-score,καθώς μπορεί να συνδυάσει πληροφορίες και από τις δύο λειτουργίες που χρησιμοποιήθηκαν(TEXT,NETWORK) κρίνοντας το πόσο σημαντικές είναι.

2.6 Sentiment Analysis of Greek Tweets and Hashtags using Sentiment Lexicon

Η παρούσα εργασία εξετάζει δεδομένα κειμένου που έχουν συλλεχθεί από ελληνικές αναρτήσεις του Twitter ,λαμβάνοντας υπόψη το συναίσθημα που εκφράζουν. Στόχος της εργασίας αυτής είναι η εφαρμογή μίας μεθοδολογίας που παρέχει διατίμηση συναισθημάτων για ελληνικά tweets, συναισθηματική αξιολόγηση hashtags και ανάλυση της μεταβολής των συναισθημάτων των hashtags με την πάροδο του χρόνου.

Τα δεδομένα συλλέχθηκαν με τη χρήση του Steaming API Twitter μέσω γλώσσας προγραμματισμού Python.Ξεκινώντας από έναν τυχαίο χρήστη δημιουργήθηκε η πρώτη λίστα χρηστών αποτελούμενη από του ακολούθους του συγκεκριμένου χρήστη. Στη συνέχεια συλλέχθηκαν τα tweets των χρηστών αυτών .καθώς και τα ids των δικών τους ακολούθων, επαναλαμβάνοντας τη διαδικασία. Οι χρήστες που χρησιμοποιήθηκαν έχουν αναρτήσεις γραμμένες στα ελληνικά ,για να εξασφαλιστεί αυτό κάθε ανάρτηση περιλαμβάνει τουλάχιστον 4 ελληνικούς χαρακτήρες Unicode.Χρησιμοποιήθηκαν οι 200 πιο πρόσφατες αναρτήσεις συμπεριλαμβανόμενης της χρονοσφραγίδας της κάθε μίας σε χρονικό διάστημα μίας εβδομάδας.

Το λεξικό συναισθήματος που χρησιμοποιήθηκε είναι το Greek Sentiment Lexicon το οποίο περιέχει 2315 καταχωρήσεις που αφορούν τα συναισθήματα του θυμού, της αηδίας, του φόβου, της χαράς και της

έκπληξης. Στο λεξικό περιλαμβάνεται συναισθηματική αξιολόγηση μέσω 4 βαθμολογητών. Το τελικό σκορ της αξιολόγησης υπολογίζεται από το μέσο όρο των βαθμολογητών.

Η προεπεξεργασία των δεδομένων έγινε με το Hadoop και το MapReduce. Τα hashtags κατηγοριοποιήθηκαν ανάλογα με την ομοιότητά τους, ενώ συμπεριλήφθηκαν και τα retweets ,δηλαδή επαναδημοσιεύσεις ενός tweet από άλλους χρήστες ,στα πλαίσια της έκφρασης κοινού συναισθήματος. Στη συνέχεια αφαιρέθηκαν τα σημεία στίξης , ενώ ο τονισμός των λέξεων αφαιρέθηκε ώστε το κείμενο να είναι συμβατό με το λεξικό.

Στη συνέχεια πραγματοποιήθηκε εκτίμηση συναισθήματος τόσο για τα tweets όσο και για τα hashtags. Τα hashtags που προσμετρήθηκαν εξετάστηκαν ανάλογα με το χρόνο σε αυξανόμενες τιμές και αφορούν μέρες που συγκεντρώθηκαν πάνω από 60 tweets.

Στο πείραμα που πραγματοποιείται στη συγκεκριμένη μελέτη επιλέχθηκαν ορισμένα Tweets ώστε να εξεταστεί καλύτερα η λειτουργία του αλγορίθμου. Οι κατηγορίες που επιλέχθηκαν περιλαμβάνουν πάνω από 1000 tweets και εκφράζουν γεγονότα για τα οποία μπορεί να υπάρξει έντονο συναίσθημα από την πλευρά των χρηστών. Επιπλέον τα αποτελέσματά τους μπορούν να εκτιμηθούν με βάση την κοινή λογική. Εν κατακλείδι ο αλγόριθμος ήταν ικανός να προβλέψει σωστά τα συναισθήματα που εκφράζονται για τα γεγονότα και τις καταστάσεις που επιλέχθηκαν.

Τα τελικά συμπεράσματα δείχνουν πως η προσέγγιση αυτή δεν είναι ικανή να ανιχνεύσει σαρκαστικές και ειρωνικές αναρτήσεις. Επιπλέον συναισθήματα που μοιάζουν μεταξύ τους δε διαχωρίζονται ιδιαίτερα αποτελεσματικά μεταξύ τους. Επιπροσθέτως τα λεξικά περιλάμβανε μόνο το 11.6% των λέξεων που χρησιμοποιήθηκαν. Κλείνοντας αξίζει να σημειωθεί ότι τα συναισθήματα του φόβου και της λύπης είχαν μικρότερες τιμές από τις υπόλοιπες αξιολογήσεις, καθώς ο μέσος χρήστης δεν προτιμά να εκφράζει τέτοιου είδους συναισθήματα στο διαδίκτυο ,ενώ και το λεξιλόγιο που τα εκφράζει είναι σχετικά πολύπλοκο ώστε να ανιχνευθεί.

2.7 A Deep Learning Approach to Aspect-Based Sentiment Prediction

Η παρούσα εργασία αφορά την πρόβλεψη συναισθημάτων η οποία ορίζεται ως ένα πρόβλημα επιτηρούμενης ταξινόμησης και αντιμετωπίζεται μέσω μίας αρχιτεκτονικής βαθιάς μάθησης, που βασίζεται σε μονάδες Bidirectional Long Short-Term Memory (BiLSTM) ,σε συνδυασμό με συνελκτικά επίπεδα προσοχής. Η συγκεκριμένη προσέγγιση εφαρμόζεται λαμβάνοντας ως δεδομένα αναρτήσεις στο Twitter , το Facebook ,το Instagram και άλλες σελίδες του ελληνικού κοινωνικού ιστού ,γραμμένες στην ελληνική γλώσσα.

Το σύστημα έχει την εξής αρχιτεκτονική: Αρχικά πραγματοποιείται μία διαδικασία data crawling.Ουσιαστικά πρόκειται για μία διαδικασία συνεχούς πρόσβασης σε διαδικτυακές πηγές και η συλλογή και το φιλτράρισμα δεδομένων, τα οποία προωθούνται σε συγκεκριμένο αποθετήριο. Στη συνέχεια υφίσταται καθαρισμός και ομογενοποίηση των δεδομένων. Σε αυτό το βήμα γίνεται εκκαθάριση σφαλμάτων ,αφαίρεση μη χρηστικού περιεχομένου της σελίδας, λόγου χάρη διαφημίσεις ή αναρτήσεις ακατάλληλου περιεχομένου. Τέλος τα δεδομένα αποθηκεύονται σε μία βάση δεδομένων, όπου μπορούν να αναγνωριστούν μοτίβα, πληροφορίες και συναισθήματα.

Στη συνέχεια επισημαίνονται τα συναισθήματα που εκφράζονται για κάθε αναφορά. Τα συναισθήματα μπορούν να χαρακτηριστούν ως θετικά, αρνητικά ή ουδέτερα. Τα κριτήρια μπορούν να είναι γενικά(κοινά συμφωνημένα κριτήρια, όπως λέξεις ή φράσεις που εκφράζουν κάποιο συναίσθημα) ή συγκεκριμένα(κανόνες που αναιρούν στην προκειμένη περίπτωση τους γενικούς και έχουν προτεραιότητα).Κάθε πτυχή λαμβάνει ως παράγοντες την πηγή των πληροφοριών, τις οντότητες, το συγγραφέα ,τις συγκυρίες και τις ώρες που αναρτήθηκε μία δημοσίευση. Η επισήμανση γίνεται με τη βοήθεια του λογισμικού και βοηθάει στον καθορισμό του συνόλου δεδομένων.

Όπως προαναφέρθηκε τα δεδομένα πάρθηκαν από αναρτήσεις του ελληνικού κοινωνικού ιστού. Πριν όμως εξεταστούν υπέστησαν κάποια προεπεξεργασία, δηλαδή αφαιρέθηκαν στοιχεία που ήταν περιττά. Επιπλέον έγινε εξομάλυνση του τονισμού των λέξεων και αφαίρεση emojis, δηλαδή εικονιδίων που εκφράζουν συναίσθημα. Επειδή όμως τα emojis έχουν την ιδιότητα να εκφράζουν αποτελεσματικά συναισθήματα, συμπεριλήφθηκαν με τη μορφή διανυσμάτων, τα οποία προσδιορίζουν τη συχνότητα εμφάνισης του κάθε emoji. Επιπλέον συμπεριλήφθηκαν και οι αναδημοσιεύσεις tweets χρηστών, υπο την προϋπόθεση ότι οι χρήστες συμμερίζονται το συναίσθημα που εκφράζεται σε αυτές.

Για τα πειράματα που πραγματοποιήθηκαν, ακολουθώντας τις διεργασίες που προαναφέρθηκαν, έγινε χρήση τριών τύπων σετ δεδομένων. Το 63,75% αφορούν δεδομένα που πάρθηκαν νωρίτερα και αποτελούν ένα σετ εκπαίδευσης. Το 11,25% είναι συνεχή στο χρόνο και αποτελούν σετ επικύρωσης. Το 25% πάρθηκε αργότερα και αποτελεί σετ δοκιμής. Η επικρατέστερη δραστηριότητα εξαγωγής χαρακτηριστικών περιλαμβάνει τα κείμενα κάθε καταγραφής της συλλογής. Αυτό γίνεται μέσω ενός BiLSTM δικτύου δύο επιπέδων, ενώ στις νευρικές ενσωματώσεις των λέξεων προστέθηκε θόρυβος Gaussian ($\mu = 0$, $\sigma = 1$). Ο βέλτιστος αριθμός μονάδων για κάθε στρώμα έχει καθοριστεί να είναι 150 με ($p = 0,3$). Έπειτα ακολουθεί ένα μονοδιάστατο συνελκτικό επίπεδο με 64 φίλτρα και παράθυρο μεγέθους 5. Η εξαγωγή χαρακτηριστικών κειμένου ολοκληρώνεται με ένα στρώμα προσοχής για την αντιμετώπιση μεγάλων προτάσεων. Τέλος, η διαδικασία εξαγωγής χαρακτηριστικών ολοκληρώνεται με ένα επίπεδο εγκατάλειψης ($p = 0,5$). Ο αριθμός επαναλήψεων φαίνεται μέσω λογαρίθμων. Το μοντέλο έχει καθοριστεί ως ένα πλήρως συνδεδεμένο τεχνητό νευρικό δίκτυο τροφοδοσίας, αποτελούμενο από δύο κρυμμένα στρώματα. Το πρώτο κρυφό στρώμα αποτελείται από 1024 νευρώνες και το δεύτερο από 128. Η εκπαίδευση στο δίκτυο βασίστηκε στον αλγόριθμο βελτιστοποίησης Adam.

Τα αποτελέσματα δείχνουν ότι όταν λαμβάνεται υπόψη μόνο το κείμενο της κάθε εγγραφής, η αποτελεσματικότητα της προτεινόμενης προσέγγισης είναι περιορισμένη. Όταν λαμβάνονται υπόψη οι σχετικές με την πτυχή πληροφορίες, η ανάκληση αυξάνεται πάνω από 6%. Η βαθμολογία F1 είναι βελτιωμένη. Τέλος όταν παρέχονται και οι τέσσερις είσοδοι (κείμενο, όψη, λογάριθμος του αριθμού των επαναλήψεων, emoji), η απόδοση του συστήματος βελτιώνεται περαιτέρω, με τις τρεις μετρήσεις να είναι πάνω (70%), προσθέτοντας περισσότερο από 8% στη συνολική απόδοση του συστήματος. Το σύστημα ανταπεξέρχεται αρκετά καλά σε αρνητικές τάξεις ενώ τα αποτελέσματα των θετικών κλάσεων είναι κάτω του μέσου όρου.

2.8 A Classifier to Distinguish Between Cypriot Greek and Standard Modern Greek

Η παρούσα εργασία πραγματεύεται το πρόβλημα διάκρισης γλωσσών και διαλέκτων που έχουν μεγάλη ομοιότητα μεταξύ τους. Παρουσιάζεται μία συλλογή διαφορετικών ελληνικών διαλέκτων και η κατασκευή ενός ταξινομητή για το διαχωρισμό μεταξύ των Κυπριακών ελληνικών (CG) και των στάνταρ νεοελληνικών (SMG). Τα σώματα κειμένων δημιουργήθηκαν από το Facebook, το Twitter και άλλα μέσα κοινωνικής δικτύωσης, όπως φόρουμ. Έγινε εξαγωγή N-gram χαρακτηριστικών και εφαρμογή τριών αλγορίθμων ταξινόμησης οι οποίοι δοκιμάστηκαν σε επισημασμένες προτάσεις: multinomial naive Bayes (NB), (SVC) και logistic regression.

Το σύστημα ταξινόμησης και ελέγχου λειτουργεί ως εξής. Επισημασμένες προτάσεις των CG και SMG φορτώνονται από το σώμα κειμένων ώστε να δεχθούν την απαραίτητη προεπεξεργασία. Στη συνέχεια εξαγάγονται επισημασμένες γραμμές κειμένου, οι οποίες εντάσσονται σε προτάσεις και έπειτα επεξεργάζονται πλήρως. Οι επισημασμένες προτάσεις είναι τυχαία χωρισμένες, έτσι ώστε το 80% είναι δεδομένα εκπαίδευσης και το 20% δεδομένα δοκιμής. Έπειτα εξαγάγονται από την εκπαίδευση χαρακτηριστικά N-gram και οι προτάσεις ελέγχονται και κωδικοποιούνται σε διανύσματα

χαρακτηριστικών. Το σετ εκπαίδευσης των διανυσμάτων χρησιμοποιείται για την εκπαίδευση των ταξινομητών, ενώ το σετ δοκιμής για την αξιολόγηση της απόδοσής τους.

Το τελικό σύνολο δεδομένων αποτελείται από 1039 προτάσεις. Ο αριθμός των λέξεων για το CG ήταν 7026 ενώ για το SMG 7100. Οι προτάσεις αφορούν διάφορα θέματα, όπως νέα, πολιτική, αθλήματα, χιούμορ. Ακολουθεί η προεπεξεργασία όπου αφαιρούνται σημεία στίξης και περιεχόμενο που εμποδίζει την κατανόηση των κειμένων από τον υπολογιστή. Μεταξύ αυτών αφαιρούνται URLs, HTML και διάφορες λέξεις κλειδιά, αφού προηγουμένως το κείμενο έχει επισημανθεί μέσω του NLTK1 της Python.

Στη συνέχεια γίνεται ανάλυση κειμένου στην οποία φαίνεται ότι τα CG και SMG διαφέρουν σημαντικά στις ορολογίες των λέξεων, στους χαρακτήρες και σε μετρήσεις n-gram. Οι δέκα πιο δημοφιλείς λέξεις του CG είναι μοναδικές στην κυπριακή διάλεκτο, συνεπώς μπορούν να συμβάλουν στην αποτελεσματική διάκριση των δύο διαλέκτων.

Κατα τη διεξαγωγή το πειράματος, οι ταξινομητές εκπαιδεύτηκαν και δοκιμάστηκαν πέντε φορές με διαφορετικούς συνδυασμούς σετ εκπαίδευσης, ώστε να προσδιοριστεί ποιός από τους ταξινομητές έχει καλύτερη απόδοση. Την καλύτερη απόδοση είχε ο multinomial NB με ακρίβεια ίση με 0.95. Στη συνέχεια εμβαθύνοντας στα χαρακτηριστικά, φαίνεται πως αυτά που προσδιορίζουν την κυπριακή διάλεκτο έχουν μεγαλύτερη βαρύτητα, καθώς είναι αυτά που διαχωρίζουν τις δύο διαλέκτους. Τέλος ακολουθούν τρία πειράματα ώστε να εξεταστεί αν με λιγότερα χαρακτηριστικά θα υπάρξει διαφορά στην επίδοση. Στο πρώτο πείραμα γίνεται αφαίρεση της επένδυσης χαρακτήρων n-gram. Τα αποτελέσματα έδειξαν ότι η επένδυση σε αυτούς τους χαρακτήρες κατέχει πολύ σημαντικό ρόλο στη διαδικασία της ταξινόμησης. Στο επόμενο πείραμα διερευνήθηκε η χρήση n-gram μόνο σε χαρακτήρες και όχι σε λέξεις. Τα αποτελέσματα δείχνουν ότι η χρήση n-gram, σε λέξεις συμβάλει στην καλύτερη ακρίβεια. Στο τελευταίο πείραμα αγνοήθηκαν χαρακτηριστικά που κατέχουν λιγότερο από το 1% των δεδομένων. Εν τέλη αυτό το πείραμα έδειξε ότι η αφαίρεση λέξεων που δεν εμφανίζονται συχνά επηρεάζει αρνητικά την ακρίβεια και αυτό

μπορεί να συμβαίνει λόγω του μικρού μεγέθους του συνόλου δεδομένων που χρησιμοποιήθηκε.

Συμπεραίνουμε ότι αυτές οι τεχνικές θα μπορούσαν να χρησιμοποιηθούν για τη διάκριση διαλέκτων , το διαχωρισμό μίας αρχαίας γλώσσας από τη σύγχρονή της ή ακόμα στην περίπτωση των ελληνικών να διαχωριστούν τα ελληνικά της καθομιλουμένης από αυτά της καθαρεύουσας. Επιπλέον αυτή η προσέγγιση θα μπορούσε να συμβάλει στην αναγνώριση των greeklish και το συσχετισμό τους με τα ελληνικά.

2.9 Argument Extraction from News, Blogs, and the Social Web

Η συγκεκριμένη εργασία πραγματεύεται την εξαγωγή επιχειρημάτων από αναρτήσεις του ελληνικού κοινωνικού ιστού. Ουσιαστικά πρόκειται για μία διεργασία εξαγωγής συμπερασμάτων και επιχειρημάτων μαζί με τους υποστηρικτικούς συντελεστές τους. Το ζήτημα που αφορά τη μελέτη αυτή είναι οι ανανεώσιμες πηγές ενέργειας. Κύριος στόχος της εργασίας δεν είναι η εύρεση του αλγόριθμου μηχανικής μάθησης με την καλύτερη απόδοση, αλλά η μελέτη των πιο σύγχρονων χαρακτηριστικών στον τομέα των κοινωνικών δικτύων.

Γίνεται χρήση του Argument Extraction (ΑΕ) όπου εφαρμόζονται δύο βήματα: α) Προσδιορίζονται τα επιχειρήματα που σχετίζονται με αυτά που έχουν συνταχθεί στο μοντέλο πολιτικής, μέσω του εργαλείου NOMAD. Τα αποτελέσματα είναι προτάσεις ή τμήματα προτάσεων που πληρούν το παραπάνω κριτήριο.

Β) Προσδιορισμός νέων συντελεστών επιχειρημάτων που δε σχετίζονται με αυτά του μοντέλου πολιτικής. Τα αποτελέσματα είναι μόνο αναγνωρισμένες προτάσεις και τμήματα προτάσεων.

Το δεύτερο βήμα είναι πιο αποδοτικό έτσι εφαρμόζεται η παρακάτω προσέγγιση: Ταυτοποιούνται προτάσεις και τμήματα προτάσεων που περιέχουν συντελεστές επιχειρήματος, λαμβάνοντας υπόψη μόνο τα γλωσσικά στοιχεία και όχι το μοντέλο πολιτικής. Τα επιχειρήματα αυτά συγκρίνονται και αντιπαραβάλλονται με αυτά του μοντέλου πολιτικής. Όποια επιχειρήματα δεν ταιριάζουν με αυτά του μοντέλου πολιτικής

,επισημαίνονται σε πιθανά νέα επιχειρήματα, για μελλοντική προσθήκη στο μοντέλο.

Θεωρώντας τον προσδιορισμό των επιχειρηματολογικών προτάσεων ως εποπτευόμενη ταξινόμηση γίνεται χρήση ταξινομητών μηχανικής μάθησης, όπως το Logistic Regression, Random Forest, Support Vector Machines και Naive Bayes.

Στη συνέχεια γίνεται χρήση Conditional Random Fields (CRFs). Το CRFs είναι ένας δομημένος αλγόριθμος πρόβλεψης, που μπορεί να ταυτοποιήσει επιχειρήματα και θέσεις, ενώ συμβάλλει στη διατήρηση της σειράς των λέξεων στην πρόταση. Έπειτα γίνεται χρήση Sentence-level argument matcher. Αυτή η μεθοδολογία μπορεί να διαχωριστεί σε μια φάση εκπαίδευσης και αξιολόγησης, ή σε κατάσταση “off-line” ή “on-line”. Στόχος της “off-line” κατάστασης είναι η δημιουργία ενός χώρου χαρακτηριστικών που επιτρέπει την κωδικοποίηση της πολιτικής σε διάνυσμα. Γίνεται χρήση TF-IDF για την επιλογή των λέξεων που αντιπροσωπεύουν καλύτερα ένα επιχείρημα πολιτικής σε σχέση με τα υπόλοιπα επιχειρήματα. Κάθε δημιουργημένο διάνυσμα χαρακτηριστικών έχει σταθερό μήκος, ίσο με το μέγεθος του χώρου δυνατοτήτων και κάθε χαρακτηριστικό είναι θετικός αριθμός, ίσος με το σκορ TF-IDF. Η “on-line” κατάσταση συσχετίζει μία επιχειρηματολογική πρόταση με το πιο πανομοιότυπο επιχείρημα πολιτικής. Κάθε επιχειρηματολογική πρόταση μετατρέπεται σε ένα δείκτη χαρακτηριστικών χρησιμοποιώντας το χώρο δυνατοτήτων της κατάστασης “off-line” και οι βαθμολογίες TF-IDF χρησιμοποιούνται ως τιμές χαρακτηριστικών. Στη συνέχεια με τη βοήθεια του k-NN, με $k = 1$, το διάνυσμα ταξινομείται σε μόνο ένα όρισμα πολιτικής.

Τα δεδομένα των πειραμάτων που πραγματοποιήθηκαν αποτελούνται από ένα σώμα 204 κειμένων του ελληνικού κοινωνικού ιστού (Bing, Google Blogger, Google Plus, Twitter,...). Δεν έγινε χρήση κειμένων micro-blogging, θεωρώντας ότι ο περιορισμένος αριθμός χαρακτήρων δεν επιτρέπει την πλήρη έκφραση του χρήστη. Τα επιλεγμένα κείμενα επισημάνθηκαν ως θετικά ή αρνητικά. Για τη διεξαγωγή των αποτελεσμάτων, αρχικά ως πρώτο βήμα εφαρμόστηκε η μεθοδολογία 10-fold cross validation στο σύνολο των 16000 προτάσεων ενώ το

σύνολο δεδομένων χωρίστηκε σε δύο τμήματα. Το πρώτο μέρος περιείχε το 70% των περιπτώσεων και χρησιμοποιήθηκε σαν σετ εκπαίδευσης ενώ το δεύτερο το 30% χρησιμοποιήθηκε σαν σετ εκπαίδευσης.

Στη συνέχεια πραγματοποιείται χρήση της τεχνικής Synthetic Minority Over-sampling Technique (SMOTE), μία τεχνική που αναγνωρίζει το γραφικό χαρακτήρα. Μέσω αυτής της τεχνικής αποδεικνύεται ότι ο συνδυασμός υπερδειγματοληψίας της μειοψηφίας και χαμηλότερης δειγματοληψίας της πλειοψηφίας, μπορεί να πετύχει καλύτερη απόδοση ταξινομητή.

Στο δεύτερο βήμα χρησιμοποιούνται τυχαία πεδία για τον εντοπισμό υπόθεσης σε πρόταση με χρήση της αναπαράστασης BIOES. Όπου B η πρώτη υπόθεση, I μία οποιαδήποτε άλλη και O οι υπόλοιπες. Έπειτα από την επισήμανση που πραγματοποιήθηκε στις προτάσεις τα αποτελέσματα και για τα δύο βήματα είναι πάνω από το βασικό σενάριο. Τέλος πραγματοποιείται αξιολόγηση προτάσεων και τμημάτων προτάσεων με τη βοήθεια του επισημασμένου σώματος κειμένων NOMAD και με τη χρήση 10-fold cross validation.

Τέλος παρουσιάζονται στον πίνακα 8 τα αποτελέσματα της αξιολόγησης προτάσεων και στον πίνακα 9 τα αποτελέσματα της αξιολόγησης τμημάτων προτάσεων με τη βοήθεια του επισημασμένου σώματος κειμένων NOMAD και με τη χρήση 10-fold cross validation.

2.10 A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets

Η παρούσα εργασία αφορά την ανίχνευση ειρωνίας σε σώματα ελληνικών κειμένων πολιτικού περιεχομένου του κοινωνικού ιστού και συγκεκριμένα του Twitter. Στόχος της είναι, συγκρίνοντας δεδομένα τα οποία έχουν εκπαιδευτεί με ημι-επιβλεπόμενες τεχνικές και δεδομένα που εκπαιδεύτηκαν με επιβλεπόμενες τεχνικές από προηγούμενη εργασία, να αποδείξει ότι μέσα από τις αναρτήσεις στο Twitter είναι ευφικτή η πρόβλεψη των πραγματικών αποτελεσμάτων των εκλογών.

Το σύνολο δεδομένων αποτελείται από 61.427 ελληνικά tweets , τα οποία συγκεντρώθηκαν μία εβδομάδα μετά τις εκλογές του 2012. Τα δεδομένα χωρίζονται σε παρατάξεις και αρχηγούς κομμάτων. Η προεπεξεργασία των δεδομένων περιλαμβάνει την αφαίρεση διπλώτυπων και αναδημοσιευμένων tweets ενώ διατηρήθηκαν tweets που περιλαμβάνουν υπερσυνδέσμους . Ο τελικός αριθμός των tweets είναι 44.438. Η ανίχνευση ειρωνίας προσεγγίστηκε ως πρόβλημα ταξινόμησης κειμένων, ενώ η απόφαση αν ένα tweet είναι ειρωνικό ως δυαδική. Τα tweets επισημάνθηκαν με βάση τα εξής χαρακτηριστικά: Ομιλία, σπανιότητα, σημασία, λεξιλογικά κριτήρια και emoticons.

Στην περίπτωση της επιβλεπόμενης τεχνικής , έγινε επισημάνση 126 tweets (74 ειρωνικά , 52 όχι) και στη συνέχεια πραγματοποιήθηκε εκπαίδευση των δεδομένων με τη χρήση των εξής αλγορίθμων: j48, SVM, νευρωνικά δίκτυα , Naive Bayes , Functional Trees, KStar και Random Forests. Την καλύτερη απόδοση είχαν τα Functional Trees. Επειδή τα δεδομένα δε φέρουν ετικέτα , γίνεται σύγκριση των πραγματικών αποτελεσμάτων των εκλογών και του όγκου της ειρωνίας που ανιχνεύτηκε για την πιο έγκυρη αξιολόγηση του μοντέλου.

Στην περίπτωση της ημι-επιβλεπόμενης τεχνικής έγινε χρήση του αλγόριθμου collective-tree , ο οποίος είναι παρόμοιος με το Random Tree, αλλά λαμβάνει υπόψη και τα επισημασμένα και τα μη επισημασμένα δεδομένα. Ο αλγόριθμος χωρίζει τα χαρακτηριστικά σε δύο μισά. Για την εκπαίδευση χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι με αυτούς της επιβλεπόμενης τεχνικής και την καλύτερη απόδοση είχε το Random Forest. Ενώ και στις δύο περιπτώσεις το Naive Bayes είχε τη χειρότερη απόδοση.

Εν κατακλείδι τα ποσοστά ειρωνίας δε συσχετίζονται άμεσα με τα αποτελέσματα των εκλογών , ωστόσο υπάρχει μία σύνδεση της ειρωνίας στις αναρτήσεις και των διακυμάνσεων των ψήφων που λαμβάνουν οι παρατάξεις. Επιπλέον τα ποσοστά ακρίβειας είναι παρόμοια και στις δύο περιπτώσεις ενώ η ανάκληση είναι χαμηλότερη στην ημι-επιβλεπόμενη τεχνική.

2.11 Political sentiment analysis of tweets before and after the Greek elections of May 2012

Η παρούσα εργασίαπραματεύεται ελληνικά tweets που αφορούν τις εκλογές του 2012.Στόχος της είναι ο συσχετισμός των προεκλογικών tweets των χρηστών και των τελικών αποτελεσμάτων εκλογών ή δημοσκοπήσεων.Επιπλέον επεξεργάζεται tweets τα οποία αναρτήθηκαν μετά τις εκλογές και μελετά την επιρροή των αποτελεσμάτων των εκλογών στην ψυχосύνθεση των ανθρώπων. Γίνεται χρήση ξεχωριστού λεξικού για κάθε παράταξη για την καλύτερη διαχείριση των δεδομένων ενώ παράλληλα οι λέξεις κλειδιά μπορεί να είναι σε διαφορετικές μορφές όπως greeklish , με κεφαλαία γράμματα ή τονισμένες ,ώστε να δημιουργηθεί μία πιο εκτενής λίστα tweets.Στη συνέχεια γίνεται σύγκριση των tweets πριν και μετά τις εκλογές , καθώς και των αποτελεσμάτων των εκλογών με τα εν λόγω tweets.Τα αποτελέσματα της έρευνας δείχνουν ότι πραγματικά είναι εφικτό να προβλέψει κανείς τα αποτελέσματα των εκλογών βάση των αναρτήσεων στο twitter με μεγαλύτερη ακρίβεια σε σχέση με τις δημοσκοπίσεις.Κατα την ανάλυση συναισθήματος κάθε καταχώρηση λεξιλογίου επισημάνθηκε ανάλογα με το αν το συναίσθημα που εκφράζεται είναι θετικό , αρνητικό ή ουδέτερο.Στη συνέχεια μέσω μη επιβλεπόμενης μάθησης τα tweets χαρακτηρίστηκαν ως θετικά,αρνητικά ή ουδέτερα ανάλογα με το συναίσθημα που εκφραζόταν από το λεξιλόγιο του καθενός.Στη συνέχεια παρατίθενται διαγράμματα που παρουσιάζουν τη διαφορά μεταξύ των θετικών και αρνητικών tweets πριν και μετά τις εκλογές .Μία σημαντική παρατήρηση είναι ότι σε tweets τα οποία αφορούν μικρότερα κόμματα εκφράζονται περισσότερα συναισθήματα σε σχέση με αυτά που αφορούν μεγαλύτερα,ενώ τα δύο άκρα δεξιά και αριστερά συγκέντρωσαν περισσότερα θετικά πριν τις εκλογές.Συμπεραίνουμε λοιπόν ότι τα tweets πέρα από ότι μπορούν να συμβάλουν στην πρόβλεψη της έκβασης των εκλογικών αποτελεσμάτων,μπορούν να μας δώσουν και μία εικόνα το πως οι ίδιες οι εκλογές επηρεάζουν τα συναισθήματα των ανθρώπων.

3.Αποτελέσματα πειραμάτων εργασιών

Λαμβάνοντας υπόψη τη βιβλιογραφική ανασκόπηση στις ερευνητικές μελέτες σχετικά με τις τεχνικές εξόρυξης γνώσης από σώματα ελληνικών κειμένων του κοινωνικού ιστού γίνεται αντιληπτό πως η μηχανική μάθηση είναι μία τεχνολογία που εξελίσσεται συνεχώς. Συνεπώς υπάρχουν αρκετές τεχνικές και διαδικασίες εξόρυξης της γνώσης που παρέχονται από τον κοινωνικό ιστό.

Οι μελέτες που διερευνήθηκαν στην παρούσα εργασία χρησιμοποιούν διάφορες προσεγγίσεις και μεθόδους. Ο συγχρωτισμός τους θα μπορούσε να συμβάλει στην αποτελεσματικότερη μελέτη των τεχνικών εξόρυξης γνώσης και να αποτελέσει τροφή για την εκπόνηση νέων μελετών πάνω στο συγκεκριμένο αντικείμενο. Παρακάτω παρατίθενται και αναλύονται τα αποτελέσματα των εργασιών που χρησιμοποιήθηκαν σε πίνακες που πάρθηκαν από τις συγκεκριμένες μελέτες.

3.1 Για την πρώτη έρευνα που αφορά τη μελέτη[1] τα δεδομένα πάρθηκαν για εμπειρική αξιολόγηση με χειροκίνητο τρόπο από ένα σύνολο 2.300 κειμένων σε χρονικό διάστημα 2 μηνών που αφορούν τον πολιτικό τομέα των ελληνικών εφημερίδων “Real News” και “Καθημερινή”. Για τον εντοπισμό αναφορών οντοτήτων η απόδοση μετρήθηκε πάνω από 93% ενώ για την πολικότητα μεταξύ τους είχε χαμηλότερα επίπεδα απόδοσης τα οποία έφταναν στο 64% ενώ η επιθυμητή τιμή τους ήταν το 85%.

Για τη δεύτερη έρευνα που αφορά τη μελέτη [1] τα δεδομένα πάρθηκαν με αυτόματο τρόπο και αφορούν δύο οντότητες την τηλεφωνική εταιρεία “Vodafone” και την τράπεζα “Alpha Bank”. Τα αποτελέσματα της εμπειρικής αξιολόγησής τους έδωσαν ακρίβεια 80%. Έτσι προέκυψαν οι δύο παρακάτω πίνακες.

Table 1. Evaluation results for the entity “Vodafone”

	Correct	Total	Accuracy
Sites	7	9	77.78 %
Blogs	18	18	100.00 %
Facebook	166	183	90.71 %
Twitter	122	158	77.22 %
Overall	313	371	84.37 %

Table 2. Evaluation results for the entity “Alpha Bank”

	Correct	Total	Accuracy
Sites	62	78	79.49 %
Blogs	55	69	79.49 %
Facebook	7	8	87.50 %
Twitter	100	129	77.52 %
Overall	223	282	79.08 %

3.2 Για τη μελέτη [2] χρησιμοποιήθηκαν διαφορετικοί συνδυασμοί από μία ομάδα χαρακτηριστικών(TF / IDF, unigrams, bigrams, POS και ετικέτες σχέσης εξάρτησης) για την πραγματοποίηση ενός πειράματος που συγκρίνει την απόδοση μεμονωμένων ταξινομητών για την αναγνώριση υβριστικής φρασεολογίας και προσβλητικού περιεχομένου σε σώματα κειμένων που αφορούν ελληνικές αναρτήσεις στο Twitter.Τα αποτελέσματα των πειραμάτων τα οποία συνοδεύονται από τους ανάλογους πίνακες είναι τα εξής:

Για την κλασσική μηχανική μάθηση:

Για τα χαρακτηριστικά TF/IDF unigram [Table 2] οι δύο γραμμικοί ταξινομητές SVM και SGDC ξεπερνούν τους άλλους με τιμή 0.80 όσον αφορά το macro-F1.Το Linear SVM έχει ελαφρώς καλύτερη βαθμολογία στην κατηγορία Not Offensive ενώ το SGDC στη κατηγορία Offensive.Στην κατηγορία Offensive είχε τα υψηλότερα αποτελέσματα

to Bernoulli Naïve Bayes.

	Not Offensive			Offensive			Weighted Average			
Model	P	R	F1	P	R	F1	P	R	F1	F1 Macro
<i>Linear SVM</i>	0.83	0.98	0.90	0.92	0.57	0.70	0.86	0.85	0.84	0.80
<i>RBF SVM</i>	0.76	0.99	0.86	0.96	0.31	0.47	0.82	0.78	0.74	0.66
<i>SGDC</i>	0.84	0.96	0.90	0.86	0.61	0.71	0.85	0.85	0.84	0.80
<i>Multinomial NB</i>	0.77	0.99	0.86	0.94	0.33	0.49	0.82	0.78	0.75	0.67
<i>Bernoulli NB</i>	0.83	0.89	0.86	0.71	0.61	0.66	0.80	0.80	0.80	0.76

Table 2: Results for offensive language detection with TF/IDF unigram features. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

Τα μοντέλα εκπαιδευμένα με χαρακτηριστικά TF/IDF bigram [Table 3] είχαν χειρότερη απόδοση με όλες τις τιμές χαμηλότερες από αυτές του TF/IDF unigram , με εξαίρεση το Multinomial Naïve Bayes που βελτιώθηκε στο macro-F1 για την κατηγορία «Μη προσβλητικό».

	Not Offensive			Offensive			Weighted Average			
Model	P	R	F1	P	R	F1	P	R	F1	F1 Macro
<i>Linear SVM</i>	0.82	0.98	0.90	0.92	0.54	0.68	0.86	0.84	0.83	0.79
<i>RBF SVM</i>	0.74	1.00	0.85	0.98	0.24	0.39	0.82	0.76	0.71	0.62
<i>SGDC</i>	0.84	0.94	0.89	0.81	0.61	0.69	0.83	0.83	0.83	0.79
<i>Multinomial NB</i>	0.77	0.99	0.87	0.93	0.32	0.48	0.82	0.79	0.75	0.67
<i>Bernoulli NB</i>	0.82	0.88	0.85	0.68	0.57	0.62	0.78	0.79	0.78	0.74

Table 3: Results for offensive language detection with TF/IDF bigram features. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

Στη συνέχεια πραγματοποιήθηκαν τρία πειράματα με γλωσσικά χαρακτηριστικά για να ελεγχθεί η αποτελεσματικότητά τους.

Στο πρώτο πείραμα[Table 4] οι λειτουργίες TF / IDF unigram συνδυάστηκαν με POS και ετικέτες σχέσης εξάρτησης. Το μοντέλο Linear SVM είχε καλύτερη βαθμολογία ανάκλησης σε σχέση με το προηγούμενο μοντέλο που εκπαιδεύτηκε με bigrams, ενώ τα υπόλοιπα εμφανίζουν σημαντική πτώση στην απόδοσή τους.

	Not Offensive			Offensive			Weighted Average			
Model	P	R	F1	P	R	F1	P	R	F1	F1 Macro
<i>Linear SVM</i>	0.84	0.96	0.90	0.88	0.58	0.70	0.85	0.85	0.83	0.80
<i>SGDC</i>	0.80	0.95	0.87	0.81	0.48	0.61	0.81	0.80	0.79	0.74
<i>Multinomial NB</i>	0.77	0.95	0.85	0.78	0.36	0.49	0.77	0.77	0.74	0.67
<i>Bernoulli NB</i>	0.80	0.78	0.79	0.54	0.58	0.56	0.72	0.72	0.72	0.68

Table 4: Results for offensive language detection with TF/IDF unigram features, POS and dependency relation tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

Στο δεύτερο πείραμα[Table 5] οι λειτουργίες TF / IDF unigram συνδυάστηκαν με POS. Το Linear SVM αποδίδει την ίδια βαθμολογία F1 Macro με το πρώτο πείραμα με οριακή βελτίωση στην κατηγορία Offensive. Η απόδοση του SGDC μειώνεται σημαντικά ενώ η βαθμολογία ανάκλησής της για την κατηγορία Offensive έχει τα χειρότερα αποτελέσματα συγκριτικά με τους άλλους ταξινομητές.

Model	Not Offensive			Offensive			Weighted Average			F1 Macro
	P	R	F1	P	R	F1	P	R	F1	
<i>Linear SVM</i>	0.84	0.97	0.90	0.91	0.58	0.71	0.86	0.85	0.84	0.80
<i>SGDC</i>	0.74	0.99	0.85	0.93	0.22	0.35	0.80	0.75	0.69	0.60
<i>Multinomial NB</i>	0.77	0.99	0.86	0.93	0.33	0.49	0.82	0.78	0.75	0.68
<i>Bernoulli NB</i>	0.83	0.86	0.84	0.66	0.61	0.63	0.78	0.78	0.78	0.74

Table 5: Results for offensive language detection with TF/IDF unigram features and POS tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

Τέλος στο τρίτο πείραμα[Table 6] οι λειτουργίες TF / IDF unigram συνδυάστηκαν με ετικέτες σχέσης εξάρτησης. Η απόδοση του Linear SVM στο F1 Macro είναι ίδια με αυτή του προηγούμενου πειράματος. Ξεπερνά κατά 0.01% την κατηγορία Offensive του προηγούμενου πειράματος ενώ υστερεί στην κατηγορία Not Offensive. Η απόδοση του το SGDC αυξήθηκε ,κάτι που ισχύει και για το Naïve Bayes καθώς και οι Multinomial και οι Bernoulli προσεγγίσεις του είχαν υψηλότερη απόδοση από το δεύτερο πείραμα.

Model	Not Offensive			Offensive			Weighted Average			F1 Macro
	P	R	F1	P	R	F1	P	R	F1	
<i>Linear SVM</i>	0.84	0.97	0.90	0.90	0.58	0.70	0.86	0.85	0.84	0.80
<i>SGDC</i>	0.87	0.66	0.75	0.51	0.78	0.61	0.76	0.70	0.71	0.68
<i>Multinomial NB</i>	0.77	0.97	0.86	0.85	0.35	0.49	0.79	0.78	0.74	0.67
<i>Bernoulli NB</i>	0.82	0.81	0.81	0.58	0.60	0.59	0.74	0.74	0.74	0.70

Table 6: Results for offensive language detection with TF/IDF unigram features and dependency relation tags. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

Για τη βαθιά μάθηση:

Το μοντέλο « LSTM and GRU with Attention» [Table 7]ξεπέρασε όλα τα υπόλοιπα μοντέλα σε ακρίβεια ,ανάκληση και F1 Macro και για τις δύο κατηγορίες Offensive και Not Offensive. Ωστόσο το μοντέλο BERT-Base Multilingual Cased δεν είχε τόσο καλά αποτελέσματα. Εν κατακλείδι το μοντέλο « LSTM and GRU with Attention» μπορεί να θεωρηθεί το

καλύτερο μοντέλο που έχει εκπαιδευτεί για το OGTD.

Model	Not Offensive			Offensive			Weighted Average			F1 Macro
	P	R	F1	P	R	F1	P	R	F1	
<i>Pooled GRU</i>	0.90	0.99	0.95	0.94	0.65	0.75	0.91	0.86	0.86	0.87
<i>Stacked LSTM with Attention</i>	0.91	0.99	0.96	0.95	0.66	0.76	0.92	0.87	0.87	0.88
<i>LSTM and GRU with Attention</i>	0.92	0.99	0.96	0.96	0.68	0.77	0.93	0.88	0.88	0.89
<i>2D Convolution with Pooling</i>	0.91	0.98	0.96	0.95	0.64	0.74	0.90	0.86	0.85	0.88
<i>GRU with Capsule</i>	0.92	0.99	0.95	0.94	0.64	0.75	0.91	0.86	0.85	0.88
<i>LSTM with Capsule and Attention</i>	0.91	0.98	0.95	0.94	0.66	0.75	0.90	0.86	0.86	0.87
<i>BERT-Base Multilingual Cased</i>	0.85	0.84	0.84	0.65	0.60	0.58	0.77	0.76	0.75	0.73

Table 7: Results for offensive language detection for Deep Learning models with Greek word embeddings. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed (best in bold).

3.3 Στη μελέτη[3] έχουμε μία σύγκριση μεταξύ δύο διαφορετικών μεθόδων ταξινόμησης :Της τεχνικής TF/IDF bag-of-words και της τεχνικής Term Occurrence (TO).Στον πίνακα 1 παρουσιάζονται τα αποτελέσματα για το TF/IDF bag-of-words .Σε αυτή την προσέγγιση έχουμε ακρίβεια 95.78 % , ένα αρκετά ικανοποιητικό ποσοστό ,καθώς έγινε σωστή κατηγοριοποίηση στις 1.724 από τις 1.800 κριτικές ξενοδοχείων.

Table 1 Results of the TF-IDF bag-of-words approach

	Predicted negative	Predicted positive
Negative cases	880	20
Positive cases	56	844
Accuracy	95.78 %	
Recall	93.78 %	
Precision	97.69 %	

Στη δεύτερη προσέγγιση έχουμε ποσοστό ακρίβειας ίσο με 71.76 %,ένα ποσοστό αρκετά μικρότερο από αυτό τις πρώτης περίπτωσης, καθώς ταξινομήθηκαν σωστά μόλις 1.222 κριτικές. Αξίζει να σημειωθεί ότι το ποσοστό ανάκλησης ήταν ίσο με 100%.

Τα αποτελέσματα παρατίθενται αναλυτικότερα στον πίνακα 2.

Table 2 Results of the TO approach

	Predicted negative	Predicted positive
Negative cases	323	481
Positive cases	0	899
Accuracy	71.76 %	
Recall	100 %	
Precision	65.14 %	

Συνεπώς οδηγούμαστε στο συμπέρασμα ότι η μέθοδος TF/IDF bag-of-words υπερτερεί της μεθόδου Term Occurrence (TO) .Παρόλο που η δεύτερη έχει ποσοστό ανάκλησης 100% το ποσοστό ακρίβειάς της είναι αρκετά χαμηλό.

3.4 Η μελέτη [4] αφορά την εύρεση του συγγραφέα ενός Blog καθώς και του φύλου του μέσω της εκπαίδευσης αλγορίθμων μηχανικής μάθησης. Αρχικά μέσω του αλγόριθμου Sequential Minimal Optimization (SMO) που αποτελεί μια βελτιστοποιημένη έκδοση του Support Vector Machines (SVMs) και μέσω μίας μεθοδολογίας 10-fold cross-validation εκπαιδεύτηκαν δεδομένα από 20 ελληνικά Blogs ,τα 10 εκ των οποίων είχαν άντρες συγγραφείς και τα άλλα 10 γυναίκες .Επιλέχθηκαν 50 αναρτήσεις από τον κάθε συγγραφέα. Τα αποτελέσματα είχαν 85.4% ακρίβεια για απόδοση συγγραφής και 82.6% για την ταυτοποίηση του φύλου.

Στη συνέχεια πραγματοποιήθηκε ένα πείραμα ώστε να διερευνηθεί κατα πόσο ο αριθμός των συγγραφέων επηρεάζει τα αποτελέσματα τις ακρίβειας. Δημιουργήθηκαν 4 ομάδες (2,4,8,16 συγγραφέων),η κάθε μία εκ των οποίων περιλάμβανε 10 διαφορετικούς συνδυασμούς συγγραφέων. Πραγματοποιήθηκαν 40 (4 x 10)πειράματα ταξινόμησης με χρήση SMO και 10-fold cross-validation. [Fig.1]Η υψηλότερη ακρίβεια επιτεύχθηκε στην περίπτωση των 2 συγγραφέων με ποσοστό 97.7% ,ενώ η χαμηλότερη στην περίπτωση των 16 με ποσοστό 86.9%.

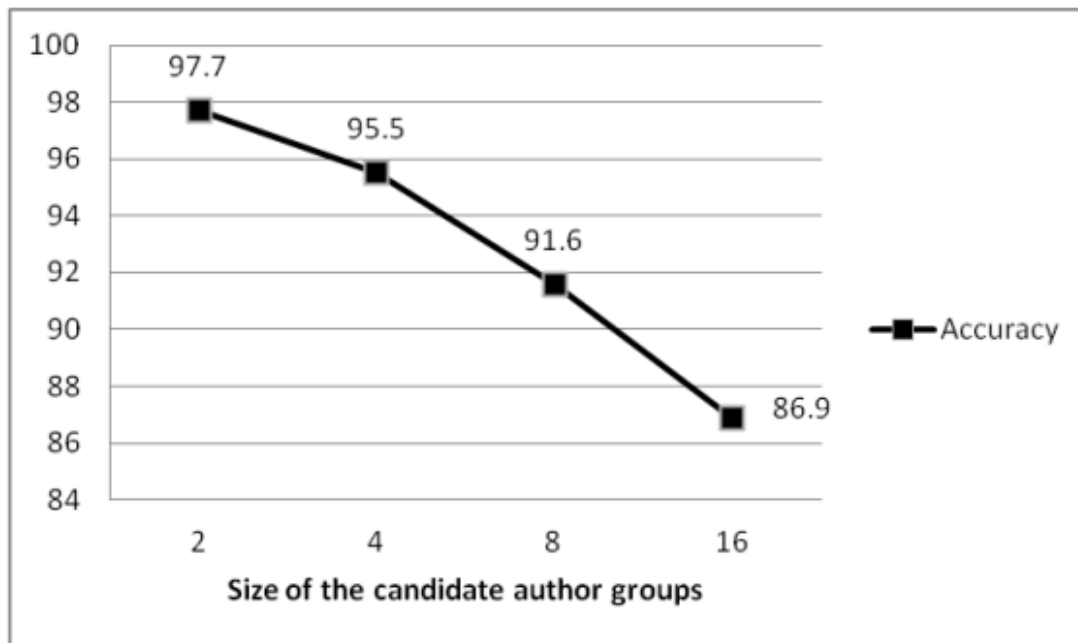


Fig. 1. Influence of the number of candidate authors

Συμπεραίνουμε λοιπόν ότι τα αποτελέσματα καταλαμβάνουν υψηλότερα ποσοστά ακρίβειας όταν ο αριθμός των συγγραφέων είναι μικρότερος, ενώ όσο ο αριθμός αυτός μεγαλώνει τα ποσοστά ακρίβειας μειώνονται γραμμικά.

3.5. Το πείραμα της συγκεκριμένης μελέτης [5], αφορά την εξόρυξη γνώσης από tweets για την πρόβλεψη της έκβασης των αποτελεσμάτων ψηφοφορίας εφόσον αυτή έχει πιθανές απαντήσεις ναι ή όχι. Γίνεται σύγκριση δύο λειτουργιών TEXT και NETWORK καθώς και των μοντέλων Convolution Kernel (SVM, MCKL) και Baselines (LR, FF, RF, SVM). Πραγματοποιήθηκε καθημερινή καταμέτρηση και τα αποτελέσματα έδειξαν ότι όσο πιο κοντά γίνονταν οι μετρήσεις στην ψηφοφορία τόσο μεγαλύτερη η ακρίβεια των αποτελεσμάτων.

Στο Figure 2 και στο Table 3 παρουσιάζονται αναλυτικά τα αποτελέσματα macro-F καθόλη τη διάρκεια των μετρήσεων.

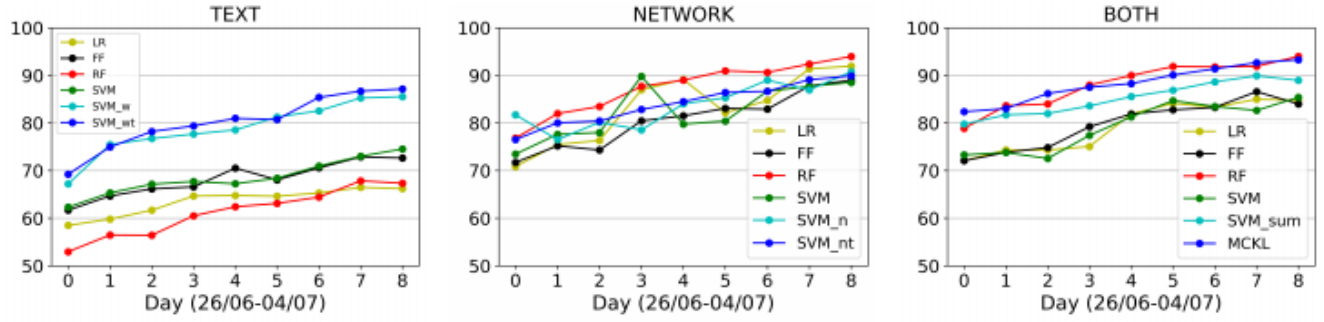


Figure 2: Macro-average F-score across all evaluation days using TEXT, NETWORK and BOTH user representations.

Table 3: Average F-score and standard deviation across all evaluation days using TEXT, NETWORK and BOTH user representations. SVM_s and SVM_{st} denote the SVM with convolution kernels (SVM_w , SVM_n) and (SVM_{wt} , SVM_{nt}), respectively.

	TEXT	NETWORK	BOTH
LR	63.55 $\pm$ 2.86	83.21 $\pm$ 7.55	79.43 $\pm$ 5.34
FF	68.19 $\pm$ 3.78	80.66 $\pm$ 5.93	79.83 $\pm$ 5.11
RF	61.27 $\pm$ 5.14	87.43 $\pm$ 5.60	88.22 $\pm$ 5.03
SVM	68.51 $\pm$ 3.80	82.43 $\pm$ 5.83	79.39 $\pm$ 5.18
SVM_s	78.91 $\pm$ 5.68	83.65 $\pm$ 4.82	–
SVM_{st}	80.30 $\pm$ 5.81	84.03 $\pm$ 4.47	–
SVM_{sum}	–	–	85.22 $\pm$ 3.64
MCKL	–	–	88.31 $\pm$ 3.95

Με τη χρήση του TEST (SVM) έχουμε 11.8% μεγαλύτερη απόλυτη βελτίωση και 17.2% σχετική βελτίωση.

Με τη χρήση του NETWORK βελτιώνεται η ακρίβεια της ταξινόμησης. Την ημέρα πριν από το δημοψήφισμα ο RF επιτυγχάνει

94% ποσοστό F-score, ενώ η χειρότερη απόδοση (FF) έχει ποσοστό F-score ίσο με 80,66%.

Την καλύτερη απόδοση επιτυγχάνει το MCKL συνδυάζοντας την προσέγγιση TEST με τη NETWORK με ποσοστό F-score ίσο με 20%.

3.6 Στη μελέτη [6] το πείραμα που πραγματοποιείται αφορά την ανάλυση συναισθήματος από ένα σύνολο αναρτήσεων στο Twitter καθώς και από το ανάλογο σύνολο hashtags. Ουσιαστικά επιλέχθηκαν ορισμένα tweets για να εξεταστεί αν οι μέθοδοι που χρησιμοποιήθηκαν για την εκτίμηση συναισθήματος είναι αποτελεσματικές.

TABLE II
EXAMPLE RATINGS

#	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1	1.00	1.00	1.00	4.75	1.00	2.75
2	3.50	3.50	1.00	1.00	1.00	2.50
3	2.58	2.00	0.79	0.79	0.95	1.63
4	1.00	1.00	1.00	3.75	1.00	2.50

TABLE III
HASHTAGS RATINGS

#	Anger	Disgust	Fear	Happiness	Sadness	Surprise
#wc14gr	1.3910	1.2862	0.9512	1.3604	0.8412	1.4552
#ekloges14	1.1627	1.1676	0.8180	1.1456	0.7219	1.2885
#kalokairipantou	0.7930	0.9158	0.7739	2.1856	0.7570	2.1084
#skouries	1.0608	1.0460	0.9399	1.0603	0.7337	1.1197
#panellinies2014	1.3900	1.3374	0.9810	1.4521	0.8153	1.4659
#vouli	1.3040	1.2608	0.7832	1.1767	0.7419	1.3122
#ert	1.0892	1.0757	0.8065	1.0242	0.6694	1.1292
#mb14gr	1.3948	1.2742	0.9510	1.3451	0.8041	1.4225
#eurovisiongr	1.3464	1.2957	0.7933	1.3533	0.7599	1.4092
#enikos	1.3189	1.2866	0.8195	1.1918	0.7616	1.3551

Όπως φαίνεται στους παραπάνω πίνακες η προσέγγιση αυτή είναι ικανή να υπολογίσει ,βάση κοινής λογικής ορθά τα συναισθήματα που αφορούν τον κάθε τομέα.

TABLE IV
SENTIMENT PEARSON CORRELATION

	Anger	Disgu st	Fear	Happi ness	Sadne ss	Surpr ise
Anger		0.827	0.500	0.002	0.384	0.465
Disgu st	0.827		0.427	-0.105	0.370	0.403
Fear	0.500	0.427		0.205	0.530	0.549
Happi ness	0.002	-0.105	0.205		0.196	0.558
Sadne ss	0.384	0.370	0.530	0.196		0.425
Surpr ise	0.465	0.403	0.549	0.558	0.425	

Επιπλέον στο TABLE IV παρατηρούμε ότι τα ζεύγη θυμός-αηδία και έκπληξη χαρά έχουν ίδιες τιμές για τις ίδιες κατηγορίες, καθώς δεν μπορούν να διαχωρισθούν.

Συνεπώς οι προσέγγιση που χρησιμοποιήθηκε είναι ικανή να προβλέψει τα συναισθήματα των χρηστών ,ωστόσο τα αποτελέσματα είναι πιο έγκυρα όσον αφορά τα συναισθήματα της χαράς και του θυμού.

3.7 Η συγκεκριμένη εργασία [7] λαμβάνει για τις ανάγκες των πειραμάτων τις δεδομένα από ελληνικές αναρτήσεις του Twitter, του facebook , του Instagram και άλλων ελληνικών κοινωνικών δικτύων. Στα πλαίσια της ανάλυσης συναισθήματος που πραγματοποιείται λαμβάνονται υπόψη και τα emojis, τα οποία εκφράζονται ως διάνυσμα που αναπαριστά τη συχνότητα εμφάνισης του καθενός.

Τα αποτελέσματα των πειραμάτων που αφορούν τεχνικές βαθιάς μάθησης είναι τα εξής:

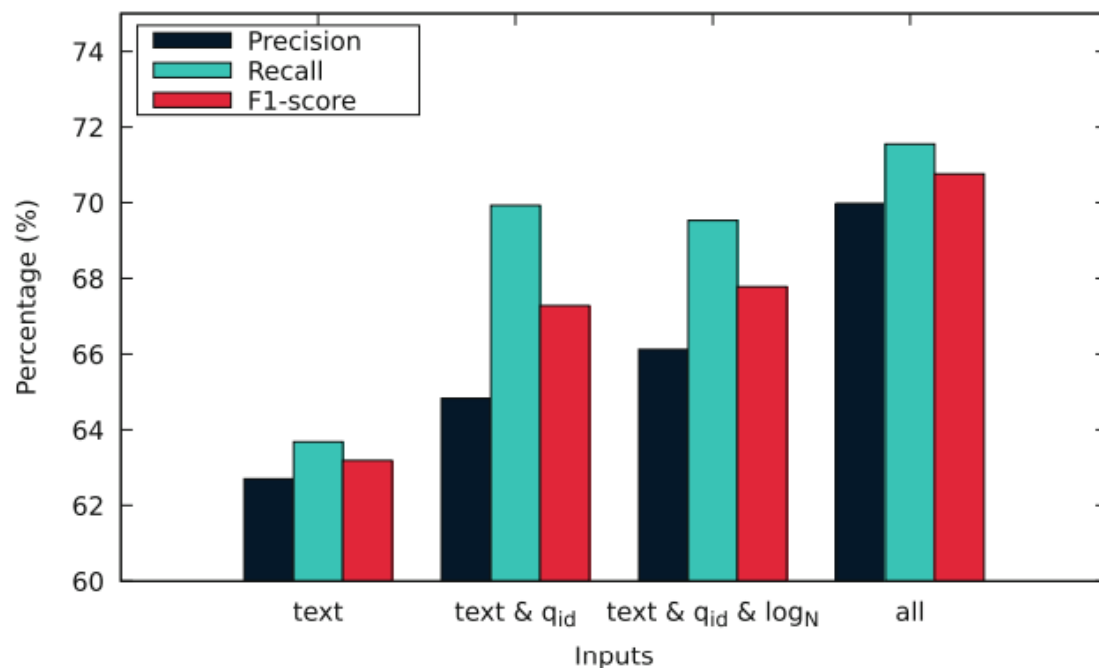


Fig. 4. System performance w.r.t different inputs

Στο Fig.4. συνοψίζεται η απόδοση του συστήματος σε σχέση με τις διαφορετικές εισόδους. Όταν λαμβάνεται υπόψη μόνο το κείμενο της κάθε εγγραφής, η αποτελεσματικότητα της προτεινόμενης προσέγγισης είναι περιορισμένη. Όταν λαμβάνονται υπόψη πληροφορίες σχετικά με τις πτυχές, η ανάκληση αυξάνεται πάνω από 6%. Η προσθήκη του λογάριθμου του αριθμού των επαναλήψεων επηρεάζει οριακά την ανάκληση, αλλά ενισχύει περαιτέρω την ακρίβεια, βελτιώνοντας τη βαθμολογία F1. Τέλος όταν παρέχονται και οι τέσσερις είσοδοι (κείμενο, όψη, λογάριθμος του αριθμού των επαναλήψεων, emoji), η απόδοση του συστήματος βελτιώνεται περαιτέρω, με τις τρεις μετρήσεις να είναι πάνω (70%), προσθέτοντας περισσότερο από 8% στη συνολική απόδοση του συστήματος. Το σύστημα ανταπεξέρχεται

αρκετά καλά σε αρνητικές τάξεις ενώ τα αποτελέσματα των θετικών κλάσεων είναι κάτω του μέσου όρου ,όπως φαίνεται και στο Fig.5. το οποίο δείχνει την απόδοση ανά τάξη των εξεταζόμενων μετρήσεων.

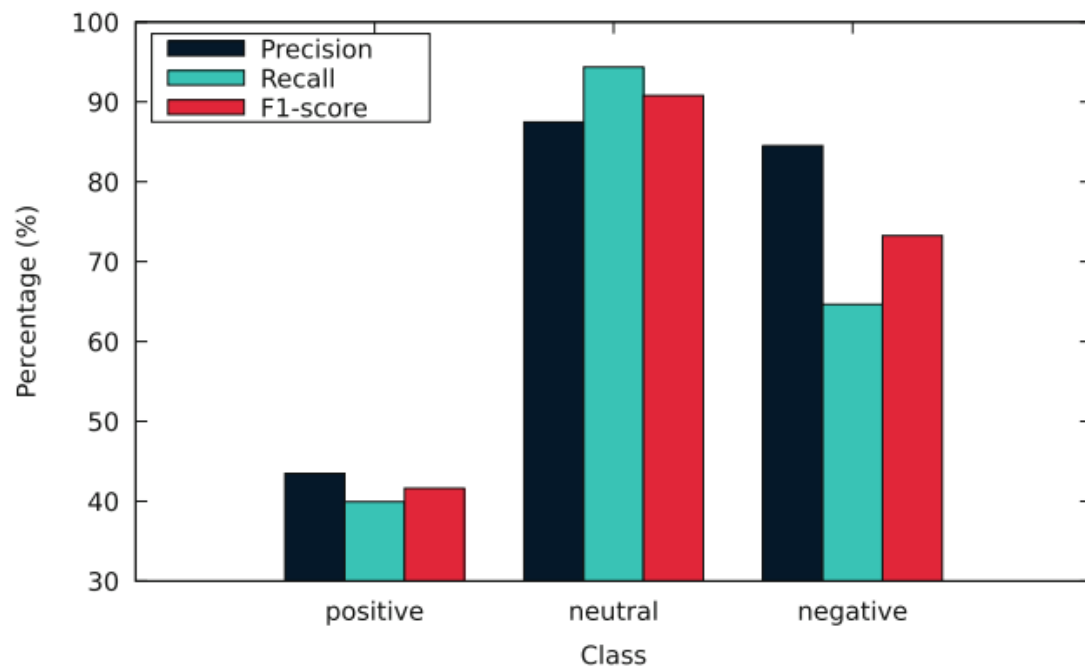


Fig. 5. System performance w.r.t different classes

3.8 Η παρούσα εργασία στοχεύει στο διαχωρισμό των στάνταρ σύγχρονων ελληνικών (SMG) και της ελληνικής κυπριακής διαλέκτου (CG). Έγινε εξαγωγή N-gram χαρακτηριστικών και εφαρμογή τριών αλγορίθμων ταξινόμησης οι οποίοι δοκιμάστηκαν σε επισημασμένες προτάσεις: multinomial naive Bayes (NB), (SVC) και logistic regression. Το πείραμα που πραγματοποιήθηκε έδειξε ότι την καλύτερη απόδοση είχε ο multinomial NB με ακρίβεια ίση με 0.95 ,όπως φαίνεται στους παρακάτω πίνακες.

		Accuracy (2 d.p.)		
		Multinomial NB	Linear SVC	Logistic Regression
Iteration	1	0.96	0.93	0.94
	2	0.95	0.93	0.94
	3	0.96	0.90	0.92
	4	0.93	0.94	0.95
	5	0.96	0.91	0.93
Mean		0.95	0.92	0.94

Table I: Mean accuracy of the classifiers over five trials

	Precision	Recall	F_1 score	Support
CG	0.97	0.95	0.96	108
SMG	0.95	0.97	0.96	100
Mean Total	0.96	0.96	0.96	208

Table II: The multinomial NB classifier's performance metrics. The final row is a support-weighted mean of precision, recall and F_1 score.

Ακολουθούν τρία πειράματα ώστε να εξεταστεί αν με λιγότερα χαρακτηριστικά θα υπάρξει διαφορά στην επίδοση. Στο πρώτο πείραμα γίνεται αφαίρεση της επένδυσης χαρακτήρων n-gram. Τα αποτελέσματα έδειξαν ότι η επένδυση σε αυτούς τους χαρακτήρες κατέχει πολύ σημαντικό ρόλο στη διαδικασία της ταξινόμησης ,καθώς η τιμή της ακρίβειας έπεσε από το 0.96 στο 0.93.Στο επόμενο πείραμα διερευνήθηκε η χρήση n-gram μόνο σε χαρακτήρες και όχι σε λέξεις. Τα αποτελέσματα δείχνουν ότι η χρήση n-gram, σε λέξεις συμβάλει στην καλύτερη ακρίβεια, καθώς η τιμή της ακρίβειας έπεσε από το 0.96 στο 0.94.Στο τελευταίο πείραμα αγνοήθηκαν χαρακτηριστικά που κατέχουν λιγότερο από το 1% των δεδομένων. Εν τέλη αυτό το πείραμα έδειξε ότι η αφαίρεση λέξεων που δεν εμφανίζονται συχνά επηρεάζει αρνητικά την ακρίβεια και αυτό μπορεί να συμβαίνει λόγω του μικρού μεγέθους

του συνόλου δεδομένων που χρησιμοποιήθηκε, καθώς η τιμή της ακρίβειας έπεσε από το 0.96 στο 0.92.

3.9 Argument Extraction from News, Blogs, and the Social Web

Η συγκεκριμένη εργασία[9] μιλά για την εξαγωγή συμπερασμάτων από προτάσεις και τμήματα προτάσεων αναρτήσεων του ελληνικού κοινωνικού ιστού , ενώ επικεντρώνεται στη μελέτη των πιο σύγχρονων χαρακτηριστικών στον τομέα των κοινωνικών δικτύων.

Πραγματοποιούνται τα εξής πειράματα : Ως πρώτο βήμα εφαρμόστηκε η μεθοδολογία 10-fold cross validation στο σύνολο των 16000 προτάσεων ενώ το σύνολο δεδομένων χωρίστηκε σε δύο τμήματα. Τα αποτελέσματα της αξιολόγησης των ταξινομητών φαίνονται στους πίνακες 2-4.

Table 2. Results of various classifiers for the first step, evaluated with 10-fold cross validation (both classes).

Step A: State of the art + new features				
	Precision	Recall	F1-Measure	Accuracy
Naive Bayes	74.10%	74.00%	74.05%	73.99%
Random Forest	74.60%	74.40%	74.50%	74.38%
Logistic Regression	77.10%	77.10%	77.10%	77.12%
Support Vector Machines	76.00%	76.00%	76.00%	76.01%
Step A: State of the art features				
	Precision	Recall	F1-Measure	Accuracy
Naive Bayes	67.40%	65.40%	66.38%	65.44%
Random Forest	64.50%	64.50%	64.50%	64.47%
Logistic Regression	68.30%	68.30%	68.30%	68.25%
Support Vector Machines	68.40%	68.10%	68.25%	68.12%

Table 3. Results of various classifiers for the first step, evaluated with 10-fold cross validation (only positive class).

Step A: State of the art + new features			
	Precision	Recall	F1-Measure
Naive Bayes	72.50%	76.10%	74.26%
Random Forest	72.70%	72.50%	72.60%
Logistic Regression	76.80%	76.90%	76.85%
Support Vector Machines	74.70%	77.70%	76.17%
Step A: State of the art features			
	Precision	Recall	F1-Measure
Naive Bayes	61.30%	81.40%	69.93%
Random Forest	64.50%	62.40%	63.43%
Logistic Regression	68.60%	65.70%	67.12%
Support Vector Machines	70.30%	61.30%	65.50%

Table 4. Results of various classifiers for the first step, evaluated with 10-fold cross validation (only negative class).

Step A: State of the art + new features			
	Precision	Recall	F1-Measure
Naive Bayes	75.50%	71.90%	73.66%
Random Forest	73.30%	73.50%	73.40%
Logistic Regression	77.40%	77.30%	77.35%
Support Vector Machines	77.40%	74.40%	75.87%
Step A: State of the art features			
	Precision	Recall	F1-Measure
Naive Bayes	73.30%	49.90%	59.38%
Random Forest	64.50%	66.50%	65.50%
Logistic Regression	67.90%	70.80%	69.32%
Support Vector Machines	66.50%	74.80%	70.41%

Στη συνέχεια πραγματοποιείται χρήση της τεχνικής (SMOTE) .

Table 5. Results of various classifiers for the first step, evaluated with 10-fold cross validation (both classes), using SMOTE over-sampling.

Step A: State of the art + new features				
	Precision	Recall	F1-Measure	Accuracy
Naive Bayes	75.50%	74.70%	75.10%	74.72%
Logistic Regression	77.50%	77.50%	77.50%	77.49%
Support Vector Machines	77.70%	77.60%	77.65%	77.64%

Στο δεύτερο βήμα χρησιμοποιούνται τυχαία πεδία για τον εντοπισμό υπόθεσης σε πρόταση με χρήση της αναπαράστασης BIO. Τα αποτελέσματα και για τα δύο βήματα είναι πάνω από το βασικό σενάριο. Η απόδοση αυξήθηκε από 20% σε 77% μέσω του ταξινομητή logistic regression.

Τέλος παρουσιάζονται στον πίνακα 8 τα αποτελέσματα της αξιολόγησης προτάσεων και στον πίνακα 9 τα αποτελέσματα της αξιολόγησης τμημάτων προτάσεων με τη βοήθεια του επισημασμένου σώματος κειμένων NOMAD και με τη χρήση 10-fold cross validation.

Table 8. Evaluation results for the Sentence-Level Argument Matcher, evaluated with 10-fold cross validation on the NOMAD manually annotated corpus.

Average	Precision	Recall	F1-Measure
micro	81.03% (± 9.94)	81.03% (± 9.94)	81.03% (± 9.94)
macro	78.58% (± 10.72)	78.32% (± 10.90)	78.45% (± 10.80)

Table 9. Evaluation results for the Segment-level Argument Matcher, evaluated with 10-fold cross validation on the NOMAD manually annotated corpus.

Average	Precision	Recall	F1-Measure
Micro	95.02% (± 2.42)	95.02% (± 2.42)	95.02% (± 2.42)
Macro	91.46% (± 4.12)	91.74% (± 4.16)	91.60% (± 4.14)

3.10 Η συγκεκριμένη εργασία[10] συγκρίνει τις ημι-επιβλεπόμενες και τις επιβλεπόμενες τεχνικές εντοπισμού ειρωνείας σε ελληνικά tweets πολιτικού περιεχομένου. Έγινε χρήση των εξής αλγορίθμων: J48, SVM, νευρωνικά δίκτυα, Naive Bayes, Functional Trees, KStar και Random Forests. Την καλύτερη απόδοση είχαν τα Functional Trees (Precision 82.4) στην περίπτωση των επιβλεπόμενων τεχνικών.

Table 3

Performance measures of the training algorithms. Green indicates the best performance, while red the worst.

Algorithms	Supervised			Semi supervised		
	Precision	Recall	F-measure	Precision	Recall	F-measure
J48	78.4	77.8	77.0	82.7	73.8	73.3
Naive bayes	72.8	71.4	71.6	69.5	59.5	57.8
Functional trees	82.4	80.2	79.1	81.2	73.0	72.6
K-Star	76.9	77.0	76.7	68.6	61.1	60.2
Random forest	77.4	77.0	76.3	83.1	74.6	74.2
SVM	73.6	73.8	73.5	71.7	65.9	65.6
Neural network	79.9	78.6	77.6	63.8	55.6	53.8

Table 4

Irony tweets that received every party and their election results. The fluctuation describes the difference between the May 2012 election results and the previous.

Party	Irony		Total	Irony/total %		2012 Greek election results %	Fluctuation % of 2012 results from 2009 greek election results
	Supervised	Semi-supervised		Supervised	Semi-supervised		
ANTARSYA	511	175	1240	41	14	1.19	0.83
PASOK	1440	592	2240	64	26	13.18	- 30.74
DHMAR	497	224	1223	41	18	6.11	New
KKE	504	218	1262	40	17	8.48	0.94
ND	252	104	560	45	18	18.80	- 14.60
SYRIZA	2649	1054	4371	61	24	16.70	12.10
XA	1040	410	1821	57	22	6.90	6.68
ANEL	280	77	680	41	11	10.60	New

Στην περίπτωση της ημι-επιβλεπόμενης τεχνικής έγινε χρήση του αλγόριθμου collective-tree. Για την εκπαίδευση χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι με αυτούς της επιβλεπόμενης τεχνικής και την καλύτερη απόδοση είχε το Random Forest (precision 83.1). Ενώ και στις δύο περιπτώσεις το Naive Bayes είχε τη χειρότερη απόδοση.

4.Συμπεράσματα

Στο κεφάλαιο αυτό μπορεί να πραγματοποιηθεί εξαγωγή συμπερασμάτων που αφορούν τη διερεύνηση τεχνικών εξόρυξης γνώσης από ελληνικά σώματα κειμένων του κοινωνικού ιστού ,μέσα από την ανάλυση των εργασιών που χρησιμοποιήθηκαν παραπάνω ,καθώς η παρούσα εργασία αποτελεί βιβλιογραφική έρευνα.

Η εξόρυξη γνώσης και η ανάλυση συναισθήματος των ανθρώπων μέσα από τον κοινωνικό ιστό αποτελούν βασικά εργαλεία για τη διαδικασία λήψης αποφάσεων.Τα δεδομένα που λαμβάνονται μέσω μηχανισμών μηχανικής μάθησης είναι αρκετά ακριβή , αντικατοπτρίζοντας μια ρεαλιστική εικόνα των απόψεων και των προτιμήσεων των χρηστών. Η χρηστικότητα τους είναι αδιαμφισβήτητη , καθώς παρέχεται η δυνατότητα παροχής πληροφοριών τόσο σε επιχειρήσεις , όσο και στον καταναλωτή-χρήστη παρέχοντας όσο το δυνατόν πιο έγκυρες και εξατομικευμένες υπηρεσίες και πληροφορίες.

Η διαδικασία εξόρυξης γνώσης δεν παρουσιάζει ιδιαίτερες διαφορές μεταξύ των εργασιών.Τα δεδομένα που συλλέγονται από τα σώματα κειμένων υφίστανται μία προεπεξεργασία ώστε να γίνονται κατονοητά από τους υπολογιστικούς κώδικες , αλλά και για να οδηγήσουν με μεγαλύτερη ταχύτητα και ευκολία στο εκάστοτε ζητούμενο.Πολλές φορές κρίνεται απαραίτητη η δημιουργία ενός νέου λεξικού,ώστε να επιτευχθούν τα επιθυμητά αποτελέσματα.Στη συνέχεια εφαρμόζονται διαφορετικές τεχνικές και εκπαιδεύονται διαφορετικά μοντέλα μηχανικής μάθησης , όπως TF / IDF, unigrams, bigrams, POS . Επιπλέον χρησιμοποιούνται αλγόριθμοι ταξινόμησης όπως Support Vector Machines(SVMs), Naive Bayes, Logistic Regression και Random Forest.Τέλος τα δεδομένα μπορούν πλέον να ερμηνευτούν , να αξιολογηθούν και να χρησιμοποιηθούν για την εύρεση της βέλτιστης λύσης σε ζητήματα λήψης αποφάσεων .Το επιθυμητό αποτέλεσμα μπορεί να προέρχεται από τη χρήση διαφορετικής τεχνικής σε κάθε εργασία , συνεπώς γίνεται σε κάθε περίπτωση εφαρμογή ποικίλων τεχνικών ώστε επιτευχθεί η μέγιστη δυνατή ακρίβεια.

Οι εργασίες που αφορούν την εξόρυξη γνώσης από τον ελληνικό κοινωνικό ιστό είναι ελάχιστες συγκριτικά με αυτές που εξετάζουν δεδομένα σε διαφορετικές γλώσσες. Συνεπώς, πολύ συχνά απαιτείται μεγαλύτερη προεπεξεργασία των κειμένων, καθώς και ανάπτυξη νέων λεξικών και χρήση μη αυτόματων τεχνικών ώστε να εκπαιδευτούν τα δεδομένα κατάλληλα, ώστε να υπάρξουν λύσεις, οι οποίες να ανταποκρίνονται στην πραγματικότητα.

Βασική ιδιαιτερότητα των εργασιών που χρησιμοποιήθηκαν είναι το ότιπραματεύονται γεγονότα και δεδομένα αρκετά εξειδικευμένα, καθώς ασχολούνται με μία συγκεκριμένη χώρα-γλώσσα. Μια γλώσσα η οποία δε χρησιμοποιείται ιδιαίτερα παγκοσμίως ενώ τα γεγονότα που εξετάζονται περιορίζονται ως επι το πλείστον στο ενδιαφέρον των χρηστών που ζουν στη συγκεκριμένη χώρα. Συμπεραίνουμε λοιπόν ότι οι δυσκολίες και οι προκλήσεις είναι μεγάλες, καθώς απαιτείται μεγαλύτερη προσπάθεια και περισσότερες ενέργειες, καθώς και διάθεση περισσότερου χρόνου για την επίτευξη των στόχων και των απαιτήσεων μίας διεργασίας εξόρυξης γνώσης από σώματα κειμένων του κοινωνικού που αφορούν μία γλώσσα όχι τόσο διαδεδομένη, όπως τα ελληνικά.

Η μηχανική μάθηση και ιδιαίτερα η εξόρυξη γνώσης και συναισθήματος αποτελούν ένα σημαντικό εργαλείο, το οποίο είναι ικανό να συμβάλει στην επίλυση προβλημάτων με τη χρήση πραγματικών απόψεων και δεδομένων. Χάρη στην ευρεία χρήση των μέσων κοινωνικής δικτύωσης, είναι πλέον αρκετά εύκολο να εντοπίσει κανείς τις απόψεις, τις πεποιθήσεις, καθώς και την οικονομική και την ψυχολογική κατάσταση του κάθε χρήστη. Συνεπώς, μελλοντικά θα μπορούσαν να πραγματοποιηθούν εργασίες που αφορούν πιο σύγχρονα κομμάτια του κοινωνικού ιστού. Παραδείγματος χάρη θα μπορούσαν να διερευνηθούν κείμενα του ελληνικού κοινωνικού ιστού, που αφορούν σχολιασμούς ή επισημάνσεις στο Instagram, το οποίο αποτελεί ένα από τα πιο δημοφιλή μέσα κοινωνικής δικτύωσης. Επιπλέον θα μπορούσαν να διερευνηθούν καταστάσεις που αφορούν τη σύγχρονη κοινωνία, που ίσως συνέβαλαν σε ζητήματα όπως η υγεία, η οικονομία, η ανεργία και η επαίδευση, συμβάλλοντας έτσι στη

βελτίωση του βιοτικού επιπέδου των ανθρώπων και την ενίσχυση της ευρύτερης οικονομίας.

Βιβλιογραφικές Αναφορές

- [1] Petasis, G., Spiliotopoulos, D., Tsirakis, N., & Tsantilas, P. (2014). Sentiment Analysis for Reputation Management: Mining the Greek Web. *Lecture Notes in Computer Science*, 327–340.
- [2] Pitenis, Zeses; Zampieri, Marcos; Ranasinghe, Tharindu (2020) Offensive Language Identification in Greek. *European Language Resources Association*, 5113—5119.
- [3] Markopoulos, G., Mikros, G., Iliadi, A., & Lontos, M. (2015). Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features. *Springer Proceedings in Business and Economics*, 373–383.
- [4] George K. Mikros.(2012). Authorship Attribution and Gender Identification in Greek Blogs. *Selected papers of the VIIIth International Conference on Quantitative Linguistics(QUALICO) in Belgrade.Serbia.Belgrade:Academic Mind .21-32*
- [5] Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the Greek Referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM.367–376*.
- [6] Dimitrios Mallis, Georgios Kalamatianos, Dimitrios Nikolaras, Symeon Symeonidis.(2015). Sentiment Analysis of Greek Tweets and Hashtags using Sentiment Lexicon. *ACM International Conference Proceeding Series / [ed] Karanikolas Nikitas N., New York, NY, United States: Association for Computing Machinery (ACM), 2015, p. 63-68*
- [7]Maglogiannis, I., Iliadis, L., & Pimenidis, E. (Eds.). (2020). A Deep Learning Approach to Aspect-Based Sentiment Prediction .*Artificial Intelligence Applications and Innovations. IFIP Advances in Information and Communication Technology*. 397-408
- [8] Sababa, H., & Stassopoulou, A. (2018). A Classifier to Distinguish Between Cypriot Greek and Standard Modern Greek. 2018 Fifth

International Conference on Social Networks Analysis, Management and Security (SNAMS).

[9] Goudas, T., Louizos, C., Petasis, G., & Karkaletsis, V. (2015). *Argument Extraction from News, Blogs, and the Social Web. International Journal on Artificial Intelligence Tools*, 24(05),

[10] Charalampakis B., Spathis D., Kouslis E., Kermanidis K. L. 2016. *A Comparison between Semi-supervised and Supervised Text Mining Techniques on Detecting Irony in Greek Political Tweets. Engineering Applications of Artificial Intelligence*, Vol. 51, pp. 50-57. Elsevier.

[11] Kermanidis K., Maragoudakis M. 2013. *Political Sentiment Analysis of Tweets before and after the Greek Elections of May 2012. International Journal on Social Network Mining*, Vol 1, No 3/4, pp. 298-317. Inderscience

[α] <https://www.igi-global.com/>

[β] [Amanpreet Singh](#) ; [Narina Thakur](#) ; [Aakanksha Sharma](#) (2016). **A review of supervised machine learning algorithms.**

[γ] [Ethem Alpaydin](#). (2020). **Introduction to Machine Learning**

[δ] <https://machinelearningmastery.com/>

Ravi, K., & Ravi, V. (2015). *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems*, 89, 14–46.

[ε] Deng, L. (2014). *Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing*, 7(3-4), 197–387.