ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι. ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση Παραπλανητικών Ειδήσεων με Χρήση Μηχανικής Μάθησης

Του φοιτητή

Παναγιώτη Καράδαη

Αρ. Μητρώου: 113732

Επιβλέπων καθηγητής

Κωνσταντίνος Διαμαντάρας

Θεσσαλονίκη 2018

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία αποσκοπεί στην παρουσίαση μιας λύσης του προβλήματος των παραπλανητικών ειδήσεων (fake news) χρησιμοποιώντας μεθόδους μηχανικής μάθησης. Παρουσιάζεται το πρόβλημα των παραπλανητικών ειδήσεων και τρόποι αντιμετωπισής του με συμβατικούς τρόπους αλλά και μεθόδους μηχανικής μάθησης. Ακόμη, παρουσιάζονται οι τρόποι αναπαράστασης κειμένου σε χώρους διανυσμάτων με word2vec και GloVe. Τελικά παρουσιάζονται τα πειράματα που εκτελέστηκαν χρησιμοποιώντας SVM για την ταξινόμηση πάνω σε δύο σετ δεδομένων και λύσεις για το παραπάνω πρόβλημα.

Λέξεις κλειδιά: Fake news, Machine Learning, Text Analysis, SVM, word2vec, GloVe

ABSTRACT

This thesis aims at presenting a solution to the problem of "Fake news" using machine learning. The problem of "Fake news" and ways of dealing with it both with conventional ways and methods of mechanical learning is presented. In addition to that, ways of representing text in vector spaces with word2vec and GloVe are presented. Finally, the experiments performed using SVM for the classification on two sets of data and solutions for the above problem are presented.

Keywords: Fake news, Machine Learning, Text Analysis, SVM, word2vec, GloVe

ΕΥΧΑΡΙΣΤΙΕΣ (προαιρετικά)

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Κωνσταντίνο Διαμαντάρα για την καθοδήηση που μου πρόσφερε στην εκπόνηση της πτυχιακής μου εργασίας, όπως επίσης και για την πολύτιμη βοήθεια του για την επίλυση διαφόρων θεμάτων.

Επίσης, θα ήθελα να ευχαριστήσω τον κ. Γεώργιο Γραβάνη για την συνεργασία και την καθοδήγηση που μου προσέφερε κατά την διάρκεια της εκπόνησης της πτυχιακής.

Τελικά θα ήθελα να ευχαριστήσω την οικογένεια μου τους φίλους και συναδέλφους για την υποστήριξη και την βοήθεια όλα αυτά τα χρόνια.

Κατάλογος περιεχομένων

Ευρετήριο πινάκων

Ευρετήριο σχημάτων

ΕΙΣΑΓΩΓΗ

Καθώς η χρήση μέσων κοινωνικής δικτύωσης αυξάνεται, ολοένα και περισσότεροι άνθρωποι τείνουν να αναζητούν και να καταναλώνουν ειδήσεις από τα κοινωνικά μέσα παρά από παραδοσιακούς οργανισμούς ειδήσεων. Οι λόγοι για αυτή την αλλαγή στις συμπεριφορές κατανάλωσης βασίζονται στη φύση των πλατφορμών κοινωνικών μέσων, συχνά είναι πιο γρήγορο το να διαβάσει και να μοιραστεί κάποιος σε σύγκριση με τα παραδοσιακά μέσα ενημέρωσης, όπως οι εφημερίδες και η τηλεόραση. Επίσης είναι πιο εύκολο να γίνει συζήτηση ή και σχολιασμός. Για παράδειγμα, σύμφωνα με το Reuters περίπου τα δύο τρίτα των Αμερικανών ενηλίκων διαβάζουν τουλάχιστον μερικές από τις ειδήσεις τους μέσω κοινωνικών μέσων μαζικής ενημέρωσης, με δύο στους δέκα να το κάνουν συχνά.

Παρά τα πλεονεκτήματα που προσφέρουν τα κοινωνικά μέσα ενημέρωσης, η ποιότητα των ειδήσεων στα κοινωνικά μέσα είναι χαμηλότερη από τους παραδοσιακούς οργανισμούς ειδήσεων. Ωστόσο, επειδή η διάδοση μέσω κοινωνικών μέσων είναι φθηνότερη, πολύ γρηγορότερη και ευκολότερη, παράγεται μεγάλος όγκος “Fake news” (ψευδείς ειδήσεις), για παράδειγμα άρθρα με σκοπό την παραπλάνηση, οικονομικό ή και πολιτικό κέρδος.

Οι στόχοι της παρούσας πτυχιακής εργασίας συνοπτικά είναι οι παρακάτω:

- Παρουσίαση του προβλήματος των παραπλανητικών ειδήσεων και υπαρχόντων λύσεων.
- Παρουσίαση τρόπων αναπαράστασης κειμένου για χρησιμοποίησή τους με μεθόδους μηχανικής μάθησης..
- Και τελικά η πρόταση ενός τρόπου πρόβλεψης παραπλανητικών ειδήσεων χρησιμοποιώντας μεθόδους μηχανικής μάθησης.

Το υπόλοιπο της εργασίας αποτελείται από τα πέντε παρακάτω κεφάλαια, στο πρώτο κεφάλαιο αναλύεται το πρόβλημα των παραπλανητικών ειδήσεων πως αντιμετωπίζεται τώρα και σχετικές δουλειές άλλων ερευνητών.

Στο δεύτερο κεφάλαιο παρουσιάζονται μέθοδοι αναπαράστασης κειμένου και αναλύονται τα word2vec και GloVe.

Στο τρίτο κεφάλαιο παρουσιάζονται οι μέθοδοι μηχανικής μάθησης και τα εργαλεία που χρησιμοποιούνται.

Στο τέταρτο κεφάλαιο κεφάλαιο παρουσιάζονται τα πειράματα που εκτελέστηκαν και αναλύονται τα αποτελέσματα.

Στο πέμπτο κεφάλαιο παρουσιάζονται συμπεράσματα, προτάσεις βελτίωσης και πιθανές μελλοντικές επεκτάσεις.

ΚΕΦΑΛΑΙΟ 1: Παραπλανητικές ειδήσεις

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο θα ορισθεί και θα αναλυθεί το πρόβλημα των παραπλανητικών ειδήσεων. Αρχικά θα δοθεί ένας ορισμός και μερικά παραδείγματα “Fake news”. Στην συνέχεια παρουσιάζονται τρόποι επίλυσης του προβλήματος από πλατφόρμες και τελικά τρόποι που έχουν προταθεί χρησιμοποιώντας μεθόδους μηχανικής μάθησης.

1.1: Το πρόβλημα των παραπλανητικών ειδήσεων (fake news)

Η έννοια των "Fake news" λαμβάνει όλο και περισσότερη προσοχή, κυρίως λόγω του Διαδικτύου και της χρήσης κοινωνικών μέσων. Με τον όρο “Fake news” αναφερόμαστε σε ένα άρθρο ειδήσεων που είναι εσκεμμένα και επαληθεύσιμα ψευδές. Το 2016 με το τέλος των εκλογών της Αμερικής εκτιμήθηκε ότι υπήρξαν πάνω από ένα εκατομμύριο ψευδείς αναρτήσεις στο twitter σχετικά με το “Pizzagate”, μία θεωρία συνομωσίας. Επίσης, την χρονιά εκείνη η λέξη “Fake news” ονομάστηκε λέξη της χρονιάς από το λεξικό Macquarie.

Η εκτεταμένη διάδοση των “Fake news” μπορεί να έχει σοβαρές αρνητικές επιπτώσεις στους ανθρώπους και την κοινωνία. Πρώτον, ψεύτικα νέα μπορούν να χαλάσουν την ισορροπία αυθεντικότητας του οικοσυστήματος ειδήσεων. Για παράδειγμα, οι πιο δημοφιλείς ψεύτικες ειδήσεις ήταν πιο ευρέως διαδεδομένες στο Facebook από τις πιο δημοφιλής αυθεντικές επικρατούσες ειδήσεις κατά τη διάρκεια ψηφίσματος του προέδρου των ΗΠΑ του 2016. Δεύτερον, τα ψεύτικα νέα πείθουν αναγνώστες να δέχονται προκατειλημμένες ή ψευδείς πεποιθήσεις. Αυτά συνήθως δημιουργούνται από προπαγανδιστές για να διαδώσουν πολιτικά μηνύματα. Μία έρευνα δείχνει ότι η Ρωσία δημιούργησε ψεύτικους λογαριασμούς

και κοινωνικά bots για να εξαπλώσει ψευδείς ιστορίες. Τρίτον, οι ψεύτικες ειδήσεις αλλάζουν τον τρόπο που ερμηνεύουν και ανταποκρίνονται οι άνθρωποι σε αληθινά νέα. Για παράδειγμα, σε μερικές περιπτώσεις δημιουργούνται ψεύτικες ειδήσεις με στόχο να προκαλέσουν τη δυσπιστία των ανθρώπων και να τους μπερδέψουν, εμποδίζοντας τις ικανότητές τους να διαφοροποιήσουν την αλήθεια από τα ψέματα.

1.2: Λύση του προβλήματος από πλατφόρμες Fact checking

Το πρόβλημα των ψευδών ειδήσεων που αναφέρθηκε είναι αρκετά πολύπλοκο πρόβλημα και λαμβάνει όλο και περισσότερη προσοχή κυρίως λόγω της αύξησης της χρήσης των κοινωνικών μέσων. Ο κύριος λόγος της πολυπλοκότητας του οφείλεται στο γεγονός ότι όπως αναφέρθηκε παραπάνω υπάρχουν πολλές μορφές ψευδών ειδήσεων και συνήθως έχουν διαφορετικά κίνητρα. Για να καταπολεμηθεί το φαινόμενο των ψευδών ειδήσεων έχουν δημιουργηθεί διάφορες πλατφόρμες οι οποίες κάνουν “Fact checking”, δηλαδή ελέγχουν δημοσιευμένα νέα για να σιγουρέψουν την αυθεντικότητα τους. Μερικές από αυτές είναι οι Politifact, Factcheck, TruthOrFiction. Ωστόσο, αυτές οι πλατφόρμες χρειάζονται πολύ χρόνο για τον έλεγχο των νέων διότι τα νέα ελέγχονται από ανθρώπους, έτσι ο αριθμός που μπορούνε να ελέγξουν είναι περιορισμένος ενώ ο όγκος των ψευδών νέων που παράγεται μεγάλος. Αυτό το πρόβλημα λύνεται μερικώς από εργαλεία που έχουν αναπτυχθεί για να βοηθάνε τα άτομα που κάνουν αυτή την δουλειά όπως το hoaxy, το check κ.α.

1.3: Λύση του προβλήματος χρησιμοποιώντας μεθόδους μηχανικής μάθησης

Παρόλο που υπάρχουν ήδη κάποιες πλατφόρμες για τον έλεγχο των νέων είναι σημαντικό να αναπτυχθούν εργαλεία που μπορούν να κάνουν αυτό τον έλεγχο τελείως αυτοματοποιημένα. Για να επιτευχθεί αυτό τα τελευταία χρόνια έχουν αναπτυχθεί μέθοδοι μηχανικής μάθησης για την ανάλυση του περιεχομένου των άρθρων και στην συνέχεια την κατηγοριοποίηση τους ανάμεσα σε ψευδή και μη. Μερικές από τις μεθόδους που έχουν αναπτυχθεί από άλλους ερευνητές παρουσιάζονται παρακάτω.

Η χρήση γλωσσικών σημαδιών σε συνδυασμό με μεθόδους μηχανικής μάθησης χρησιμοποιήθηκαν από τους Benjamin D. Horne και Sibel Adali. Χρησιμοποιώντας Support Vector Machines για την ταξινόμηση των νέων κατάφεραν να φτάσουν ακρίβεια 78% χωρίζοντας ψευδή από αληθινά νέα. Ο Hadeer Ahmed χρησιμοποίησε n-grams και TF-IDF για την εξαγωγή χαρακτηριστικών σε συνδυασμό με τεχνικές μηχανικής μάθησης για την αναγνώριση ψευδών ειδήσεων. Στο άρθρο του χρησιμοποίησε μία σειρά από διαφόρους ταξινομητές με τον καλύτερο να είναι ο Linear SVM όπου επιτεύχθηκε ακρίβεια 92%. Οι Yang Yang κ.α προτείνουν ένα μοντέλο το οποίο χρησιμοποιεί συνελικτικά νευρωνικά δίκτυα (CNNS), συνδυάζοντας διαφορετικά δίκτυα για εικόνες και κείμενο πέτυχαν επίσης 92% ακρίβεια.

ΚΕΦΑΛΑΙΟ 2: Μέθοδοι αναπαράστασης κειμένου (Word representations)

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο παρουσιάζονται οι μέθοδοι αναπαράστασης κειμένου. Ο όρος αναπαράσταση κειμένου περιγράφει μια σειρά από τεχνικές που μετατρέπουν το αρχικό κείμενο σε μία αναπαράσταση διανυσμάτων την οποία μπορούμε να διαχειριστούμε αποτελεσματικά με μεθόδους μηχανικής μάθησης.

Θα επικεντρωθούμε σε ενσωματώσεις λέξεων (word embeddings) καθώς τα πειράματα που ακολουθούν βασίζονται πάνω τους. Αρχικά θα γίνει μία ιστορική αναδρομή στις μεθόδους αναπαράστασης κειμένου και στην συνέχεια θα αναλυθούν με περισσότερη λεπτομέρεια οι μέθοδοι word2vec και GloVe οι οποίοι χρησιμοποιούνται αργότερα για την εξαγωγή των γνωρισμάτων από τα σύνολα δεδομένων.

2.1: Word embeddings

Τα word embeddings είναι πυκνές αναπαραστάσεις λέξεων σε χώρο διανυσμάτων χαμηλών διαστάσεων. Χρησιμοποιούνται ως μοντέλα νευρωνικής γλώσσας (neural language models) σε πολλές εργασίες επεξεργασίας φυσικής γλώσσας (Natural Language Processing) και είναι χρήσιμα γιατί μπορούν να χρησιμοποιηθούν κατευθείαν σαν είσοδος σε νευρωνικά δίκτυα. Παραδοσιακά κατά την επεξεργασία φυσικής γλώσσας οι λέξεις αναπαριστώνται ως διακριτοί αριθμοί για παράδειγμα η λέξη γάτα μπορεί να αποθηκευτεί ως ‘123’ και ο σκύλος σαν ‘444’. Αυτοί οι αριθμοί ωστόσο, δεν έχουν καμία συντακτική ή σημασιολογική πληροφορία. Έτσι τα μοντέλα αυτά δεν μπορούν να χρησιμοποιήσουν πληροφορίες που έχουν για άλλες λέξεις όταν επεξεργάζονται κάποια λέξη. Αυτό οδηγεί στην σποραδικότητα των δεδομένων και συνήθως σημαίνει ότι ίσως χρειαστούμε περισσότερα δεδομένα προκειμένου να εκπαιδεύσουμε επιτυχώς τα στατιστικά μοντέλα.

Το πρόβλημα που προαναφέρθηκε λύνεται χρησιμοποιώντας κατανεμημένες αναπαραστάσεις των λέξεων. Χρησιμοποιώντας την παραπάνω τεχνική οι λέξεις που έχουν παρόμοια σημασία θα είναι σε κοντινή θέση στον χώρο διανυσμάτων. Με αυτό τον τρόπο, οι λέξεις πλέον δεν είναι τελείως ανεξάρτητες η μία από την άλλη διότι η κάθε λέξη έχει πληροφορίες και για τις υπόλοιπες.

2.2: Ιστορική αναδρομή

Η τεχνική του να αναπαρίστανται λέξεις ως διανύσματα έχει τις ρίζες κάπου στην δεκαετία του 1970 με την ανάπτυξη του Vector Space Model το οποίο είναι ένα μοντέλο που αναπαριστά κείμενο ως διανύσματα γνωρισμάτων και χρησιμοποιείται για φιλτράρισμα ή ανάκτηση πληροφορίας.

Αργότερα στα τέλη της δεκαετίας του 1980, η μείωση των διαστάσεων χρησιμοποιώντας αποδόμηση μοναδιαίας τιμής (Singular Value Decomposition) εισήγαγε την λανθάνουσα σημασιολογική ανάλυση (Latent Semantic Analysis) η οποία είναι μία τεχνική με την οποία αναλύονται σχέσεις μεταξύ εγγράφων και οι όροι που υπάρχουν σε αυτά και πρόγονος των τεχνικών που χρησιμοποιούνται σήμερα. Η τεχνική αυτή υποθέτει πως λέξεις που έχουν παρόμοιο νόημα βρίσκονται σε παρόμοια κομμάτια κειμένου.

Στις αρχές του 2000 προτάθηκαν από τον Bengio κ.α μια σειρά από νευρωνικά πιθανολογικά μοντέλα γλώσσας για να μειώσει τον μεγάλο αριθμό διαστάσεων των αναπαραστάσεων των λέξεων μαθαίνοντας μια κατανεμημένη αναπαράσταση για τις λέξεις. Ο όρος word embeddings αρχικά χρησιμοποιήθηκε από τον Bengio το 2003 όταν εκπαίδευσε ένα νευρικό μοντέλο γλώσσας μαζί με τις παραμέτρους του μοντέλου.

Το 2008 οι Collobert και Weston ήταν οι πρώτοι που πραγματικά έδειξαν τις δυνατότητες των word embeddings αποδεικνύοντας ότι είναι ένα πολύ καλό και αποδοτικό εργαλείο. Ακόμη, παρουσιάζουν μια αρχιτεκτονική νευρωνικών δικτύων που είναι η βάση πολλών αρχιτεκτονικών σήμερα.

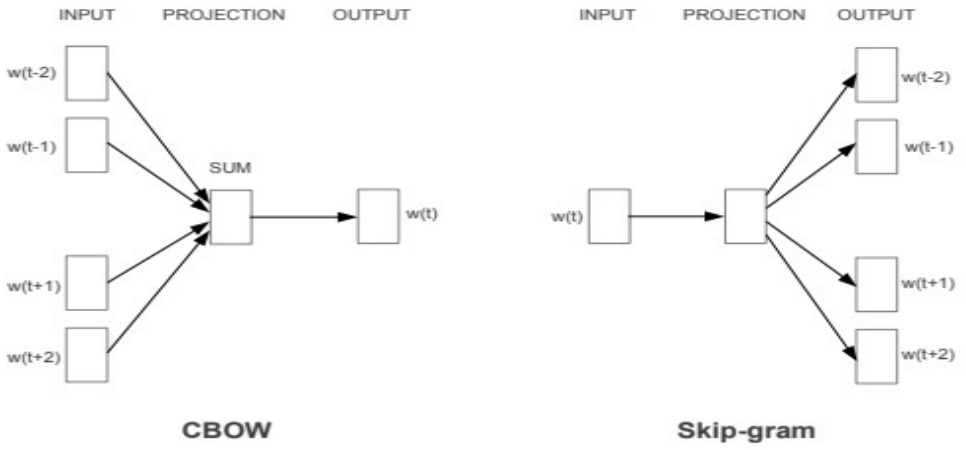
Ωστόσο, αυτός που πραγματικά έφερε στο προσκήνιο τα word embeddings ήταν ο Mikolov το 2013 με την δημιουργία του word2vec μία εργαλειοθήκη που επιτρέπει την εκπαίδευση και χρήση word embeddings. Ένα χρόνο αργότερα ο Pennington

κ.α δημοσίευσαν το GloVe μια παρόμοια ανταγωνιστική εργαλειοθήκη. Οι δύο προαναφερθέντες μέθοδοι για την δημιουργία και χρήση word embeddings θα αναλυθούν με παραπάνω λεπτομέρεια παρακάτω.

2.3: Word2vec

Το Word2vec είναι μια ομάδα ιδιαίτερα αποτελεσματικών υπολογιστικών μοντέλων για την εκμάθηση word embeddings από ακατέργαστο κείμενο. Αυτά τα μοντέλα είναι ρηχά νευρωνικά δίκτυα δύο στρωμάτων (layers) που εκπαιδεύονται από ένα μεγάλο όγκου κειμένου και παράγουν ένα χώρο διανυσμάτων συνήθως μερικών εκατοντάδων διαστάσεων. Η κάθε μοναδική λέξη του κειμένου αντιστοιχίζεται σε ένα διάνυσμα σε αυτό τον χώρο, λέξεις που έχουν παρόμοια σημασία είναι σε κοντινή απόσταση σε αυτό τον χώρο.

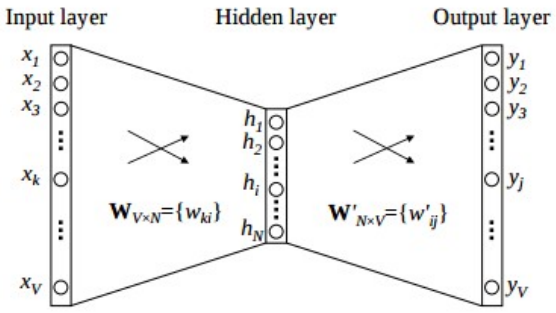
Το Word2vec μπορεί να χρησιμοποιήσει δύο αρχιτεκτονικές μοντέλων για την δημιουργία των word embeddings. Οι αρχιτεκτονικές αυτές είναι οι continuous bag-of-words (CBOW) και η skip-gram. Στην περίπτωση που χρησιμοποιείται CBOW το μοντέλο προβλέπει την τρέχουσα λέξη από ένα παράθυρο γειτονικών λέξεων, η σειρά των λέξεων δεν επηρεάζει την πρόβλεψη αυτή. Στην περίπτωση που χρησιμοποιείται skip-gram το μοντέλο χρησιμοποιεί την τρέχων λέξη για να προβλέψει τις γειτονικές λέξεις. Σε αυτή την αρχιτεκτονική οι λέξεις που είναι πλησιέστερες σε σημασία έχουν μεγαλύτερη επιρροή από λέξεις που λέξεις που είναι πιο απομακρυσμένες σε σημασία. Στην εικόνα που ακολουθεί απεικονίζονται οι δύο αυτές αρχιτεκτονικές.



Σχήμα 1: Οι 2 Αρχιτεκτονικές του word2vec

Η αρχιτεκτονική CBOW

Όπως αναφέρθηκε παραπάνω η αρχιτεκτονική CBOW προσπαθεί προβλέψει την πιθανότητα μιας λέξης δεδομένου ενός context. Το context αυτό μπορεί να είναι μια ενιαία λέξη ή μια ομάδα λέξεων. Στην παρακάτω εικόνα το context είναι ίσο με ένα. Τα δεδομένα στέλνονται σε ένα ρηχό νευρωνικό δίκτυο αποτελούμενο από τρία στρώματα ένα στρώμα εισόδου ένα κρυφό στρώμα και ένα στρώμα εξόδου.

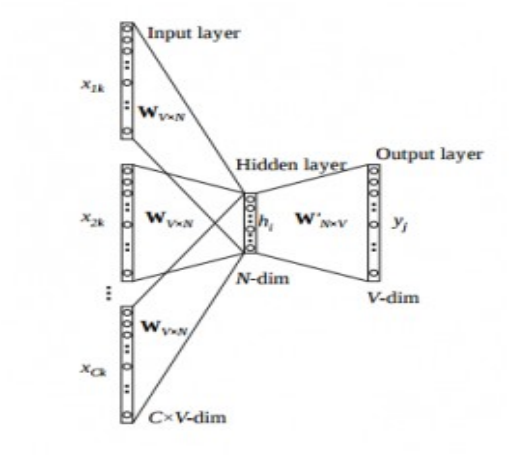


Σχήμα 2: Η αρχιτεκτονική CBOW

Ο τρόπος με τον οποίο υπολογίζονται οι αναπαραστάσεις στην περίπτωση που το context είναι μία λέξη συνοπτικά είναι ο παρακάτω.

1. Το στρώμα εισόδου και εξόδου είναι και τα 2 ένα διάνυσμα one-hot encoded μεγέθους [1 X V] όπου V είναι το μέγεθος του λεξικού μας.
2. Υπάρχουν δύο πίνακες βαρών W και W'. Ο ένας είναι μεταξύ του στρώματος εισόδου και του κρυφού και ο δεύτερος μεταξύ του κρυφού και του στρώματος εξόδου. Σαν είσοδος στο κρυφό στρώμα δίνεται το W το οποίο ένας πίνακας μεγέθους [V X N] Σαν είσοδος στο κρυφό στρώμα δίνεται το W το οποίο ένας πίνακας μεγέθους [V X N] , η έξοδος του κρυφού στρώματος είναι το W' και το μέγεθος του είναι [N X V] όπου N είναι ο αριθμός των διαστάσεων που επιλέγουμε να αναπαραστήσουμε την κάθε λέξη. Το N είναι υπερ-παράμετρος του νευρωνικού δικτύου και μπορεί να διαφέρει, συνήθως επιλέγεται N μεταξύ του 50 και 500. Επίσης το N είναι ο αριθμός των νευρώνων στο κρυφό στρώμα.
3. Το στρώμα εισόδου πολλαπλασιάζεται με τα βάρη W και ονομάζεται κρυφή ενεργοποίηση.
4. Στην συνέχεια το αποτέλεσμα από το κρυφό στρώμα πολλαπλασιάζεται με τα βάρη W' και υπολογίζεται η έξοδος.
5. Αργότερα υπολογίζεται το λάθος μεταξύ στόχου και του αποτελέσματος που υπολογίσθηκε και με back-propagation προσαρμόζονται τα βάρη.
6. Το βάρος μεταξύ του κρυφού στρώματος και του στρώματος εξόδου λαμβάνεται ως αναπαράσταση διανύσματος της λέξης.

Στην περίπτωση που χρησιμοποιούνται παραπάνω από μία λέξεις σαν context η αρχιτεκτονική δείχνει ως παρακάτω.



Σχήμα 3: Η αρχιτεκτονική CBOW με context ίσο με 1

Τα βήματα παραμένουν τα ίδια με την περίπτωση που έχουμε μόνο μία λέξη σαν context, αλλάζει μόνο ο υπολογισμός της κρυφής ενεργοποίησης. Αντί να αντιγράψουμε απλώς τις αντίστοιχες σειρές του κρυμμένου πίνακα μήκους εισόδου στο κρυφό στρώμα, λαμβάνεται ένας μέσος όρος για όλες τις αντίστοιχες σειρές του πίνακα. Το μέσο διάνυσμα που υπολογίζεται γίνεται η κρυφή ενεργοποίηση.

Η αρχιτεκτονική Skip-gram

Στην αρχιτεκτονική Skip-gram προσπαθούμε να προβλέψουμε το context δεδομένης μιας λέξης, είναι σαν να γυρνάμε την αρχιτεκτονική CBOW ανάποδα. Το διάνυσμα εισόδου είναι παρόμοιο με την αρχιτεκτονική CBOW. Επίσης οι υπολογισμοί μέχρι τις ενεργοποιήσεις του κρυφού στρώματος είναι οι ίδιοι. Η διαφορά είναι στον στόχο, στην περίπτωση που έχουμε context ίσο με ένα προσπαθούμε να προβλέψουμε την λέξη πριν και μετά από την λέξη που δόθηκε σαν είσοδος οπότε θα έχουμε δυό one-hot-encoded διανύσματα σαν έξοδο.

Υπολογίζονται και προστίθενται τα λάθη για τα διανύσματα στόχους και στην συνέχεια χρησιμοποιούνται για να προσαρμόσουν τα βάρη. Τελικά σαν αναπαράσταση της λέξης παίρνονται τα βάρη μεταξύ του στρώματος εισόδου και του κρυφού.

2.4: Glove

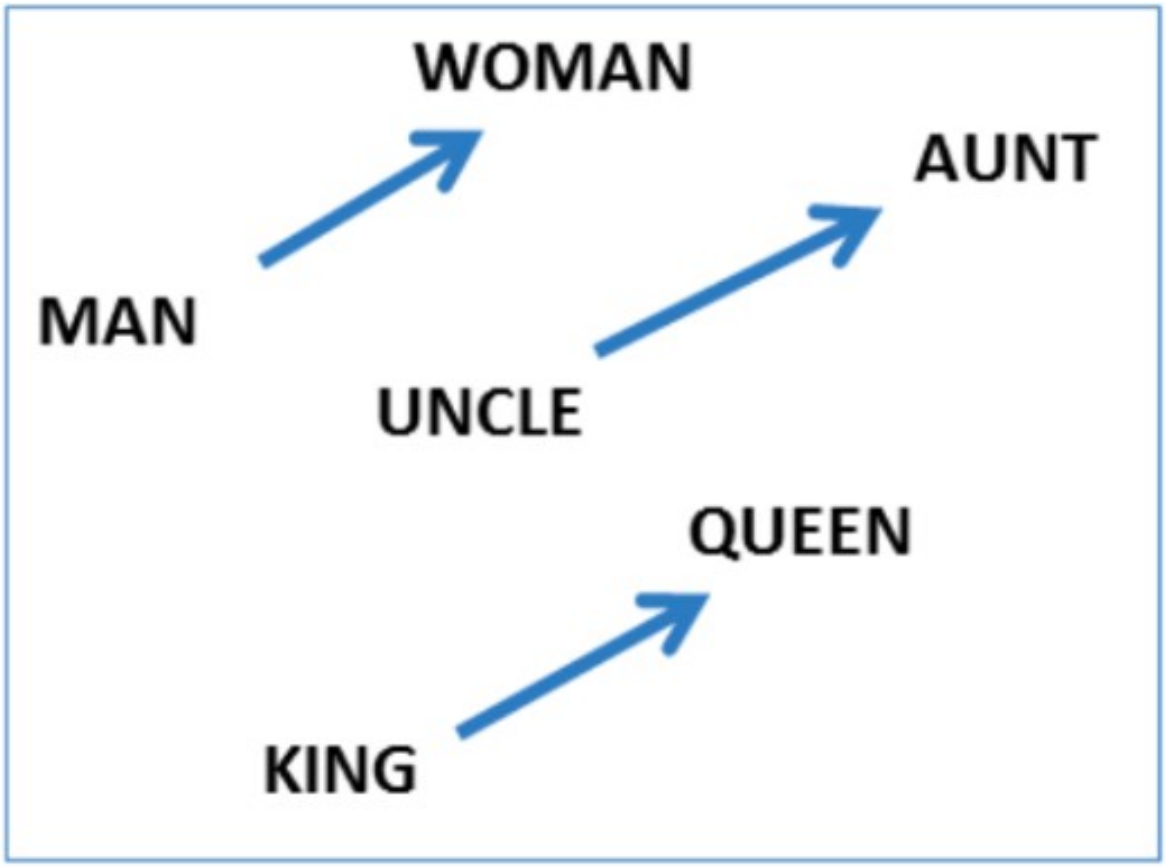
Το GloVe (Global Vectors for Word Representation) είναι ένα άλλο μοντέλο για την αναπαράσταση λέξεων ως διανύσματα. Όπως και το word2vec η μάθηση γίνεται χωρίς επίβλεψη. Η διαφορά σε αυτή την περίπτωση είναι ότι η εκπαίδευση πραγματοποιείται σε συγκεντρωτικά στατιστικά στοιχεία συνεμφάνισης λέξεων σε αντίθεση με το word2vec που εκπαιδεύεται στο να βρίσκει το πλαίσιο δεδομένης μια λέξης. Αφού εκπαιδευτεί ένα τέτοιο μοντέλο οι προκύπτουσες αναπαραστάσεις παρουσιάζουν ενδιαφέρουσες γραμμικές υποσυνθέσεις του χώρου διανυσμάτων λέξης οι οποίες παρουσιάζονται στο επόμενο υποκεφάλαιο.

Ο στόχος του GloVe κατά την εκπαίδευση είναι να μάθει διανύσματα από λέξεις έτσι ώστε το το εσωτερικό γινόμενο τους να ισούται με τον λογάριθμο της πιθανότητας συνεμφάνισης των λέξεων αυτών. Λόγω του γεγονότος ότι ο λογάριθμος ενός λόγου ισούται με τη διαφορά των λογαρίθμων, ο στόχος αυτός συσχετίζει (τους λογάριθμους) τις αναλογίες των πιθανών συνεμφανίσεων με διαφορές διανυσμάτων στον χώρο διανυσμάτων λέξεων. Επειδή αυτές οι αναλογίες μπορούν να κωδικοποιήσουν κάποια μορφή νοήματος, αυτές οι πληροφορίες κωδικοποιούνται και ως διανυσματικές διαφορές. Για το λόγο αυτό, τα διάνυσμα λέξεων που προκύπτουν είναι πολύ καλά σε εργασίες αναλογίας λέξεων.

2.5: Linguistic regularities

Αφού έχει εκπαιδευτεί ένα μοντέλο από ένα σώμα κειμένου μπορούν να παρατηρηθούν κάποιες γραμμικές ιδιαιτερότητες. Οι αναπαραστάσεις που μαθαίνει το μοντέλο στην πραγματικότητα καταγράφουν συντακτικές και σημασιολογικές ιδιαιτερότητες με έναν πολύ απλό τρόπο. Συγκεκριμένα, οι ιδιαιτερότητες παρατηρούνται ως σταθερές αντισταθμίσεις διανυσμάτων μεταξύ ζευγών λέξεων που μοιράζονται μια συγκεκριμένη σχέση. Τα διανύσματα αυτά είναι πολύ καλά στο να απαντούν σε ερωτήσεις αναλογίας όπως το

α είναι στο β οπως το γ είναι στο; Για παράδειγμα, ο άντρας είναι στην γυναίκα όπως ο θείος σε; Η απάντηση είναι θεία. Η παραπάνω αναλογία υπολογίζεται χρησιμοποιώντας μία μέθοδο μετατόπισης διανύσματος βασισμένη στην απόσταση συνημιτόνου.



Σχήμα 4: Γραμμικές ιδιαιτερότητες λέξεων

Ακόμη μπορούμε να κάνουμε αλγεβρικές πράξεις πάνω στα διανύσματα για παράδειγμα το διάνυσμα("Βασιλιάς") - διάνυσμα(“Άντρας”) + διάνυσμα("Γυναίκα") καταλήγει σε ένα διάνυσμα που είναι πιο κοντά στο διάνυσμα που αναπαριστά την λέξη Βασίλισσα.

ΚΕΦΑΛΑΙΟ 3: Μέθοδοι Μηχανικής Μάθησης και εργαλεία

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο παρουσιάζονται οι μέθοδοι μηχανικής μάθησης καθώς και τα εργαλεία που χρησιμοποιήθηκαν για την ανάπτυξη και την εκτέλεση των πειραμάτων που ακολουθούν στο επόμενο κεφάλαιο.

3.1: Support vector machines

Τα Support Vector Machines (SVMs) είναι μία ομάδα αλγορίθμων επιτηρούμενης μάθησης που αρχικά χρησιμοποιήθηκαν για κατηγοριοποίηση ενώ αργότερα εφαρμόστηκαν και σε προβλήματα παλινδρόμησης. Αναπτύχθηκαν για πρώτη φορά από τον Vapnik και τους συνεργάτες τους στο AT&T Bell Labs. Απέκτησαν γρήγορα δημοσιότητα καθώς παρουσίασαν μεγάλη ικανότητα γενίκευσης σε σχέση με άλλες παραδοσιακές μεθόδους ταξινόμησης. Η κατηγοριοποίηση των δεδομένων στηρίζεται στην εύρεση ενός βέλτιστου υπερεπιπέδου που διαχωρίζει τα δεδομένα δημιουργώντας το μέγιστο περιθώριο. Στην περίπτωση που ο γραμμικός διαχωρισμός είναι αδύνατος, γίνεται χρήση κατάλληλων απεικονίσεων που μεταφέρουν το σύνολο των δεδομένων σε μεγαλύτερη διάσταση ώστε να επιτευχθεί τελικά ο διαχωρισμός τους. Η ικανότητα γενίκευσης της χρήσης των SVM σε μη γραμμικά δεδομένα στηρίζεται στο τέχνασμα του πυρήνα (kernel trick). Κάθε μηχανή διανυσμάτων υποστήριξης είναι ένας δυαδικός ταξινομητής, έχει δηλαδή τη δυνατότητα κατηγοριοποίησης σε δύο κλάσεις. Εάν οι κλάσεις είναι περισσότερες, τότε κρίνεται απαραίτητη η χρήση περισσότερων SVMs και η εφαρμογή διαφόρων τεχνικών.

Στην παρούσα εργασία τα SVMs με διαφορετικούς πυρήνες (kernels) είναι ο μόνος τρόπος μηχανικής μάθησης που θα χρησιμοποιηθεί για την ταξινόμηση των άρθρων σε ψευδή και μη.

3.2: Εργαλεία και γλώσσες προγραμματισμού

Python

Για όλα τα πειράματα και την προεπεξεργασία των άρθρων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python, επίσης όλες οι βιβλιοθήκες που ακολουθούν είναι σε Python.

SciPy και Numpy

Το SciPy είναι ένα οικοσύστημα ανοιχτού κώδικα γραμμένο σε Python για τα μαθηματικά, την επιστήμη και τη μηχανική και χρησιμοποιείται για επιστημονικούς υπολογισμούς το NumPy είναι ένα από τα βασικά πακέτα του SciPy.

Το NumPy είναι το θεμελιώδες πακέτο για επιστημονικό υπολογισμό με την Python. Μας επιτρέπει να κάνουμε αποτελεσματικούς υπολογισμούς και προσφέρει πολλές λειτουργίες γραμμικής άλγεβρας επίσης μπορεί να χρησιμοποιηθεί ως πολυδιάστατο δοχείο γενικών δεδομένων και μπορούμε να ορίσουμε σύνθετους τύπους δεδομένων.

Nltk (Natural Language Toolkit

Το NLTK είναι μια πλατφόρμα για την κατασκευή προγραμμάτων Python για την επεξεργασία δεδομένων που έχουν να κάνουν με την ανθρώπινη γλώσσα. Παρέχει εύχρηστες διεπαφές σε πάνω από 50 κορμούς και λεξικά μέσα όπως το WordNet, μαζί με μια σουίτα βιβλιοθηκών επεξεργασίας κειμένου classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

Gensim

Το Gensim είναι ένα ισχυρό εργαλείο μοντελοποίησης διανυσματικών χώρων ανοιχτού κώδικα και εργαλείων μοντελοποίησης. Χρησιμοποιεί τα πακέτα NumPy, SciPy και προαιρετικά Cython για απόδοση. Το Gensim έχει σχεδιαστεί ειδικά για να χειρίζεται μεγάλες συλλογές κειμένων, χρησιμοποιώντας streaming δεδομένων και αποδοτικούς αυξητικούς αλγόριθμους, το οποίο το διαφοροποιεί από τα

περισσότερα πακέτα επιστημονικού λογισμικού που στοχεύουν μόνο επεξεργασία παρτίδας(batch) και εντός της μνήμης.

Scikit-learn

Το Scikit-learn είναι μια βιβλιοθήκη μηχανικής μάθησης η οποία προσφέρει υλοποιήσεις από πολλούς αλγόριθμους όπως support vector machines, random forests, gradient boosting, k-means and DBSCAN και είναι φτιαγμένη ώστε να λειτουργεί με τα NumPy και SciPy.

Pandas

Το pandas είναι βιβλιοθήκη ανοικτού κώδικα που παρέχει υψηλής απόδοσης, εύκολες στη χρήση δομές δεδομένων και εργαλεία ανάλυσης δεδομένων για τη γλώσσα προγραμματισμού Python.

Matplotlib

Το Matplotlib είναι μια βιβλιοθήκη σχεδίασης Python 2D, η οποία χρησιμοποιείται για το σχεδιασμό γραφημάτων και σχεδίων

ΚΕΦΑΛΑΙΟ 4: Πειραματική εφαρμογή και αποτελέσματα

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο αρχικά περιγράφονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τα πειράματα. Στην συνέχεια περιγράφεται ο τρόπος προ επεξεργασίας των δεδομένων για να χρησιμοποιηθούν από μεθόδους μηχανικής μάθησης και τελικά περιγράφονται και ερμηνεύονται τα αποτελέσματα που προκύπτουν από τα πειράματα.

4.1: Σύνολα δεδομένων

Για τα πειράματα που εκτελέσθηκαν χρησιμοποιήθηκαν δύο διαφορετικά σύνολα δεδομένων. Τα σύνολα αυτά συλλέχθηκαν από τον κ. Γραβάνη Γεώργιο και συγκεκριμένα είναι τα UNBFakeNews και KaggleEXT.

Στην περίπτωση του UNBFakeNews τα ψευδή νέα συλλέχθηκαν από οργανισμούς που κάνουν fact checking, μερικοί από αυτούς είναι οι Snopes.com, TruthOrFiction.com και Politifact.org. Για τα μη ψευδή νέα χρησιμοποιήθηκαν άρθρα από ακριβής και αξιόπιστες οργανώσεις δημοσιογραφίας όπως οι The New York Times, The Wall Street Journal, The Washington Post κ.α. Σημαντικό να αναφερθεί είναι ότι τα άρθρα που χρησιμοποιήθηκαν είναι από διάφορες κατηγορίες όπως το ταξίδι, η τεχνολογία, ο αθλητισμός, η ζωή, πολιτικά κ.α. Το σύνολο δεδομένων αυτό συνολικά περιέχει 3400 άρθρα, από τα οποία τα 1400 είναι ψευδή και τα υπόλοιπα 2000 πραγματικά.

Στην περίπτωση του KaggleEXT χρησιμοποιείται σαν βάση το σύνολο δεδομένων με τα ψευδή νέα από το kaggle το οποίο στην συνέχεια εμπλουτίστηκε με πραγματικά νέα με παρόμοιο τρόπο όπως και το UNB. Σε αυτό το σύνολο υπάρχουν συνολικά 18600 άρθρα από τα οποία τα 12400 είναι ψευδή ενώ τα υπόλοιπα 6200 πραγματικά.

4.2: Προεπεξεργασία Δεδομένων

Τα δεδομένα έπρεπε να αναπαρασταθούν με τους δύο τρόπους που αναφέρονται στο δεύτερο κεφάλαιο, ο πρώτος είναι με word2vec και ο δευτέρος με GloVe. Για

την περίπτωση του word2vec χρησιμοποιήθηκε ένα ήδη εκπαιδευμένο μοντέλο πάνω στα Google News το οποίο περιέχει διανύσματα 300 διαστάσεων για 3 εκατομμύρια λέξεις. Στην περίπτωση του GloVe χρησιμοποιήθηκε ένα ήδη εκπαιδευμένο μοντέλο πάνω σε δεδομένα από το B τα διανύσματα είναι 300 διαστάσεων και οι λέξεις που περιέχει είναι 1.9 εκατομμύρια.

Το πρώτο βήμα είναι να φορτώσουμε το αντίστοιχο μοντέλο που μόλις αναφέρθηκε, η υπόλοιπη διαδικασία για την προ επεξεργασία είναι η ίδια ανεξαρτήτως του μοντέλου. Καθώς τα δεδομένα ήταν σε μορφή βάσης δεδομένων το πρώτο βήμα ήταν να φορτωθούν τα δεδομένα από την βάση. Το επόμενο βήμα ήταν, χρησιμοποιώντας το εργαλείο nltk, να σπάσουμε το κάθε άρθρο σε λέξεις η διαδικασία αυτή είναι γνωστή ως tokenization, ταυτόχρονα μετατράπηκαν και όλες οι λέξεις σε πεζά. Στην συνέχεια χρησιμοποιώντας το ίδιο εργαλείο αφαιρέθηκαν από το κάθε άρθρο stop words και αριθμοί, stop words είναι οι λέξεις που εμφανίζονται πολύ συχνά και δεν έχουν κάποια ιδιαίτερη σημασία μερικές από αυτές τις λέξεις είναι οι the, a, and, that κ.α. Στην συνέχεια για τις λέξεις του κάθε άρθρου ψάχνουμε στο αντίστοιχο μοντέλο που έχει φορτωθεί το διάνυσμα για την κάθε λέξη και τελικά προσθέτουμε όλα τα διανύσματα για το κάθε άρθρο. Έτσι αναπαριστάται ένα άρθρο ως το άθροισμα των διανυσμάτων των λέξεων του.

Τελικά, δημιουργήθηκαν αναπαραστάσεις δεδομένων χρησιμοποιώντας τις αναπαραστάσεις που δημιουργήθηκαν για κάθε άρθρο με word2vec είτε GloVe σε συνδυασμό με ένα σύνολο γλωσσικών χαρακτηριστικών τα οποία δημιουργήθηκαν από τον κ. Γεώργιο Γραβάνη σε αυτή την περίπτωση το κάθε άρθρο έχει τα 300 γνωρισμάτα από το GloVe ή word2vec και επιπρόσθετα 92 από τα γλωσσικά χαρακτηριστικά.

4.3: Μεθοδολογία μάθησης

Για όλα τα πειράματα που εκτελέστηκαν χρησιμοποιήθηκαν διανύσματα υποστήριξης μηχανής (Support Vector Machines) για την κατηγοριοποίηση των άρθρων σε ψευδή και και πραγματικά. Κάθε σύνολο δεδομένων χωρίστηκε σε δύο κομμάτια το ένα κομμάτι που είναι το 75% του συνόλου χρησιμοποιήθηκε για την εκπαίδευση ενώ το υπόλοιπο 25% χρησιμοποιήθηκε για την δοκιμή και επαλήθευση του ταξινομητή. Για την εύρεση των καλύτερων παραμέτρων για το κάθε σύνολο δεδομένων και διαφορετικό μοντέλο χρησιμοποιήθηκε grid search με 7 fold cross-validation. Ακόμη στην περίπτωση της σύνθετης αναπαράστασης πριν την εκτέλεση του grid search έγινε και κανονικοποίηση των δεδομένων. Οι παράμετροι που χρησιμοποιήθηκαν παρουσιάζονται στον παρακάτω πίνακα.

Kernel	Linear, rbf
C	1, 10, 100, 1000
Gamma	1.00E-01, 1.00E-02, 1.00E-03, 1.00E-04, 1.00E-05, 1.00E-06

Πίνακας 1: Παράμετροι grid search

4.4: Πειράματα και αποτελέσματα

Πειράματα στο σύνολο δεδομένων UNBFakeNews

Word2vec

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.931	0.931	0.931	0.931

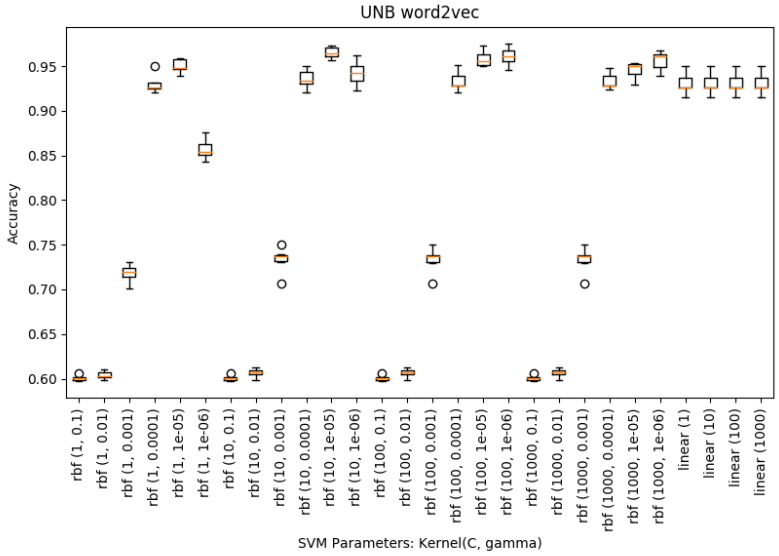
Πίνακας 2: Αποτελέσματα grid search για linear kernel

	C			
Gamma	1	10	100	1000
1.00E-01	0.601	0.601	0.601	0.601
1.00E-02	0.604	0.607	0.607	0.607
1.00E-03	0.718	0.733	0.733	0.733
1.00E-04	0.930	0.936	0.934	0.934
1.00E-05	0.951	**0.966**	0.958	0.946
1.00E-06	0.857	0.942	0.961	0.956

Πίνακας 3: Αποτελέσματα grid search για kernel rbf.

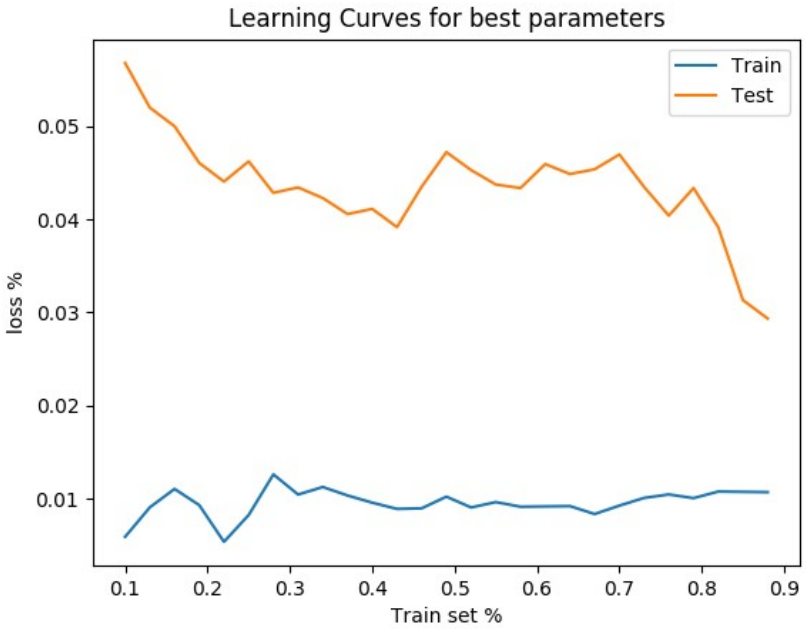
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-05 και C = 10.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 5: Boxplots grid search για word2vec

Στο παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 6: Learning curves για τις καλύτερες παραμέτρους στο word2vec

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.96.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.961	0.969	0.965
Ψευδή	0.959	0.949	0.954
Σύνολο / Μέσος όρος	0.960	0.960	0.960

Πίνακας 4: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για word2vec

Glove

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.941	0.941	0.941	0.941

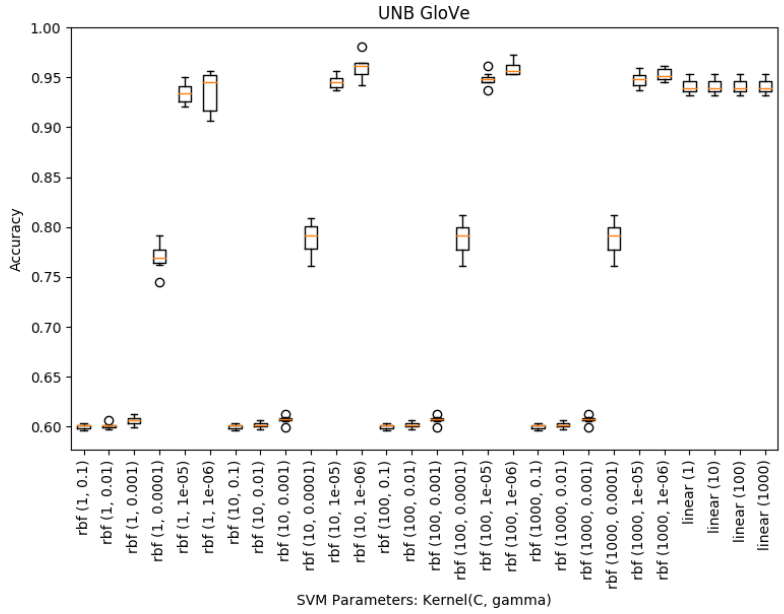
Πίνακας 5: Αποτελέσματα grid search για kernel linear.

	C			
Gamma	1	10	100	1000
1.00E-01	0.600	0.600	0.600	0.600
1.00E-02	0.601	0.602	0.602	0.602
1.00E-03	0.606	0.607	0.607	0.607
1.00E-04	0.770	0.788	0.788	0.788
1.00E-05	0.934	0.945	0.948	0.948
1.00E-06	0.935	**0.960**	0.959	0.953

Πίνακας 6: Αποτελέσματα grid search για kernel rbf.

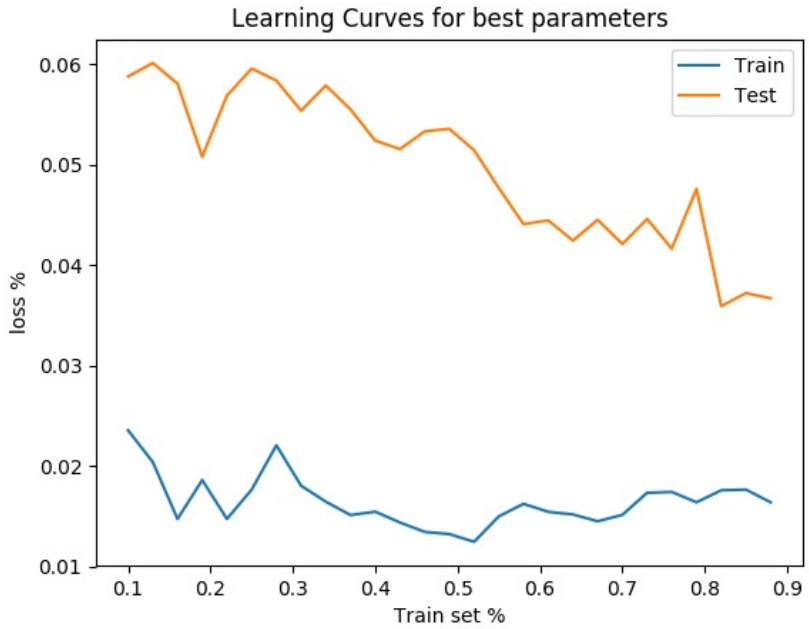
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-06 και C = 10.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 7: Boxplots grid search για glove

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 8: Learning curves για τις καλύτερες παραμέτρους στο GloVe

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.958.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.961	0.965	0.963
Ψευδή	0.954	0.949	0.951
Σύνολο / Μέσος όρος	0.958	0.958	0.958

Πίνακας 7: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για GloVe

Word2vec+linguistic

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.737	0.851	0.940	**0.957**

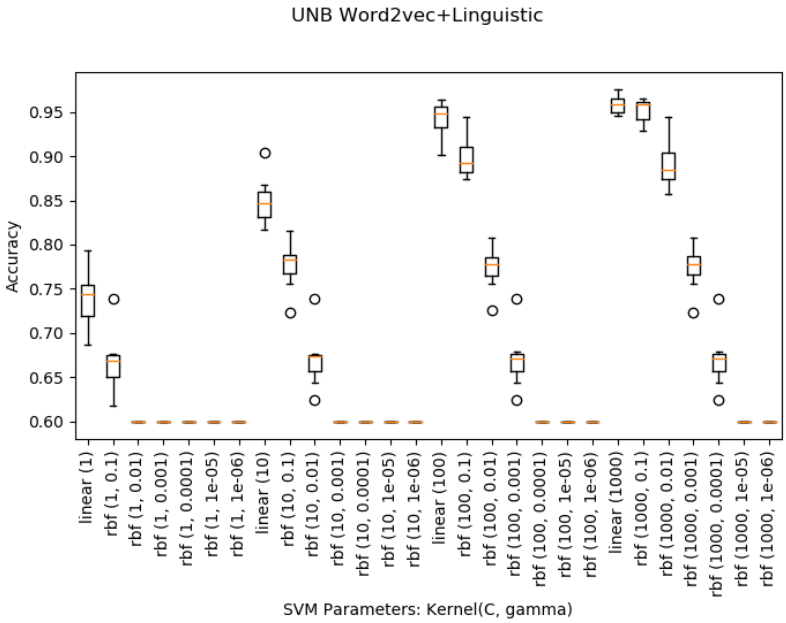
Πίνακας 8: Αποτελέσματα grid search για kernel linear.

	C			
Gamma	1	10	100	1000
1.00E-01	0.670	0.777	0.900	0.949
1.00E-02	0.599	0.673	0.774	0.891
1.00E-03	0.599	0.599	0.673	0.775
1.00E-04	0.599	0.599	0.599	0.673
1.00E-05	0.737	0.599	0.599	0.599
1.00E-06	0.670	0.599	0.599	0.599

Πίνακας 9: Αποτελέσματα grid search για kernel rbf.

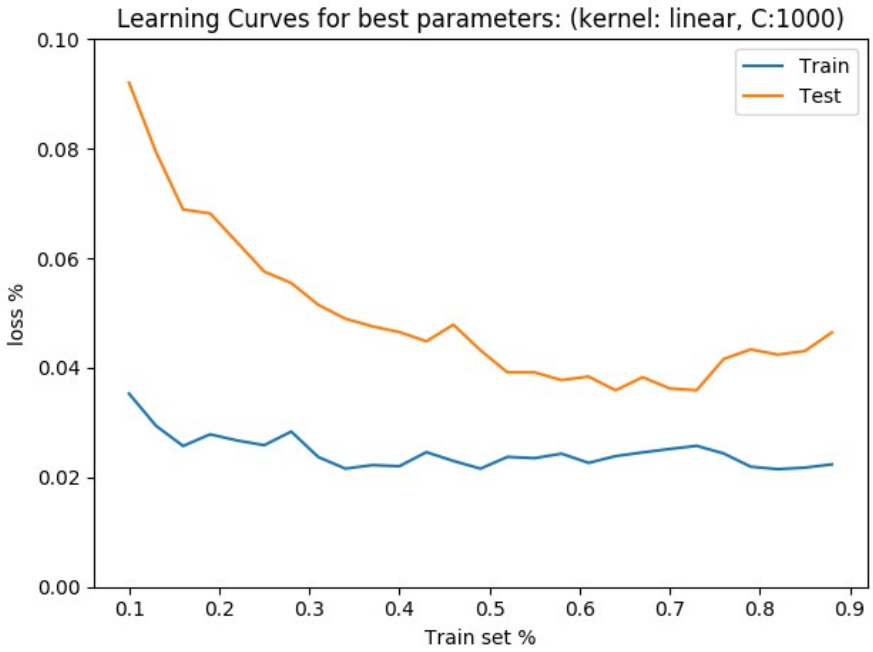
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = linear με C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 9: Boxplots grid search για word2vec+Linguistic

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 10: Learning curves για τις καλύτερες παραμέτρους στο word2vec+Linguistic

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια 0.962, ανάκληση και f1 score ίσο με 0.962.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.974	0.956	0.965
Ψευδή	0.945	0.968	0.957
Σύνολο / Μέσος όρος	0.962	0.961	0.961

Πίνακας 10: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για word2vec+Linguistic

Glove+linguistic

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.807	0.926	0.954	**0.957**

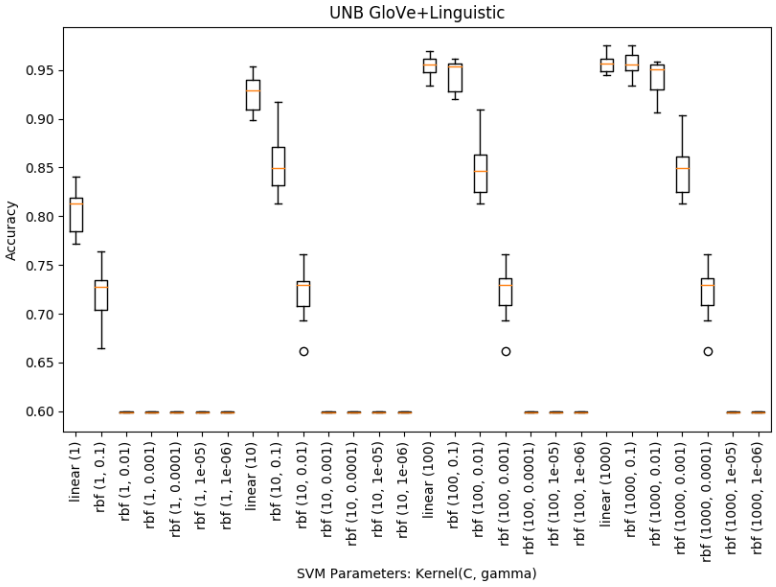
Πίνακας 11: Αποτελέσματα grid search για kernel linear

	C			
Gamma	1	10	100	1000
1.00E-01	0.719	0.855	0.944	0.957
1.00E-02	0.599	0.719	0.849	0.941
1.00E-03	0.599	0.599	0.721	0.848
1.00E-04	0.599	0.599	0.599	0.721
1.00E-05	0.599	0.599	0.599	0.599
1.00E-06	0.599	0.599	0.599	0.599

Πίνακας 12: Αποτελέσματα grid search για kernel rbf

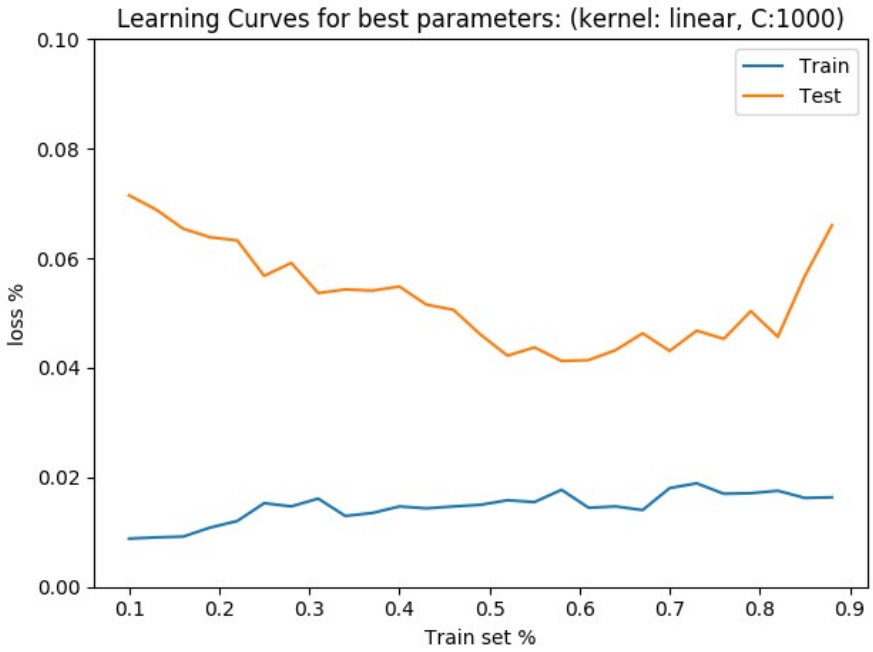
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = linear με C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 11: Boxplots grid search για GloVe+Linguistic

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 12: Learning curves για τις καλύτερες παραμέτρους στο GloVe+Linguistic

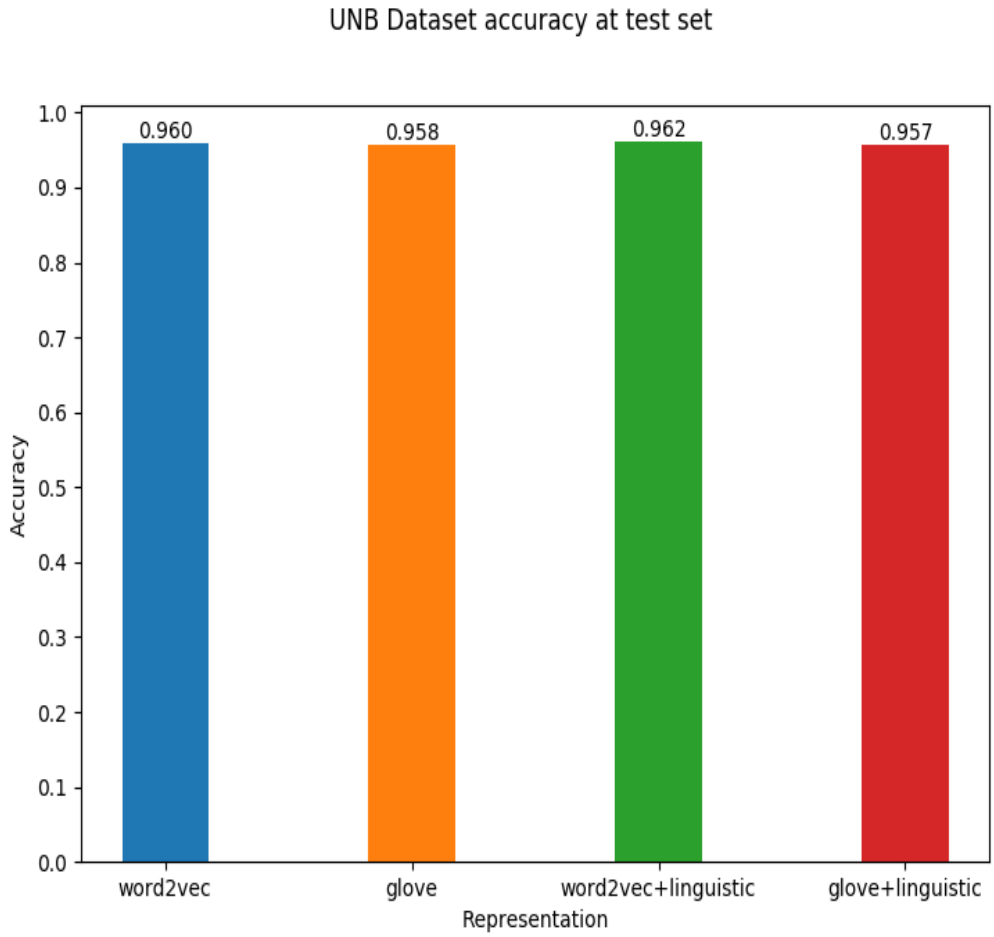
Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.957.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.972	0.949	0.961
Ψευδή	0.938	0.965	0.952
Σύνολο / Μέσος όρος	0.957	0.957	0.957

Πίνακας 13: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για GloVe+Linguistic

Σύγκριση αποτελεσμάτων

Συγκρίνωντας τα αποτελέσματα από όλους τους παραπάνω συνδυασμούς βλέπουμε πως η ακρίβεια που πετυχένουμε είναι σχεδόν ίδια σε ολές τις περιπτώσεις με καλύτερο συνδυασμό να είναι το wor2vec+linguistic όπου πετυχένουμε ακρίβεια 0.962.



Σχήμα 13: Σύγκριση αποτελεσμάτων διάφορων αναπαραστάσεων

Πειράματα στο σύνολο δεδομένων KaggleEXT

Word2vec

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.932	0.941	0.940	0.939

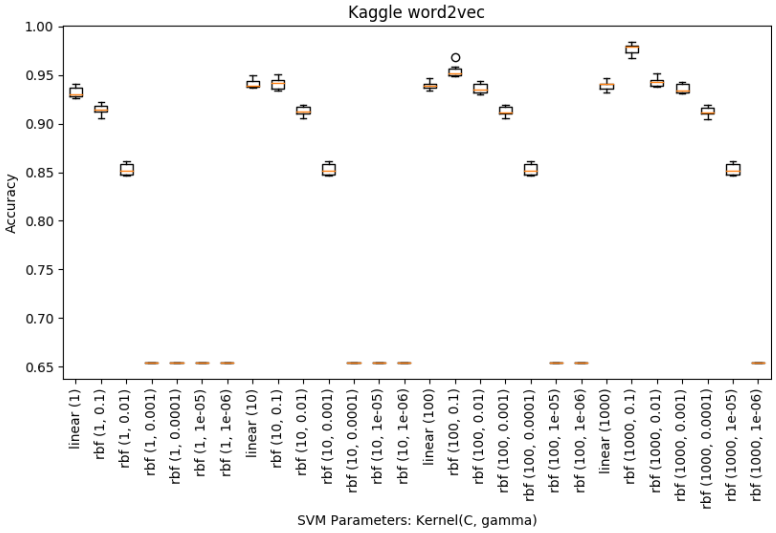
Πίνακας 14: Αποτελέσματα grid search για kernel linear

	C			
Gamma	1	10	100	1000
1.00E-01	0.915	0.941	0.955	**0.977**
1.00E-02	0.853	0.913	0.936	0.943
1.00E-03	0.654	0.853	0.913	0.936
1.00E-04	0.654	0.654	0.853	0.913
1.00E-05	0.654	0.654	0.654	0.853
1.00E-06	0.654	0.654	0.654	0.654

Πίνακας 15: Αποτελέσματα grid search για kernel rbf

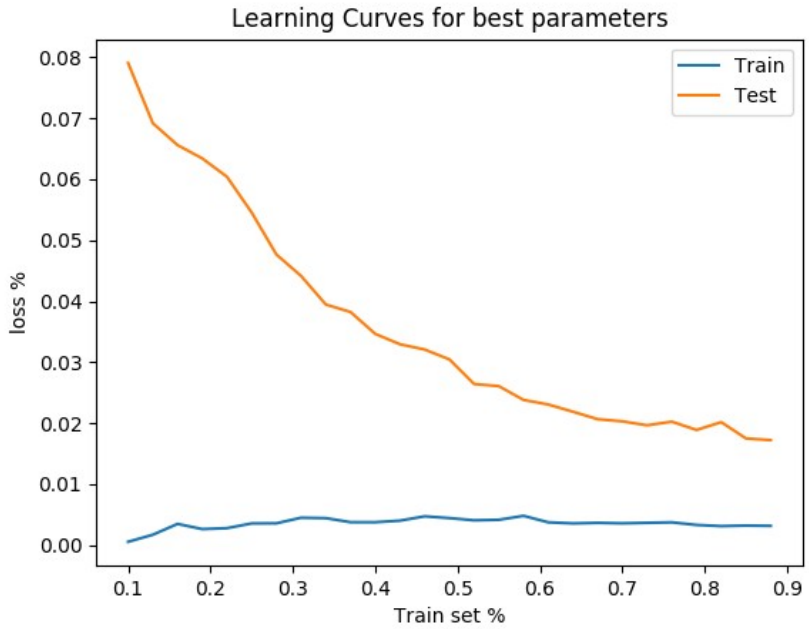
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-01 και C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 14: Boxplots grid search για word2vec

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 15: Learning curves για τις καλύτερες παραμέτρους στο word2vec

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.979.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.961	0.980	0.970
Ψευδή	0.989	0.978	0.984
Σύνολο / Μέσος όρος	0.979	0.979	0.979

Πίνακας 16: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για word2vec

Glove

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.915	0.936	0.940	0.939

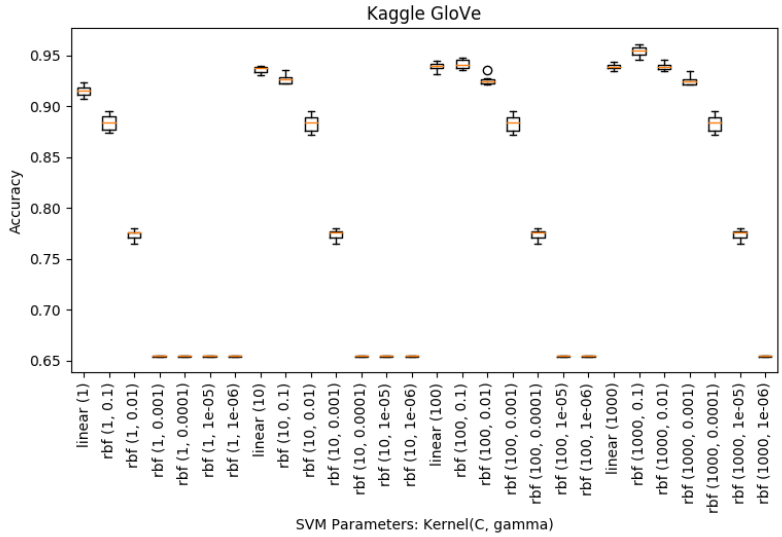
Πίνακας 17: Αποτελέσματα grid search για kernel linear

	C			
Gamma	1	10	100	1000
1.00E-01	0.884	0.927	0.942	**0.954**
1.00E-02	0.774	0.884	0.926	0.939
1.00E-03	0.654	0.774	0.884	0.926
1.00E-04	0.654	0.654	0.774	0.884
1.00E-05	0.654	0.654	0.654	0.774
1.00E-06	0.654	0.654	0.654	0.654

Πίνακας 18: Αποτελέσματα grid search για kernel rbf

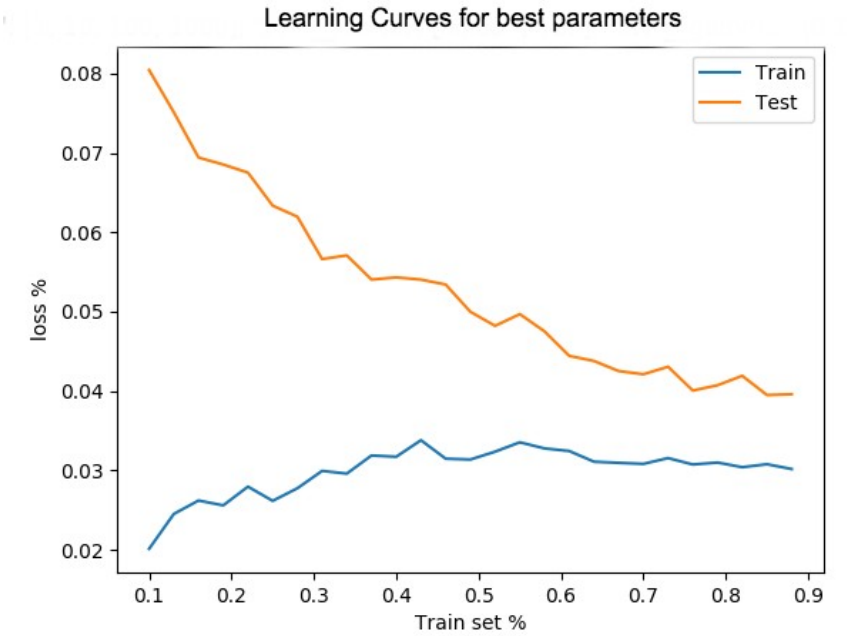
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-01 και C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 16: Boxplots grid search για GloVe

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 17: Learning curves για τις καλύτερες παραμέτρους στο GloVe

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.96.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.944	0.943	0.943
Ψευδή	0.969	0.970	0.969
Σύνολο / Μέσος όρος	0.960	0.960	0.960

Πίνακας 19: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης για GloVe

Word2vec+linguistic

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.938	0.959	0.824	0.990

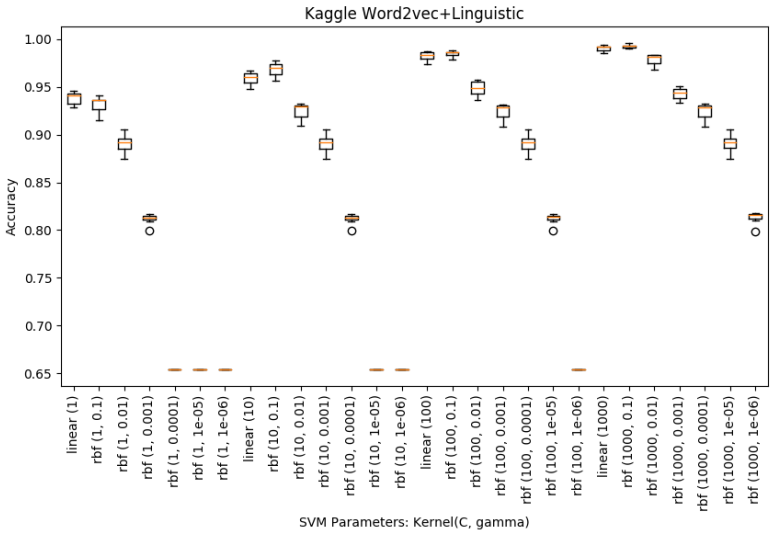
Πίνακας 20: Αποτελέσματα grid search για kernel linear

	C			
Gamma	1	10	100	1000
1.00E-01	0.931	0.968	0.984	**0.992**
1.00E-02	0.891	0.924	0.948	0.979
1.00E-03	0.812	0.891	0.924	0.943
1.00E-04	0.654	0.812	0.891	0.924
1.00E-05	0.654	0.654	0.812	0.891
1.00E-06	0.654	0.654	0.654	0.813

Πίνακας 21: Αποτελέσματα grid search για kernel rbf

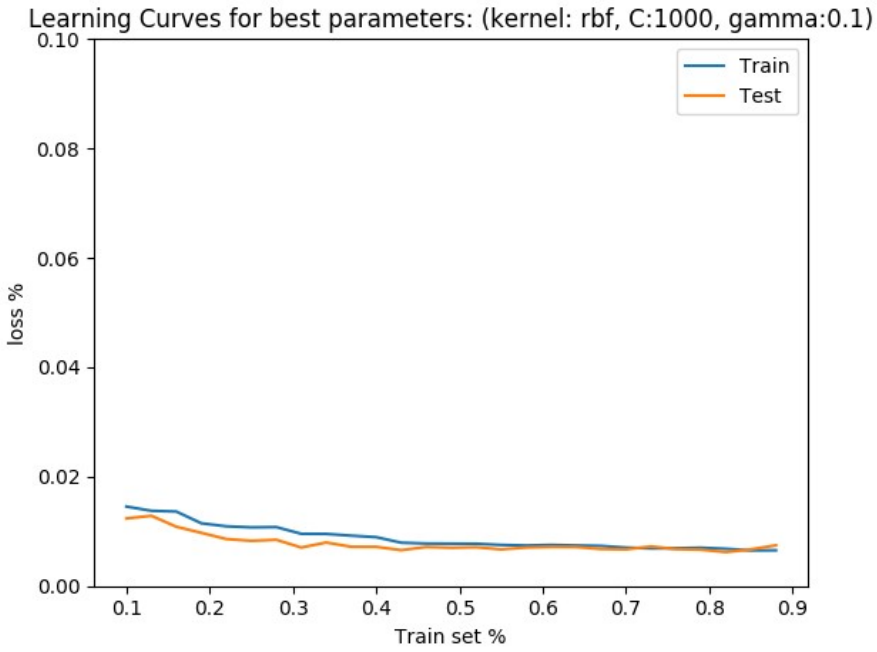
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-01 και C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 18: Boxplots grid search για Word2vec+Linguistic

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε μικρότερο λάθος από οτι στο σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 19: Learning curves για τις καλύτερες παραμέτρους στο Word2vec+Linguistic

Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.993.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.992	0.988	0.990
Ψευδή	0.993	0.996	0.995
Σύνολο / Μέσος όρος	0.993	0.993	0.993

Πίνακας 22: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης

glove+linguistic

Στους παρακάτω 2 πίνακες παρουσιάζονται τα αποτελέσματα από το grid search για kernel ίσο με linear και rbf.

C			
1	10	100	1000
0.958	0.969	0.984	0.991

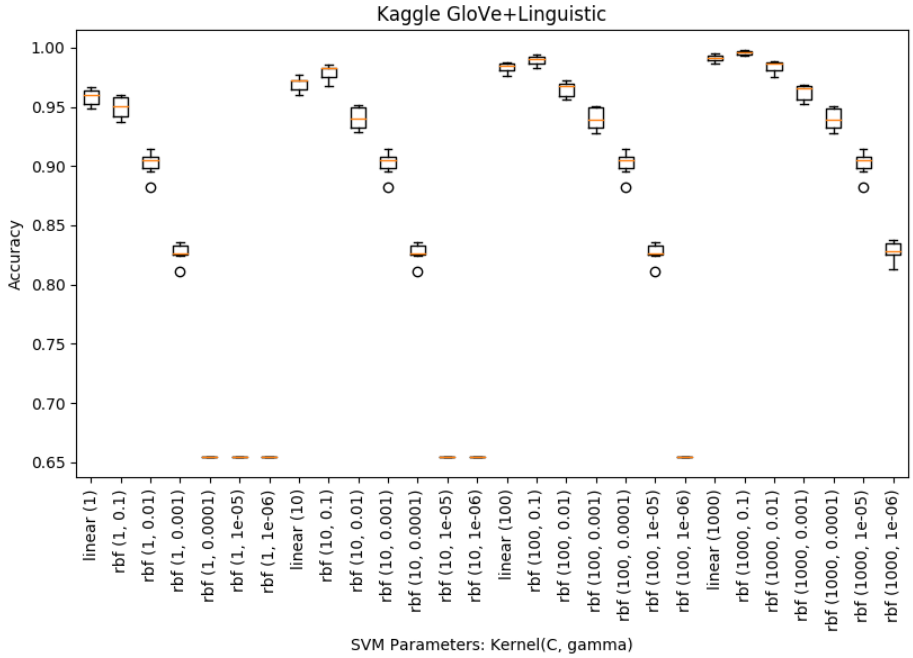
Πίνακας 23: Αποτελέσματα grid search για kernel linear

	C			
Gamma	1	10	100	1000
1.00E-01	0.950	0.979	0.989	**0.996**
1.00E-02	0.902	0.941	0.965	0.984
1.00E-03	0.827	0.902	0.940	0.962
1.00E-04	0.654	0.827	0.902	0.940
1.00E-05	0.654	0.654	0.827	0.902
1.00E-06	0.654	0.654	0.654	0.828

Πίνακας 24: Αποτελέσματα grid search για kernel rbf

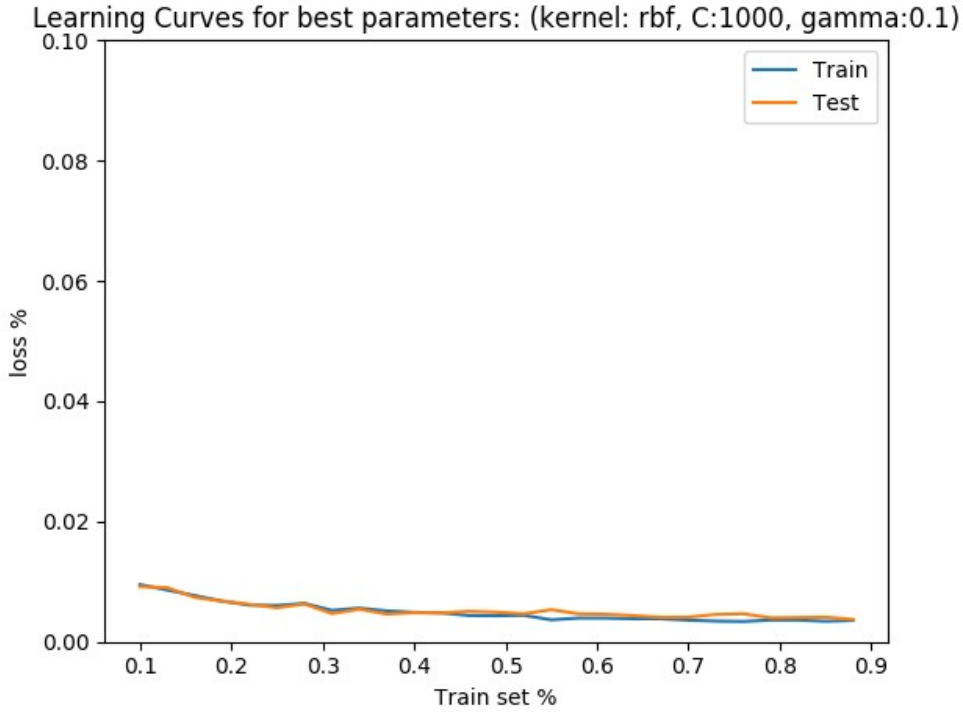
Όπως φαίνεται από τον παραπάνω πίνακα οι καλύτεροι παράμετροι που προκύπτουν είναι kernel = rbf με gamma = 1.00E-01 και C = 1000.

Στο παρακάτω σχήμα παρουσιάζονται boxplots για όλες τις παραμέτρους από όπου μπορούμε να δούμε και να συγκρίνουμε την τυπική απόκλιση και την μέση τιμή που πετυχένουμε με κάθε παραμέτρο.



Σχήμα 20: Boxplots grid search για GloVe+Linguistic

Στα παρακάτω σχήμα παρουσιάζονται learning curves για τις καλύτερες παραμέτρους από όπου μπορούμε να δούμε και να ελέγξουμε αν το μοντέλο μας κάνει overfitting όπως φαίνεται στο σύνολο που χρησιμοποιούμε για την εκπαίδευση έχουμε σχεδόν το ίδιο λάθος με το σύνολο που χρησιμοποιούμε για την επαλήθευση το οποίο δείχνει ότι ότι το μοντέλο μας δεν κάνει overfitting.



Σχήμα 21: Learning curves για τις καλύτερες παραμέτρους στο GloVe+Linguistic

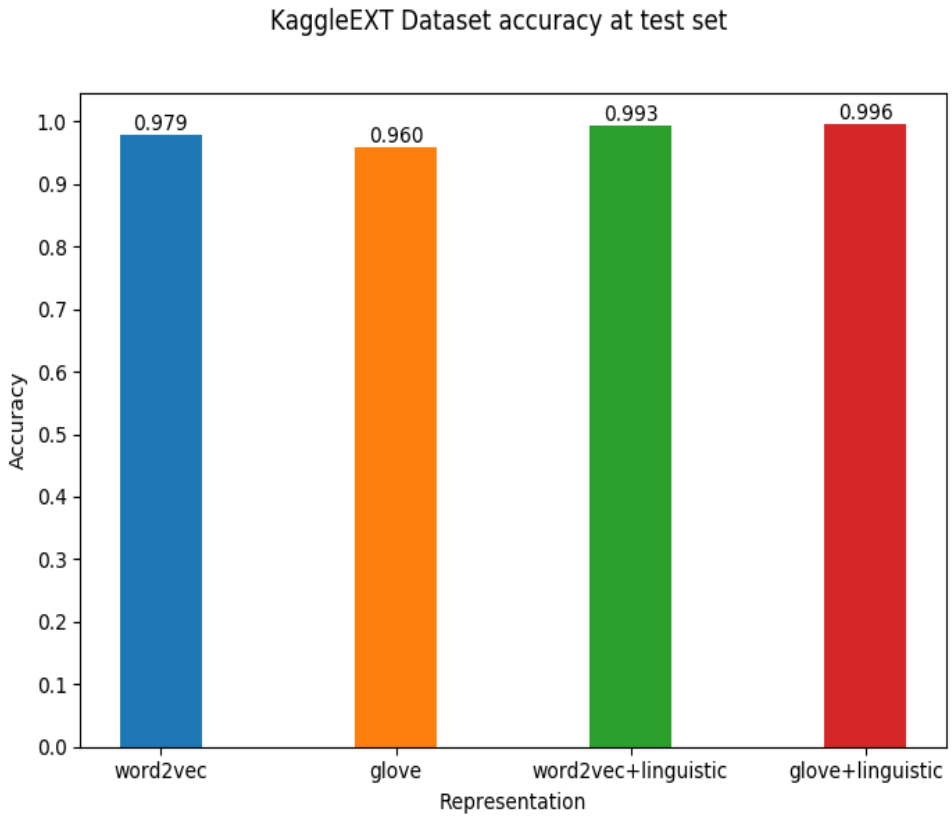
Στον παρακάτω πίνακα παρουσιάζεται η αναφορά ταξινόμησης για τις καλύτερες παραμέτρους που προέκυψαν από το grid search στην οποία βλέπουμε πως κατά μέσο όρο πετυχένουμε ακρίβεια, ανάκληση και f1 score ίσο με 0.996.

Τάξη	Ακρίβεια	Ανάκληση	f1
Πραγματικά	0.994	0.993	0.994
Ψευδή	0.996	0.997	0.997
Σύνολο / Μέσος όρος	0.996	0.996	0.996

Πίνακας 25: Αναφορά ταξινόμησης για τις καλύτερες παραμέτρους στο σύνολο δεδομένων δοκιμής και επαλήθευσης

Σύγκριση αποτελεσμάτων

Συγκρίνωντας τα αποτελέσματα από όλους τους παραπάνω συνδυασμούς βλέπουμε πως η ακρίβεια που πετυχένουμε είναι λίγο μεγαλύτερη στους συνδυσμούς που περιέχουν γνωρίσματα linguistic με τον συνδυασμό glove+linguistic να πετυχαίνει την υψηλότερη ακρίβεια με 0.996.



Σχήμα 22: Σύγκριση αποτελεσμάτων διάφορων αναπαραστάσεων

ΚΕΦΑΛΑΙΟ 5: Συμπεράσματα, προτάσεις βελτίωσης, ιδέες για μελλοντική επέκταση

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο παρουσιάζονται τα συμπεράσματα που προκύπτουν από την πτυχιακή εργασία καθώς και προτάσεις για μελλοντική επέκταση της εργασίας.

ΥΠΟΚΕΦΑΛΑΙΟ 5.1: Συμπεράσματα

Η παρούσα πτυχιακή εργασία έδειξε ότι τα word embeddings είναι ένα πολύ καλό και αποδοτικό εργαλείο για την αναπαράσταση των άρθρων και στην συνέχεια την επεξεργασία και κατηγοριοποίηση τους σε ψευδή και μη. Όπως φάνηκε από τα πειράματα που εκτελέστηκαν στο τέταρτο κεφάλαιο επιτεύχθηκαν πολύ υψηλές ακρίβειες στην ταξινόμηση των άρθρων με όλους τους τρόπους αναπαράστασης.

ΥΠΟΚΕΦΑΛΑΙΟ 5.2: Μελλοντικές επεκτάσεις και προτάσεις βελτίωσης

Για την αναπαράσταση του κειμένου χρησιμοποιήθηκε word2vec, GloVe και ένας συνδυασμός των παραπάνω με ένα σύνολο γλωσσικών χαρακτηριστικών. Πέρα από αυτά θα μπορούσε να χρησιμοποιηθεί το Doc2Vec για την αναπαράσταση του κειμένου το οποίο είναι μια ειδική περίπτωση η επέκταση του word2vec κατά την οποία μαθαίνονται αναπαραστάσεις ολόκληρων προτάσεων παραγράφων ή ακόμα και άρθρων.

Καθώς χρησιμοποιήθηκαν μόνο Support Vector Machines για την ταξινόμηση των άρθρων σε ψευδή ή μη, προτείνεται η χρήση επιπλέον μεθόδων ταξινόμησης, ακόμη θα προτεινόταν η χρήση μοντέλων ensemble καθώς συνήθως τα μοντέλα αυτά αυξάνουν την ακρίβεια.

Τελικά καθώς χρησιμοποιήθηκαν μόνο δύο σύνολα δεδομένων για την εκπαίδευση προτείνεται η χρήση περισσότερων συνόλων δεδομένων από διάφορες πηγές όπως το kaggle.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Άρθρα και βιβλία

Ahmed, Hadeer & Traore, Issa & Saad, Sherif. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques.

Bengio Y., Schwenk H., Senécal JS., Morin F., Gauvain JL. (2006) Neural Probabilistic Language Models. In: Holmes D.E., Jain L.C. (eds) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol 194. Springer, Berlin, Heidelberg.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. ICML.

Horne, B.D., & Adali, S. (2017). This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. CoRR, abs/1703.09398.

Mikolov, T., & Sutskever, I. (2013). Representations of Words and Phrases and their Compositionality.

Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781

Mikolov, T., Yih, W.-t. & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations.. HLT-NAACL(p./pp. 746--751).

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin, 29(5), 665-675. DOI: 10.1177/0146167203029005010

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. EMNLP (p./pp. 1532--1543).

Salton, G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing. Commun. ACM, 18, 613-620.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explorations, 19, 22-36.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P.S. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection. CoRR, abs/1806.00749.

Ιστοσελίδες

(2017). Two-thirds of American adults get news from social media: survey. Ανακτήθηκε 5 Ιουνίου 2018, από, https://www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8

(2017). Fact-checking Tools for Exhausted Journalists. Ανακτήθηκε 5 Ιουνίου 2018, από, https://medium.com/dear-laura/fact-checking-tools-for-overwhelmed-journalists-8ca337c1aef5

(2018). Check Verify breaking news online. Ανακτήθηκε 5 Ιουνίου 2018, από, https://meedan.com/en/check/

(2018). An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec. Ανακτήθηκε 5 Ιουνίου 2018, από, https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

(2018). Python programming language. Ανακτήθηκε 5 Ιουνίου 2018, από, https://www.python.org/

(2018). NumPy. Ανακτήθηκε 5 Ιουνίου 2018, από, http://www.numpy.org/

(2018). Natural Language Toolkit. Ανακτήθηκε 5 Ιουνίου 2018, από, https://www.nltk.org/

(2018). Gensim topic modelling for humans. Ανακτήθηκε 5 Ιουνίου 2018, από, https://radimrehurek.com/gensim/

(2018). Scikit-learn Machine Learning in Python. Ανακτήθηκε 5 Ιουνίου 2018, από, http://scikit-learn.org/stable/

(2018). Pandas - Python Data Analysis Library. Ανακτήθηκε 5 Ιουνίου 2018, από, https://pandas.pydata.org/

(2018). Matplotlib - Python 2D Plotting library. Ανακτήθηκε 5 Ιουνίου 2018, από, https://matplotlib.org/