Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής &
Υπολογιστών

Ανίχνευση Προπαγανδιστικών και Εξτρεμιστικών Απόψεων στα Κοινωνικά Δίκτυα με Μηχανική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΣΤΑΣΙΟΣ ΑΡΣΕΝΟΣ

Επιβλέπων : Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής & Υπολογιστών

Ανίχνευση Προπαγανδιστικών και Εξτρεμιστικών Απόψεων στα Κοινωνικά Δίκτυα με Μηχανική Μάθηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΣΤΑΣΙΟΣ ΑΡΣΕΝΟΣ

Επιβλέπων : Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10η Σεπτεμβρίου 2020.

.....	.....	.....
Ανδρέας Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αν. Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2020

Αναστάσιος Αρσένος

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

Copyright © Αναστάσιος Αρσένος, 2020.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η Ανίχνευση και Κατηγοριοποίηση προπαγανδιστικών σχολίων στα κοινωνικά δίκτυα με χρήση μοντέλων βαθιάς και μηχανικής μάθησης. Το πρόβλημα της ανίχνευσης και κατηγοριοποίησης προπαγανδιστικών μεθόδων αποτελεί ένα ανοιχτό πρόβλημα στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας λόγω της εκτεταμένης χρήσης τους στα κοινωνικά δίκτυα. Συνεπώς, υπάρχει μεγάλη ανάγκη για την αντιμετώπισή του με αυτοματοποιημένο τρόπο. Το διεθνές συνέδριο Workshop for Semantic Evaluation 2020 (SemEval) εισήγαγε αυτό το πρόβλημα σε διαγωνισμό στον οποίο συμμετείχαμε, προτείνοντας τις μεθόδους που αναλύουμε στην παρούσα εργασία, λαμβάνοντας αξιόλογα αποτελέσματα.

Το πρόβλημα χωρίζεται σε δύο επιμέρους υπο-προβλήματα, την Ανίχνευση των συγκεκριμένων κομματιών του κειμένου που εμφανίζονται προπαγανδιστικές τεχνικές και στην Κατηγοριοποίηση αυτών των αποσπασμάτων σε μία ή περισσότερες από τις 14 κύριες προπαγανδιστικές τεχνικές. Στα πλαίσια της διπλωματικής εργασίας αυτά τα προβλήματα αντιμετωπίζονται με προηγμένες τεχνικές βαθιάς νευρωνικής μάθησης αλλά και με κλασικές τεχνικές μηχανικής μάθησης όπως η Λογιστική Παλινδρόμηση.

Αρχικά, διερευνώνται μέσω εκτεταμένης βιβλιογραφικής έρευνας επιστημονικές προσεγγίσεις στον τομέα της ανίχνευσης προπαγανδιστικών, ρατσιστικών και εξτρεμιστικών σχολίων και ψευδών ειδήσεων. Στην βιβλιογραφική αυτή ανασκόπηση δίνεται ιδιαίτερη έμφαση τόσο στην μέθοδο εξαγωγής χαρακτηριστικών όσο και στους αλγόριθμους μηχανικής μάθησης που χρησιμοποιούνται. Στη συνέχεια περιγράφονται συνοπτικά οι βασικές θεωρητικές αρχές πάνω στις οποίες βασίζονται οι προσεγγίσεις με αξιοσημείωτα αποτελέσματα στην κάθε κατηγορία.

Στη συνέχεια, μετά την περιγραφή του γενικού πλαισίου και της ιστορίας, μεταβαίνουμε στην παρουσίαση της περίπτωσης του προβλήματος υπό εξέταση. Ακολουθεί η περιγραφή των δεδομένων, η ανάλυση των βασικών χαρακτηριστικών τους και η επεξεργασία που απαιτήθηκε για να είναι δυνατή η χρήση τους από τα μοντέλα βαθιάς και μηχανικής μάθησης.

Μέχρι στιγμής, το γενικότερο πρόβλημα ανίχνευσης κακόβουλων σχολίων είτε αυτά ήταν προπαγανδιστικού, ρατσιστικού ή εξτρεμιστικού περιεχομένου αντιμετωπιζόταν σε επίπεδο κειμένου. Πιο αναλυτικά, οι αλγόριθμοι που χρησιμοποιούνται για την αυτόματη ανίχνευση τέτοιων τεχνικών ταξινομούν ολόκληρο το κείμενο αν περιέχει σημαντικό ποσοστό αυτών των τεχνικών και δεν εντοπίζει τα σημεία ακριβώς του κείμενου στα οποία βρίσκονται.

Σε αντίθεση με τη συνήθη αντιμετώπιση, στην παρούσα εργασία στόχος είναι αφενός η στοχευμένη ανίχνευση των προπαγανδιστικών τεχνικών και αφετέρου η κατηγοριοποίησή τους σε μία ή περισσότερες από τις 14 κατηγορίες προπαγάνδας. Για το σκοπό αυτό δημιουργήθηκαν σταδιακά τρεις αρχιτεκτονικές για το κάθε υποπρόβλημα (της Ανίχνευσης και της Κατηγοριοποίησης) τα οποία επιτυγχάνουν αξιοσημείωτα αποτελέσματα στο προκλητικό πεδίο της αναγνώρισης προπαγάνδας

Λέξεις κλειδιά

Αναγνώριση προπαγάνδας, μηχανική μάθηση, λογιστική παλινδρόμηση, ταξινόμηση πολλών κλάσεων, σύνολο δεδομένων, επεξεργασία φυσικής γλώσσας, βαθιά μάθηση, τεχνητά νευρωνικά δίκτυα

Abstract

The subject of the diploma thesis is Propaganda Detection and Categorisation utilising deep learning techniques. Identification and classification of propaganda techniques is an open problem in the broader field of *Natural Language Processing* (NLP) that is heavily addressed especially in the past few years because of the use of these techniques in social media. There is more than ever the need to automatically detect malicious messages like propaganda, misinformation and fake news. Our task is part of the Workshop of Semantic Evaluation 2020 and is divided in Subtask A : Span Identification (SI) and B: Technique Classification (TC). Given a plain-text document, identify those specific fragments which contain at least one propaganda technique. This is a binary sequence tagging task and we refer to as SI. In addition, given a text fragment identified as propaganda and its document context, identify the applied propaganda technique in the fragment. Since there are overlapping spans, formally this is a a multilabel multiclass classification problem. However, whenever a span is associated with multiple techniques, the input file will have multiple copies of such fragments, so the problem can be algorithmically treated as a multiclass classification problem, which we refer to as TC. In this thesis these problems are dealt with advanced deep and machine learning techniques (BiLSTM, logistic regression, Elmo).

Initially, the state-of-the-art technologies in the field of propaganda, offensive language, hyperpartisanship are explored through extensive bibliographical research, and the most important approaches are documented. The emphasis of the review is placed on both on the feature extraction methodologies and the machine learning algorithms being utilised. In the next chapter, the basic theoretical principles and techniques on which the proposed approach is based are presented.

Following the general description we specifically and in detail present the task's dataset, the features that we created, the evaluation metrics and our approaches' architectures for both subtasks.

Until recently, the problem of detecting malicious content (propaganda, extremism, racism) was addressed mainly in the article level. More specifically, the algorithms proposed on many previous tasks aimed to identify if the whole article or post on social media included a high number of these techniques.

In contrast with the usual approaches, our task's goal is to identify specific text fragments and classify them in an automated way. For this purpose we introduce three novel architectures for each subtask.

Key words

Deep learning, machine learning, propaganda detection, natural language processing, Logistic Regression, Recurrent Neural Networks, LSTM, ELMo

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές ειλικρινείς ευχαριστίες μου στον επιβλέποντα της πτυχιακής μου, Καθηγητή κ. Ανδρέα Σταφυλοπάτη για την εμπιστοσύνη που μου έδειξε από τα χρόνια των προπτυχιακών μου σπουδών και τον πολύτιμο χρόνο που μου αφιέρωσε στην καθοδήγηση μου κατά τη διάρκεια εκπόνησης της πτυχιακής μου, τόσο σε επιστημονικά θέματα όσο και σε θέματα χαρακτήρα και ήθους. Θα ήθελα επίσης να τον ευχαριστήσω γιατί μου έδωσε την δυνατότητα να συμμετάσχω στον διαγωνισμό SemEval 2020.

Κατά την εκπόνηση της πτυχιακής μου εργασίας καθοριστική ήταν η συμβολή του συνεπιβλέποντα κ. Γιεώργιου Σιόλα τόσο σε επίπεδο συμβουλών και καθοδήγησης στα πρώτα βήματα όσο και σε επίπεδο ερευνητικής συνεργασίας. Πολύτιμη επίσης ήταν η ηθική συμπαράσταση και η ενθάρρυνσή του καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου.

Τέλος, θα ήθελα να εκφράσω τη βαθιά μου ευγνωμοσύνη στα μέλη της οικογένειας μου και ειδικότερα στη μητέρα Ζωή για την αγάπη, τη συμπαράσταση και την ανεξάντλητη υπομονή της.

Αναστάσιος Αρσένος,

Αθήνα, 10η Σεπτεμβρίου 2020

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 9

Περιεχόμενα 11

Κατάλογος πινάκων 13

Κατάλογος σχημάτων 15

1. Εισαγωγή 17
Εισαγωγή 17
1.1 Το πρόβλημα της Ανίχνευσης Προπαγάνδας 17
1.2 Αντικείμενο της διπλωματικής 17
1.3 Κίνητρο 18
1.4 Γενική Προσέγγιση του Προβλήματος 19
1.5 Δομή Διπλωματικής Εργασίας 20

2. Συγγενείς Επιστημονικές Εργασίες 21
Συγγενείς Επιστημονικές Εργασίες 21
2.1 Αναγνώριση Προπαγάνδας 21
2.2 Αναγνώριση Εξτρεμιστικών Σχολίων 22
2.3 Αναγνώριση Ρατσιστικών και Σεξιστικών Σχολίων 23

3. Θεωρητικό Υπόβαθρο 25
Θεωρητικό Υπόβαθρο 25
3.1 Μηχανική Μάθηση 25
3.1.1 Ορισμός Μηχανικής Μάθησης 25
3.1.2 Βασικά είδη μάθησης και Χαρακτηριστικοί Αλγόριθμοι 25
3.1.3 Ορισμός Βαθιάς Μάθησης 26
3.1.4 Νευρωνικά Δίκτυα 27
3.1.5 Συναρτήσεις Ενεργοποίησης 28
3.1.6 Πολυεπίπεδα νευρωνικά Δίκτυα 29
3.1.7 Συναρτήσεις Κόστους (Cost functions) 30
3.1.8 Αλγόριθμος Κατάβασης Κλίσης (Gradient Descent Algorithm) 31
3.2 Επαναλαμβανόμενα νευρωνικά Δίκτυα (Recurrent Neural Networks) 32
3.2.1 Περιληπτική Αναφορά στα Δίκτυα Εμπρόσθιας τροφοδότησης (Feedforward Neural Nets)) 33
3.2.2 Οπίσθια Διάδοση στο Χρόνο 35
3.2.3 Το πρόβλημα του (Vanishing (and Explonding) Gradients) 35
3.3 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long short term Memory Networks) . 36

3.3.1 Bidirectional LSTMs . . . 38
3.4 Επεξεργασία κείμενου . . . 39
3.4.1 Σύνολα από λέξεις (Bag of words) . . . 39
3.4.2 Bag of n-grams . . . 40
3.4.3 Term Frequency-Inverse Term Frequency (TF-IDF) . . . 40
3.4.4 Word2vec . . . 40
3.4.5 Λειτουργία και μηχανισμοί του αλγορίθμου Word2Vec . . . 41
3.4.6 Μέθοδος Δημιουργίας Διανυσμάτων Λέξεων GloVe (Global Vectors) . . . 42
3.4.7 Διανύσματα λέξεων από Γλωσσικά Μοντέλα (ELMo: Embeddings from Language Models) . . . 44

4. Αρχιτεκτονικές Μοντέλων Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας . . . 49
Αρχιτεκτονικές Μοντέλων Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας . . . 49
4.1 Ανίχνευση Προπαγανδιστικών Κομματιών . . . 49
4.1.1 Λογιστική Παλινδρόμηση . . . 49
4.1.2 Αμφίδρομο LSTM . . . 49
4.1.3 Αμφίδρομο residual LSTM . . . 50
4.2 Κατηγοριοποίηση προπαγανδιστικών τεχνικών . . . 51
4.3 Παρατηρήσεις ως προς την Υλοποίηση . . . 52

5. Επισκόπηση του Συστήματος Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας . 53
Επισκόπηση του Συστήματος Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας . . . 53
5.1 Δεδομένα . . . 53
5.2 Μετρικές Αξιολόγησης . . . 55
5.2.1 Ανίχνευση Προπαγανδιστικών Τμημάτων (SI task) . . . 55
5.2.2 Κατηγοριοποίηση Τεχνικών (TC task) . . . 56
5.3 Προ-επεξεργασία Δεδομένων . . . 57
5.4 Δημιουργία Χαρακτηριστικών . . . 58
5.5 Αξιολόγηση Αλγορίθμων . . . 60
5.5.1 Ανίχνευση Προπαγάνδας . . . 60
5.5.2 Κατηγοριοποίηση Τεχνικών . . . 61

6. Συμπεράσματα . . . 63
Συμπεράσματα . . . 63
6.1 Συμπεράσματα . . . 63
6.2 Προτάσεις για Μελλοντική Έρευνα . . . 64

Κατάλογος πινάκων

3.1 Αποτελέσματα χρήσης του μοντέλου Elmo σε διαφορετικές κατηγορίες προβλημάτων στον τομέα Επεξεργασίας Φυσικής Γλώσσας 47

5.1 Στατιστικά στοιχεία που αφορούν τις προπαγανδιστικές τεχνικές και τα αποτελέσματα στα σύνολα test και dev για το πρόβλημα Κατηγοριοποίησης 55
5.2 Αποτελέσματα για το πρόβλημα της Ανίχνευσης στο σύνολο δεδομένων του dev φακέλου 57
5.3 Αποτελέσματα για το πρόβλημα της κατηγοριοποίησης τεχνικών στα δεδομένα του φακέλου dev 58
5.4 Τελικά αποτελέσματα για το πρόβλημα της Ανίχνευσης σε σύγκριση με την νικήτρια ομάδα του διαγωνισμού 60

Κατάλογος σχημάτων

3.1 Βαθιά Μάθηση σε σύγκριση με άλλες τεχνικές μάθησης . . . 27
3.2 Συστατικά μέρη ενός τεχνητού νευρώνα . . . 29
3.3 Σιγμοειδής συνάρτηση . . . 30
3.4 Πολυεπίπεδο Νευρωνικό Δίκτυο . . . 30
3.5 Αλγόριθμος πίσω Διάδοσης Σφάλματος (Backpropagation algorithm) . . . 32
3.6 Δίκτυα Εμπρόσθιας τροφοδότησης . . . 33
3.7 Δομή Επαναλαβανόμενου Νευρωνικού Δικτύου . . . 34
3.8 Πολλαπλή εφαρμογή της σιγμοειδούς συνάρτησης στο πρώτο επίπεδο ενός δικτύου . . . 36
3.9 Ροή πληροφορίας σε ένα κύτταρο μνήμης LSTM δικτύου . . . 37
3.10 Σύγκριση κυττάρων απλού επαναλαμβανόμενου δικτύου (αριστερά) και LSTM (δεξιά)) . . . 37
3.11 Αναπαράσταση των πυλών ενός δικτύου LSTM) . . . 38
3.12 Αμφίδρομο Επαναλαμβανόμενο Νευρωνικό Δίκτυο . . . 39
3.13 Μέθοδος Continuous Bag of Words (CBoW) . . . 41
3.14 Μέθοδος Skip-gram . . . 42
3.15 Παράδειγμα πίνακα συνεμφάνισης (co-occurance matrix . . . 43
3.16 Παράδειγμα που δείχνει τη μέθοδο εύρεσης της σημασιολογικής ομοιότητας . . . 43
3.17 Ένα αμφίδρομο LSTM δύο επιπέδων. Τα κόκκινα τετράγωνα αναπαριστούν το εμπρόσθιο επαναλαμβανόμενο τμήμα και τα γαλάζια το οπίσθιο επαναλαμβανόμενο τμήμα. . . . 45
3.18 Η προσθήκη της υπολλειπόμμενης (residual) σύνδεσης μεταξύ πρώτου και δεύτερου επιπέδου του LSTM. Η είσοδος του πρώτου επιπέδου προστίθεται απευθείας στην έξοδό του πριν περαστεί σαν είσοδος στο δεύτερο επίπεδο. . . . 46
3.19 Μετασχηματισμός που εφαρμόζεται σε κάθε λέξη πριν αυτή διοχετευθεί σαν είσοδος στο πρώτο επίπεδο LSTM. Εδώ φαίνεται ενδεικτικά η διαδικασία για τη λέξη the. . . . 47
3.20 Ένα παράδειγμα συνδυασμού των αναπαραστασέων των κρυφών επιπέδων και της λεξικής αναπαράστασης έτσι ώστε η λέξη "happy" να αποκτήσει την αναπαράστασή της στο μοντέλο ELMo. . . . 48

4.1 Σχηματική αναπαράσταση του τελικού μοντέλου για το υπο-πρόβλημα της Ανίχνευσης που περιλαμβάνει αμφίδρομο residual LSTM και το μοντέλο ELMo . . . 50
4.2 Σχηματική αναπαράσταση του μοντέλου Κατηγοριοποίησης με χρήση αμφίδρομου LSTM και διανύσματα λέξεων από το μοντέλο GloVe . . . 51

5.1 Παράδειγμα επισημειώσεων (άνω μέρος) και των προβλέψεων ενός μοντέλου μάθησης με επίβλεψη (κάτω μέρος) σε ένα κείμενο που αναπαριστάται ως μια ακολουθία χαρακτήρων. . . . 57
5.2 Παράδειγμα επισημειώσεων . . . 58
5.3 Κατανομή μήκους των προτάσεων των κειμένων . . . 59

Acronyms

NLP Natural Language Processing

NER Name Entity Recognition

SI Span Identification

TC Technique Classification

SLC Sentence Level Classification

FLC Fragment Level Classification

OLID Offensive Language Identification Dataset

RNN Recurrent Neuaral Network

LSMN Long Short Memory Network

FNN Feedforward Neural Net

LSTM Long Short Term Memory

TF-IDF Term Frequency-Invers Term Frequency

ELMo Embedding from Language Models

ΠΤΔ Πρόβλημα Ταυτοποίησης Διαστήματος

MLP Multilayer Perceptron

Κεφάλαιο 1

Εισαγωγή

1.1 Το πρόβλημα της Ανίχνευσης Προπαγάνδας

Οι προπαγανδιστικές καμπάνιες έχουν ως στόχο να επηρεάσουν την γνώμη των ανθρώπων με σκοπό να προωθήσουν μια συγκεκριμένη ατζέντα. Εκμεταλλευόμενοι την ανωνυμία που παρέχει το διαδίκτυο, την εύκολη δημιουργία ψεύτικων λογαριασμών στα κοινωνικά δίκτυα και την ευκολία στην αυτοματοποιημένη οργάνωση και διαχείριση μεγάλου όγκου τέτοιων ψεύτικων λογαριασμών είναι πολύ εύκολο κάποιοι, να στείλουν προπαγανδιστικά μηνύματα σε εκατομμύρια χρήστες. Συνέπεια αυτής της διαδικασίας είναι να επηρεάζεται μεγάλο κομμάτι του πληθυσμού χωρίς να το συνειδητοποιεί. Είναι προφανές λοιπόν ότι χρειάζεται ένας αποτελεσματικός τρόπος ανίχνευσης της προπαγάνδας προκειμένου να αποφευχθούν φαινόμενα παραπλάνησης της κοινής γνώμης.

1.2 Αντικείμενο της διπλωματικής

Στην παρούσα διπλωματική διατυπώνουμε το πρόβλημα της ανίχνευσης προπαγάνδας. Το πρόβλημα αυτό αποτελεί ανοικτό πρόβλημα στο ευρύτερο ερευνητικό πεδίο της ανίχνευσης κακόβουλων σχολίων και ψευδών ειδήσεων. Με τον όρο προπαγάνδα αναφερόμαστε στην έκφραση μιας γνώμης ή μιας πράξης από μεμονωμένα άτομα ή από ομάδες ατόμων τα οποία εσκεμμένα προσπαθούν να επηρεάσουν την γνώμη άλλων ανθρώπων χρησιμοποιώντας αθέμιτες τεχνικές Miller, C. R (1937-1938). Συχνά συνδυάζεται με παραπληροφόρηση και ψευδείς ειδήσεις. Οι κύριες μεθοδολογίες για την αντιμετώπιση του συγκεκριμένου προβλήματος εστιάζονται στη εύρεση λέξεων που έχουν μεγάλη πιθανότητα προπαγανδιστικής χρήσης όπως οι άσεμνοι χαρακτηρισμοί προσώπων, λέξεις που δηλώνουν επίκληση στο φόβο και η συχνή χρήση υπερθετικού βαθμού που δείχνει υπερβολή στο λόγο. Έτσι εντοπίζοντας έναν αρκετά μεγάλο αριθμό τέτοιων λέξεων μέσω αλγορίθμων βαθιάς και μηχανικής μάθησης γίνεται προσπάθεια ταξινόμησης του κειμένου σε προπαγανδιστικό ή μη ανάλογα με το ποσοστό τέτοιων λέξεων που βρεθήκανε συνολικά στο κείμενο. Τα τελευταία χρόνια γίνεται προσπάθεια ανίχνευσης κακόβουλων τεχνικών όπως η προπαγάνδα όχι σε επίπεδο κειμένου αλλά σε επίπεδο συγκεκριμένων σημείων αυτού.

Έτσι στο διαγωνισμό του SemEval 2020 (Semantic Evaluation) Da San Martino et al. (2020) προτάθηκε η εύρεση και η κατηγοριοποίηση προπαγανδιστικών σχολίων σε ειδησεογραφικά άρθρα. Η διαδικασία χωρίστηκε σε δύο ανεξάρτητα υπο-προβλήματα, το πρόβλημα της Ανίχνευσης και το πρόβλημα της Κατηγοριοποίησης της προπαγάνδας. Η μετρική αξιολόγησης για το πρόβλημα της Ανίχνευσης προπαγάνδας (Span identification (SI)) είναι μια παραλλαγή της μετρικής F1 κατάλληλη για να λάβει υπόψιν μερική επικάλυψη μεταξύ των προπαγανδιστικών σχολίων (spans).

Αναλυτική περιγραφή αυτής της μετρικής γίνεται στο κεφάλαιο 5. Το πρόβλημα της Κατηγοριοποίησης Τεχνικών (Technique Classification (TC)) αποτελείται από κομμάτια κειμένου τα οποία ανήκουν σε μία από τις 14 μορφές (κλάσεις) προπαγάνδας όπως Αναφορά σε Όνόματα (Name Calling) και Φορτισμένη Γλώσσα (Loaded Language). Ο στόχος είναι η ταξινόμηση αυτών των κομματιών στις σωστές κλάσεις, αυτοματοποιημένα με αλγορίθμους μηχανικής και βαθιάς μάθησης. Για το σκοπό αυτό δόθηκε από τους διοργανωτές ένα σύνολο 550 ειδησεογραφικών άρθρων τα οποία έχουν επεξεργαστεί και εντοπιστεί τα σημεία που υπάρχουν προπαγανδιστικά σχόλια αλλά και η κατηγορία αυτών. Αυτή η διαδικασία ανατέθηκε στην εταιρία DataPro η οποία ειδικεύεται στη δημιουργία συνόλου δεδομένων με τα οποία μπορούν να εκπαιδευτούν αλγόριθμοι βαθιάς και μηχανικής μάθησης.

Στην παρούσα εργασία, προτείνουμε δύο καινοτόμες αρχιτεκτονικές βαθιάς μάθησης για την αντιμετώπιση των δύο αυτών προβλημάτων του διαγωνισμού Arsenos and Siolas (2020). Επιπρόσθετα, συγκρίνουμε τις μεθόδους μας με κλασσικές μεθόδους μηχανικής μάθησης όπως η Λογιστική Παλινδρόμηση στις οποίες παρέχονται διανύσματα λέξεων προ-εκπαιδευμένα σε τεράστιου όγκου εξωτερικά δεδομένα. Η μάθηση με μεταφορά γνώσης (Transfer Learning) έχει δώσει υποσχόμενα αποτελέσματα σε πολλαπλά προβλήματα του τομέα της Επεξεργασίας Φυσικής Γλώσσας, κάτι που αποτέλεσε παρακίνηση και αισθητήριο στο να συμπεριλάβουμε αυτού του είδους την τεχνική μάθησης και σε αυτό το απαιτητικό πρόβλημα.

1.3 Κίνητρο

Το Διαδίκτυο προσφέρει σε όλους τη δυνατότητα να δημιουργήσουν μια σελίδα ή ένα blog το οποίο θα λειτουργήσει ως μέσο διακίνησης ειδήσεων. Αναμφίβολα, αυτό είναι μια ιδιαίτερα θετική εξέλιξη αφού ενισχύει την ελευθερία του λόγου δίνοντας την ευκαιρία σε όλους να κάνουν τη φωνή τους να ακουστεί. Με την άνοδο της δημοτικότητας των κοινωνικών δικτύων όπως του Twitter και του Facebook ο καθένας μπορεί να έχει πρόσβαση σε ένα πολύ μεγάλο κοινό για να μεταλαμπαδεύσει τις απόψεις του, κάτι που μέχρι πολύ πρόσφατα ήταν προσβάσιμο μόνο από ειδησεογραφικούς κολοσσούς και τηλεοπτικά κανάλια.

Παρόλα αυτά, αυτή η εξέλιξη της άμεσης έκφρασης έφερε και μη αναμενόμενες συνέπειες. Η πιο χαρακτηριστική είναι ότι άφησε την κοινωνία απροστάτευτη απέναντι σε πιθανή προσπάθεια χειραγώγησης από πολλαπλές πηγές. Σε αυτό το φαινόμενο δόθηκε ιδιαίτερη προσοχή από την κοινή γνώμη το 2016, μια χρονιά που στιγματίστηκε από παραπληροφόρηση και ψευδείς διαδικτυακές ειδήσεις σε πρωτόγνωρο βαθμό, κυρίως σε συσχέτιση με το δημοψήφισμα για το Brexit και τις αμερικανικές προεδρικές εκλογές. Έπειτα, το 2020 η πανδημία του COVID-19 έδωσε την ευκαιρία σε εκτεταμένη διάδοση παραπληροφόρησης κάτι που έγινε γνωστό με τον όρο infodemic σε διεθνές επίπεδο. Η μετάδοση παραπληροφόρησης μεταμφιεσμένης σε ειδήσεις δημιουργεί την ψευδαίσθηση στους αναγνώστες ότι πρόκειται για αξιόπιστες πληροφορίες που πρέπει να τις λάβουν σοβαρά υπόψιν και μειώνει παράλληλα τα αντανακλαστικά τους για κριτική σκέψη από ότι αν αυτές οι πληροφορίες προέρχονταν από άλλες πηγές. Παρόλο που οι ψευδείς δηλώσεις δεν είναι καινούργιο φαινόμενο, (ο κίτρινος τύπος υπάρχει εδώ και δεκαετίες) αλλά αυτή τη φορά τα πράγματα είναι αρκετά διαφορετικά σε ότι αφορά την κλίμακα και την αποτελεσματικότητα χάρις στα κοινωνικά δίκτυα, τα οποία παρέχουν και το μέσο πρόσβασης σε εκατομμύρια χρήστες αλλά και ένα εύκολο τρόπο να στοχεύσουν συγκεκριμένες ομάδες χρηστών και ψηφοφόρων βάση

γεωγραφικών, δημογραφικών, ψυχολογικών και πολιτικών δεδομένων που υπάρχουν στους λογαριασμούς.

Μια σημαντική πτυχή του προβλήματος που συχνά αγνοείται είναι ο μηχανισμός μέσω του οποίου προσπαθούν να πείσουν τους χρήστες με παραπληροφόρηση, που είναι οι προπαγανδιστικές τεχνικές (*propaganda techniques*). Αυτές συμπεριλαμβάνουν συγκεκριμένες θεωρητικές και ψυχολογικές τεχνικές όπως *loaded language*, *flag waving*, *appeal to authority*, *slogans*, και *cliches*, λογικές πλάνες όπως *straw men* (η λανθασμένη παρουσίαση της γνώμης κάποιου), *red herring* (παρουσίαση άσχετων δεδομένων), *black and white fallacy* (παρουσίαση δυο εκδοχών ως τις μοναδικές επιλογές) και *whataboutism*. Επιπλέον, το πρόβλημα περιπλέκεται λαμβάνοντας υπόψιν το γεγονός ότι η προπαγάνδα δεν χρειάζεται κατ' ανάγκη να περιέχει ψέματα. Μπορεί απλά να κάνει επίκληση στο συναίσθημα του χρήστη και να παρουσιάζει τη μισή αλήθεια. Για αυτό πιστεύουμε ότι η έρευνα που κάναμε και το μοντέλα που δημιουργήσαμε συνεισφέρουν στη μάχη ενάντια στην διαδικτυακή προπαγάνδα.

Η προπαγάνδα είναι κάτι διαφορετικό από την παραπληροφόρηση. Πιο αναλυτικά, σε ότι αφορά την φερεγγυότητα και το στόχο της διαχειριζόμενης πληροφορίας η παραπληροφόρηση χρησιμοποιεί ψευδή και με πρόθεση να βλάψουν στοιχεία. Αυτή η πρόθεση να βλάψει έγινε κυρίως αισθητή κατά τη διάρκεια του Brexit και των Αμερικανικών εκλογών. Αντιθέτως, από θεωρητικής άποψης η προπαγάνδα μπορεί να βασίζεται σε ψευδή αλλά και αληθή στοιχεία και οι προθέσεις μπορεί να είναι είτε βλαπτικές είτε όχι. Στην πράξη βέβαια, η προπαγάνδα και η παραπληροφόρηση χρησιμοποιούνται συνεργατικά για να πετύχουν συγκεκριμένους στόχους, μετατρέποντας τα κοινωνικά δίκτυα σε ένα όπλο. Μια ακόμη συγγενής έννοια είναι αυτή των ψευδών ειδήσεων ("fake news") η οποία εστιάζει στην δημιουργία πληροφοριών που δεν αντικατοπτρίζουν την πραγματικότητα. Σε αυτή την έννοια έχει δοθεί ο όρος μαύρη προπαγάνδα από διάφορους συγγραφείς.

1.4 Γενική Προσέγγιση του Προβλήματος

Όπως ορίστηκε παραπάνω το πρόβλημα της Ανίχνευσης Προπαγάνδας έχει απασχολήσει σε μεγάλο βαθμό την επιστημονική έρευνα. Λαμβάνοντας υπόψιν ότι πρόκειται για μια καινοτόμα προσπάθεια ανίχνευσης συγκεκριμένων κομματιών του κειμένου η προσέγγισή μας διαμορφώθηκε έπειτα από συγκεκριμένες παραδοχές και πειράματα.

Πιο αναλυτικά, οι διοργανωτές μας παρείχαν ένα σύνολο δεδομένων από ειδησεογραφικά άρθρα χωρισμένα σε τρεις φακέλους. Το μεγαλύτερο μέρος των κειμένων βρίσκονταν στον φάκελο εκπαίδευσης (train folder) με τον οποίο θα γινόταν η εκπαίδευση των αλγορίθμων βαθιάς και μηχανικής μάθησης. Επιπλέον, δόθηκαν οι φάκελοι development και test οι οποίοι περιείχαν τα άρθρα για την αξιολόγηση των αλγορίθμων. Οι επισημειώσεις (annotations) των προπαγανδιστικών κομματιών είχαν δοθεί σε επίπεδο χαρακτήρων, δηλαδή κάθε χαρακτήρας του κειμένου είχε ένα δείκτη (offset). Ο αρχικός χαρακτήρας της πρώτης λέξης του κειμένου είχε δείκτη 0 και οι επόμενοι $+1$ διαδοχικά. Εμείς αποφασίσαμε να μετατρέψουμε τις επισημειώσεις από επίπεδων χαρακτήρων σε επίπεδο λέξεων αφού όταν ο πρώτος χαρακτήρας μιας λέξης είχε επισημειωθεί ως προπαγάνδα τότε και όλη η λέξη θα είχε επισημειωθεί αναλόγως. Έτσι αφαιρέσαμε όλα τα σημεία στίξης που δεν προσδίδουν χρήσιμες πληροφορίες και αντιστοιχίσαμε τις επισημειώσεις σε επίπεδο λέξεων (token-level annotations). Παρόλο που υπήρχε μια μικρή απώλεια πληροφορίας με την έννοια ότι μερικοί δείκτες είχαν μετατοπιστεί σε ένα εύρος ± 3 αυτό δεν επηρέασε την εκπαίδευση των αλ-

γορίθμων σημαντικά. Κύρια παρότρυνση για αυτό τον μετασχηματισμό ήταν η προσαρμογή του προβλήματος Ανίχνευσης στη μορφή προβλημάτων Ανίχνευσης Οντοτήτων με Ονόματα (Named Entity Recognition (NER)).

Η κατηγορία προβλημάτων NER έχει ως στόχο να εξάγει πληροφορίες εντοπίζοντας και ταξινομώντας οντότητες από μη δομημένα κείμενα σε προκαθορισμένες κλάσεις όπως ονόματα ανθρώπων, τοποθεσίες, χρονικές εκφράσεις και άλλα. Στην δική μας περίπτωση αυτές οι οντότητες είναι οι κλάσεις Προπαγάνδα και μη-Προπαγάνδα στις οποίες θα ταξινομηθούν οι λέξεις του κάθε άρθρου. Εκτός από αυτό το μετασχηματισμό ιδιαίτερη έμφαση δόθηκε στη δημιουργία των χαρακτηριστικών μέσω προ-εκπαιδευμένων διανυσμάτων λέξεων αλλά και η βέλτιστη χρήση αυτών. Τέλος, η αρχιτεκτονική του μοντέλου, δηλαδή η επιλογή και η δομή στοιχείων όπως τα αμφίδρομα επαναλαμβανόμενα δίκτυα, λειτούργησε καθοριστικά στη διαμόρφωση ενός αξιόλογου αποτελέσματος.

1.5 Δομή Διπλωματικής Εργασίας

Στο κεφάλαιο 2 καταγράφουμε το σύνολο των συγγενών εργασιών, χωρισμένων σε τρεις κατηγορίες, την Αναγνώριση Προπαγάνδας Εξτρεμιστικών και Ρατσιστικών-Σεξιστικών σχολίων.

Στο Κεφάλαιο 3, περιγράφεται το τεχνικό υπόβαθρο, με τις αντίστοιχες παραπομπές στα πρωτότυπα κείμενα, που απαιτείται για την πλήρη κατανόηση της συγκεκριμένης εργασίας. Στο κεφάλαιο αυτό, λοιπόν, δίνεται μια εκτενέστερη αναφορά στα εργαλεία που χρησιμοποιήθηκαν στα τελικά μοντέλα και είχαν τη μεγαλύτερη συμβολή στα τελικά αποτελέσματα όπως τα αμφίδρομα επαναλαμβανόμενα δίκτυα (BiLSTMs) και τα διανύσματα λέξεων από το μοντέλο ELMo.

Στο κεφάλαιο 4, αποτυπώνεται η περιγραφή των αρχιτεκτονικών που δημιουργήθηκαν στα πλαίσια αυτής της εργασίας και τα στάδια εξέλιξης του μοντέλου, από την πρώτη πρόχειρη δοκιμαστική μορφή του μέχρι την τελική μορφή της αρχιτεκτονικής του κάθε υπο-προβλήματος.

Στο κεφάλαιο-5 δίνεται αρχικά, η περιγραφή του τρόπου απόκτησης των δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση της επιλεγμένης τεχνικής. Στη συνέχεια δίνονται αναλυτικά τα χαρακτηριστικά και οι παράμετροι που χρησιμοποιήθηκαν καθώς και η παρουσίαση των μετρικών αξιολόγησης. Ακολουθεί η περιγραφή των δεδομένων που χρησιμοποιήθηκαν.

Τέλος, στο κεφάλαιο-6 καταγράφουμε τα βασικά συμπεράσματα της εργασίας καθώς και τα μελλοντικά μας ερευνητικά σχέδια στο θέμα της αναγνώρισης προπαγάνδας.

Κεφάλαιο 2

Συγγενείς Επιστημονικές Εργασίες

Στο παρόν κεφάλαιο παρουσιάζονται παρόμοιες και συγγενείς εργασίες στον τομέα της Αναγνώρισης Προπαγάνδας, Ψευδών Ειδήσεων, Εξτρεμιστικών και Ρατσιστικών σχολίων. Ιδιαίτερη έμφαση έχει δοθεί τα τελευταία χρόνια στο διεθνές Συνέδριο Workshop of Semantic Evaluation στους συγκεκριμένους τομείς με διαγωνισμούς στους οποίους αξιολογούνται αλγόριθμοι τεχνητής νοημοσύνης στην αποτελεσματική αναγνώριση των παραπάνω κακόβουλων τεχνικών. Παρακάτω διατυπώνονται ενδεικτικά σχετικές εργασίες σε τρεις κύριες κατηγορίες.

2.1 Αναγνώριση Προπαγάνδας

Ένας πολύ παρόμοιος διαγωνισμός σε σχέση με αυτόν που πραγματεύεται η παρούσα εργασία έλαβε χώρα στο 9ο συνέδριο με τίτλο International Joint Conference on Natural Language Processing, EMNLP-IJCNLP Da San Martino et al. (2019e). Αυτός ο διαγωνισμός αποτελούνταν από δύο υπο-προβλήματα το Sentence Level Classification(SLC) και το Fragment Level Classification(FLC). Το SLC πρόβλημα προσπαθεί να προβλέψει αν σε μια πρόταση υπάρχει έστω και μια μορφή προπαγάνδας, ενώ το FLC έχει ως στόχο να βρει το συγκεκριμένο σημείο και να προσδιορίσει τη μορφή της προπαγάνδας ανάμεσα από τις 18 που υπάρχουν συνολικά. Τα δεδομένα που χρησιμοποιήθηκαν είναι ουσιαστικά ένα υποσύνολο των δεδομένων της παρούσας εργασίας και το FLC αποτελεί ένα συνδυασμό αναγνώρισης και κατηγοριοποίησης προπαγανδιστικών σχολίων. Κάτι τέτοιο βέβαια αποδείχθηκε ιδιαίτερα δύσκολο να επιτευχθεί και τα αποτελέσματα των διαγωνιζόμενων κυμαίνονταν μεταξύ 0.10 και 0.24 στη μετρική F1 score.

Πιο αναλυτικά, η νικήτρια ομάδα του διαγωνισμού στο FLC task Yoosuf and Yang (2019) χρησιμοποίησε ταξινόμηση σε επίπεδο λέξεων βασιζόμενη στη μέθοδο BERT όπου η κάθε λέξη μπορούσε να ανήκει σε μία από τις 18 κλάσεις σε καμία ή σε μια βοηθητική κλάση. Η ομάδα αυτή τροφοδοτούσε στο μοντέλο της μία πρόταση τη φορά για να μειώσει την ανάγκη για υπολογιστική ισχύ. Επιπλέον, προσπάθησαν να προσθέσουν δείγματα μέσω μεθόδων υπερδειγματοληψίας (oversampling) για να ισορροπήσουν τις κλάσεις του προβλήματος κάτι που φαίνεται να είχε καθοριστική σημασία στην απόδοση του αλγορίθμου.

Μία άλλη ομάδα, υψηλά στην κατάταξη, με το όνομα Stalin Ek and Ghanimifard (2019), χρησιμοποίησε προ εκπαιδευμένα και ευαίσθητα στα συμφραζόμενα, μοντέλα αναπαράστασης λέξεων όπως το ELMo και ο BERT τα οποία εκπαιδεύτηκαν με αναδρομικά δίκτυα τύπου LSTM. Ομοίως και αυτή η ομάδα προσπάθησε μέσω υπερδειγματοληψίας των κλάσεων μειοψηφίας να αντιμετωπίσει το πρόβλημα της ανισορροπίας των δειγμάτων των κλάσεων.

Ο διαγωνισμός του SemEval 2020 όπου συμμετείχαμε αποτελεί συνέχεια αυτού του διαγωνισμού με τη διαφορά ότι έγιναν κάποιες προσθήκες στο σύνολο των δεδομένων και διαχωρίστηκε το FLC σε δυο ανεξάρτητα υπο-προβληματα. Η ομάδα που διοργάνωσε και τους δύο αυτούς διαγωνισμούς έχει δημιουργήσει επίσης στο πλαίσιο του data analysis project το σύστημα PROPPY Da San Martino et al. (2019a) το οποίο οργανώνει τα άρθρα και καταδεικνύει το ποσοστό του προπαγανδιστικού τους περιεχομένου. Στόχος αυτού του συστήματος είναι να εφιστήσει την προσοχή στους αναγνώστες και παράλληλα παρέχει ένα εργαλείο για την οργάνωση της παρακολούθησης πολλών ειδησεογραφικών άρθρων.

Τα τελευταία χρόνια η προπαγάνδα είχε αντιμετωπιστεί κυρίως σε επίπεδο άρθρων. Ο Rashkin Rashkin et al. (2017) δημιούργησε ένα σύνολο δεδομένων από ειδησεογραφικά άρθρα με τις εξής ετικέτες προπαγάνδα, αξιόπιστο, σάτιρα, απάτη. Ο Da San Martino et al. (2019b) πειραματίστηκε με μια παραλλαγή αυτού του συνόλου δεδομένων, ομαδοποιώντας τις τρεις κλάσεις εκτός της προπαγάνδας σε μια. Έτσι δημιούργησε ένα πρόβλημα δυαδικής ταξινόμησης (προπαγάνδα -μη-προπαγανδα). Η ίδια ομάδα δημιούργησε εκ νέου ένα τεράστιο δείγμα προπαγανδιστικών κειμένων και πρότεινε μια μέθοδο στηριζόμενη στην ανάλυση χαρακτηριστικών για την ταξινόμηση αυτών. Κύριο μειονέκτημα και των τριών προαναφερθέντων προσεγγίσεων είναι ότι οι ετικέτες των κειμένων δόθηκαν με μέθοδο απομακρυσμένης επίβλεψης (distance supervision). Πιο συγκεκριμένα, τα άρθρα κατηγοριοποιήθηκαν σύμφωνα με την εικόνα που επικρατεί για το μέσο που τα δημοσίευσε, δηλαδή αν η εφημερίδα θεωρείται ότι έχει προπαγανδιστικές τάσεις τότε και όλα τα άρθρα της επισημειώνονται αυτόματα ως προπαγανδιστικά.

2.2 Αναγνώριση Εξτρεμιστικών Σχολίων

Στο συνέδριο SemEval 2019 έλαβε χώρα ο διαγωνισμός Hyperpartisan News Detection Kiesel et al. (2019) που στόχο είχε να αναγνωρίσει αν ένα κείμενο περιείχε ή όχι εξτρεμιστικά σχόλια. Σε αυτό το διαγωνισμό χρησιμοποιήθηκαν δυο σύνολα δεδομένων. Το ένα ήταν σχετικά μικρό ως προς τον αριθμό των άρθρων αλλά είχε δημιουργηθεί χειροκίνητα από δημοσιογράφους και οι ετικέτες που ανατέθηκαν στα κείμενα ως εκ τούτου ήταν κατά πάσα πιθανότητα ορθές. Παράλληλα, δόθηκε και ένα τεράστιο σύνολο δεδομένων το οποίο είχε ταξινομηθεί λαμβάνοντας υπόψιν μόνο την πηγή από την οποία τα άρθρα προέρχονταν, δηλαδή αν μια εφημερίδα έχει τη φήμη ότι προσπαθεί να διαδώσει εξτρεμιστικά μηνύματα τότε όλα τα άρθρα της, χωρίς περαιτέρω επεξεργασία ταξινομούνται ως εξτρεμιστικά. Το ενδιαφέρον σε αυτή την περίπτωση είναι ότι όλοι ανεξαιρέτως οι διαγωνιζόμενοι μετά από σειρά πειραμάτων αποφάνθηκαν ότι το τεράστιο σύνολο δεδομένων όχι μόνο δεν βοήθησε στο να εκπαιδευτούν καλύτερα οι αλγόριθμοι βαθιάς και μηχανικής μάθησης αλλά είχε μάλιστα αρνητικό αντίκτυπο στα αποτελέσματα. Αυτό καταδεικνύει την σημασία δημιουργίας ενός αξιόπιστου συνόλου δεδομένων για την επιτυχία και την εξαγωγή χρήσιμων συμπερασμάτων στους διαγωνισμούς τεχνητής νοημοσύνης.

Πολύ ενδιαφέρον παρουσιάζει το γεγονός ότι οι δυο πρώτες ομάδες του διαγωνισμού πρότειναν δυο εκ δια μέτρου αντίθετες προσεγγίσεις με πολύ κοντινά αποτελέσματα. Πιο αναλυτικά η ομάδα Bertha von Suttner Jiang et al. (2019) χρησιμοποίησε αναπαραστάσεις προτάσεων από το μέσο όρο των διανυσμάτων λέξεων της κάθε πρότασης. Τα διανύσματα των λέξεων προέρχονταν από το μοντέλο ELMo Peters et al. (2018). Το μοντέλο περιείχε επίσης συνελικτικά νευρωνικά δίκτυα και Κανονικοποίηση Τμημάτων (Batch Normalization). Κάθε άρθρο επεξεργάστηκε ως μια ακολουθία

από διανύσματα τέτοιων προτάσεων μέσα από μια αρχιτεκτονική νευρωνικού δικτύου που περιείχε 5 παράλληλα συνελικτικά επίπεδα με πυρήνες μεγέθους 2,3,4,5,6 και 512 χαρακτηριστικά, ακολουθούμενα κάθε ένα από μια μη-γραμμική συνάρτηση τύπου ReLU, μεθόδους κανονικοποίησης και max-pooling. Τα αποτελέσματα εισάγονταν έπειτα στο πλήρως συνδεδεμένο τελικό επίπεδο όπου γινόταν η δυαδική ταξινόμηση μέσω σιγμοειδούς συνάρτησης ενεργοποίησης.

Από την άλλη μεριά, η ομάδα Vernon-fenwick έδωσε έμφαση στη δημιουργία χαρακτηριστικών "με το χέρι" (handcrafted) αντί υλοποίησης μιας αρχιτεκτονικής με νευρωνικά δίκτυα. Πιο συγκεκριμένα το σύστημά τους αποτελούνταν από τρία κύρια μέρη την προ-επεξεργασία, την δημιουργία χαρακτηριστικών και τέλος την εισαγωγή αυτών σε ένα μοντέλο μηχανικής μάθησης για την τελική ταξινόμηση. Στο πρώτο στάδιο προσπάθησαν να απο-θορυβοποιήσουν τα άρθρα διαγράφοντας τις παραπομπές. Έπειτα, μέσω των εξής λεξικών: bias lexicon από το Neutral Point of View corpus της Wikipadia, MPQA Subjectivity lexicon, προσπάθησαν να ποσοτικοποιήσουν το βαθμό bias και το βαθμό πόλωσης αντίστοιχα μετρώντας τη συχνότητα με την οποία λέξεις του κάθε άρθρου περιέχονται σε αυτά τα λεξικά. Ακόμα, μέτρησαν τη συχνότητα εμφάνισης υπερθετικού βαθμού και των συγκρίσεων σε κάθε άρθρο μέσω της βιβλιοθήκης Part of Speech Tagger του NLTK. Τέλος η σημασιολογική αναπαράσταση των λέξεων και των προτάσεων έγινε με μοντέλα όπως το GloVe και Doc2Vec (Document to Vector) και του Universal Sentence Encoder. Σαν ταξινομητής επιλέχθηκε η μέθοδος της Λογιστικής Παλινδρόμησης.

Συμπερασματικά, παρόλο που και οι δυο μέθοδοι είχαν αξιοσημείωτα αποτελέσματα, η προσέγγιση με ανάλυση χαρακτηριστικών μέσω λεξικών περιορίζει τη χρήση του μοντέλου ακόμα και σε πολύ συγγενή προβλήματα. Αντιθέτως η χρήση του μοντέλου της ομάδας Bett-VonSUttner μπορεί να χρησιμοποιηθεί με πολύ μικρές αλλαγές σε άλλα παρόμοια προβλήματα και να βελτιστοποιηθεί περαιτέρω.

2.3 Αναγνώριση Ρατσιστικών και Σεξιστικών Σχολίων

Σε αυτό το σημείο θα παρουσιάσουμε τα ευρήματα του διαγωνισμού με τίτλο Ανίχνευση και Κατηγοριοποίηση Προσβλητικής Γλώσσας στα Κοινωνικά Δίκτυα (SemEval-2019 Task 6) Oberstrass et al. (2019). Για την προσέγγιση αυτού του προβλήματος δημιουργήθηκε ένα σύνολο δεδομένων με τον τίτλο Offensive Language Identification Dataset (OLID), το οποίο περιέχει πάνω από 14.000 tweets στην αγγλική γλώσσα. Το πρόβλημα χωρίστηκε σε τρία μέρη:

- Α: στην διάκριση αν είναι ή όχι προσβλητική η τοποθέτηση,
- Β: στον τύπο της προσβολής
- Γ: στην εύρεση του στόχο στον οποίο απευθυνόταν η προσβολή

Τα συστατικά μέρη των μοντέλων που προτάθηκαν ποικίλουν από κλασικές μεθόδους μηχανικής μάθησης όπως SVM και Λογιστική Παλινδρόμηση, σε βαθιά νευρωνικά δίκτυα όπως CNN, biLSTM, RNN συνδυάζοντάς τα με μηχανισμούς προσοχής (attention mechanisms). Ακόμα υπήρχαν μέθοδοι με καινοτόμα στοιχεία όπως η μέθοδος ELMo και BERT. Ξεκάθαρα υπερίσχυσαν σε δημοτικότητα και σε αποτελέσματα οι προσεγγίσεις βαθιάς μάθησης και στα τρία υπο-προβλήματα. Πιο αναλυτικά η ομάδα NLPR@SRPOL Seganti et al. (2019) τερμάτισε 8η, 9η, και 12η στα υπό προβλήματα

Α, Β, και Γ αντίστοιχα με χρήση μεθόδων όπως τα Τυχαία Δένδρα (Random Forest), μετασχηματιστές Transformer Universal encoder ELMo και συνδυασμό άλλων διανυσμάτων αναπαράστασης λέξεων όπως το fast-Text. Επίσης, εκπαίδευσαν το μοντέλο τους όχι μόνο στο δοσμένο σύνολο δεδομένων αλλά και σε άλλα παρόμοια σύνολα από προηγούμενες ερευνητικές προσπάθειες σε παρόμοια προβλήματα. Η ομάδα NULI Liu et al. (2019) η οποία κατατάχθηκε 1η, 4η, και 18η στα υπο-προβλήματα Α, Β και Γ αντίστοιχα πειραματίστηκε με αρκετά μοντέλα τα οποία περιείχαν γραμμικά μοντέλα, LSTMs και προ-εκπαιδευμένη μέθοδο BERT προσαρμοσμένη στο σύνολο OLID. Ιδιαίτερη έμφαση δόθηκε και στο στάδιο της προ επεξεργασίας με την αφαίρεση σχολίων μετά από δίεση (hashtag) η αντικατάσταση εικονιδίων (emoji) με κείμενα αντίστοιχης πληροφορίας.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

3.1 Μηχανική Μάθηση

3.1.1 Ορισμός Μηχανικής Μάθησης

Το φαινόμενο κατά το οποίο ένα σύστημα βελτιώνει την απόδοσή του κατά την εκτέλεση μιας συγκεκριμένης εργασίας, χωρίς να υπάρχει ανάγκη να προγραμματιστεί εκ νέου ονομάζεται Μηχανική Μάθηση. Πρόκειται για ένα κλάδο της τεχνητής νοημοσύνης που ως σκοπό έχει τη δημιουργία υπολογιστικών προγραμμάτων ικανών να μαθαίνουν, να βελτιώνουν δηλαδή την απόδοσή τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Mitchell (1997). Εμπειρία αποτελούν τα δεδομένα εισόδου τα οποία δέχονται την απαραίτητη προ-επεξεργασία προκειμένου να έρθουν σε κατάλληλη μορφή για να μπορεί το πρόγραμμα να αξιοποιήσει την πληροφορία που περιέχουν.

3.1.2 Βασικά είδη μάθησης και Χαρακτηριστικοί Αλγόριθμοι

- **Επιβλεπόμενη μάθηση**
- **Μη επιβλεπόμενη μάθηση**
- **Ενισχυτική μάθηση**

Στην πρώτη περίπτωση, το πρόγραμμα λαμβάνει δεδομένα εισόδου αλλά και την επιθυμητή τους έξοδο. Στα δεδομένα εισόδου δηλαδή έχουν δοθεί ετικέτες οι οποίες αντιπροσωπεύουν την πραγματική κλάση στην οποία ανήκουν. Έτσι μέσω του αλγορίθμου μάθησης το πρόγραμμα χρησιμοποιεί τα δεδομένα εισόδου και εξόδου (ετικέτες) για να εντοπίσει πρότυπα που θα αποτελέσουν τη βάση για την κατηγοριοποίηση νέων δεδομένων που θα τους δοθούν χωρίς τις επιθυμητές τους εξόδους (ετικέτες).

Στη δεύτερη περίπτωση, τα δεδομένα δεν διαθέτουν τις επιθυμητές εξόδους. Ο αλγόριθμος προσπαθεί να εντοπίσει πρότυπα, να βρει κανόνες και να συσχετίσει τα δεδομένα εισόδου μεταξύ τους.

Στην τρίτη περίπτωση, το ζητούμενο είναι να γίνουν οι απαιτούμενες δράσεις από τον αλγόριθμο οι οποίες θα επιφέρουν τη μεγαλύτερη επιβράβευση για μια συγκεκριμένη κατάσταση. Η απουσία δεδομένων εισόδου (εκπαίδευσης), έχει ως αποτέλεσμα ο αλγόριθμος να μαθαίνει μόνο από την εμπειρία που αποκτά. Τα κύρια σημεία της ενισχυτικής μάθησης είναι τα εξής:

- **Είσοδος**: Η είσοδος είναι μια αρχική κατάσταση από την οποία θα ξεκινήσει το μοντέλο
- **Έξοδος**: Υπάρχουν πολλές πιθανές έξοδοι αφού υπάρχει ποικιλία από πιθανές λύσεις

- **Εκπαίδευση**: Η εκπαίδευση βασίζεται στην είσοδο, το μοντέλο θα επιστρέψει μια κατάσταση για την οποία ο χρήστης θα αποφασίσει αν θα επιβραβεύσει ή θα τιμωρήσει το μοντέλο ανάλογα με την έξοδο.

- **Το μοντέλο συνεχίζει να μαθαίνει**

- **Η καλύτερη λύση αποφασίζεται με βάση τη μέγιστη επιβράβευση**

Η κύρια διαφορά με την επιβλεπόμενη μάθηση είναι ότι στην ενισχυτική μάθηση οι αποφάσεις λαμβάνονται ακολουθιακά. Με άλλα λόγια, μπορούμε να πούμε πως η έξοδος εξαρτάται από την κατάσταση της τρέχουσας εισόδου και η επόμενη είσοδος εξαρτάται από την έξοδο της προηγούμενης εισόδου. Οι αποφάσεις είναι εξαρτώμενες η μία με την άλλη. Στην επιβλεπόμενη μάθηση είναι ανεξάρτητες μεταξύ τους και οι ετικέτες δίνονται σε κάθε απόφαση. Reinforcemnt learning

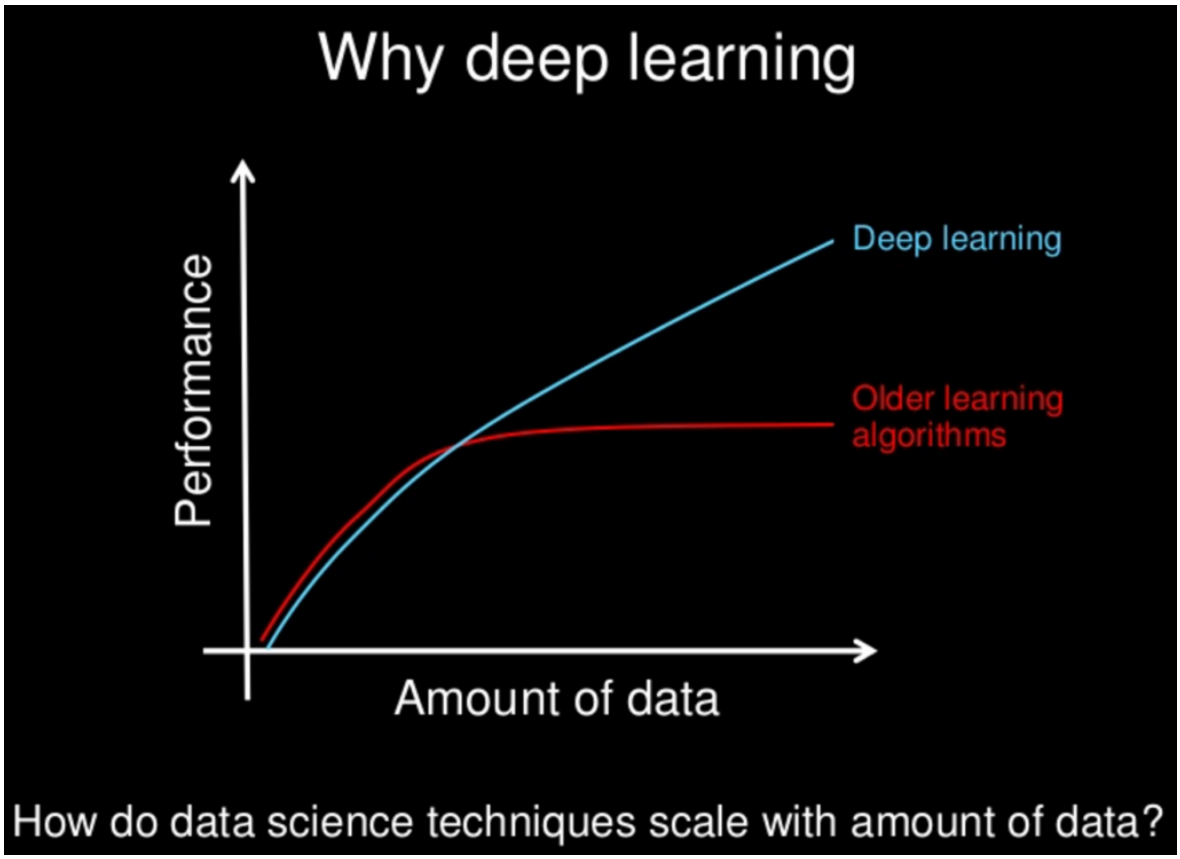
3.1.3 Ορισμός Βαθιάς Μάθησης

Η Βαθιά Μάθηση είναι ένα πεδίο της Μηχανικής Μάθησης που αφορά αλγορίθμους εμπνευσμένους από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Οι αλγόριθμοι αυτοί ονομάζονται τεχνητά νευρωνικά δίκτυα. Σε αυτή την ενότητα παρατίθενται ορισμοί και επεξηγήσεις για το τι είναι, πως λειτουργεί και πως προέκυψε η επιτυχία αυτής της μεθόδου, αναφέροντας ιδέες και τοποθετήσεις από διάσημους και διακεκριμένους επιστήμονες του κλάδου της τεχνητής νοημοσύνης.

Το 2013 σε ομιλία του ο Andrew Ng με τίτλο " Deep Learning, Self-Taught Learning and Unsupervised Feature Learning" περιέγραψε την ιδέα της Βαθιάς Μάθησης ως εξής:" Με τη χρήση προσομοίωσης του εγκεφάλου ελπίζουμε ότι θα κάνουμε τους αλγορίθμους μάθησης καλύτερους και πιο εύχρηστους πετυχαίνοντας επαναστατική πρόοδο στον τομέα της τεχνητής νοημοσύνης." Ο πυρήνας της Βαθιάς Μάθησης σύμφωνα με τον Andpew Ng είναι ότι τώρα πια υπάρχουν υπολογιστές αρκετά γρήγοροι και επαρκή (σε όγκο) δεδομένα για να εκπαιδεύσουν τεράστια νευρωνικά δίκτυα. Το σημείο κλειδί είναι η κλίμακα στη μέθοδο της Βαθιάς Μάθησης. Όπως σχολιάζει ο ίδιος, όσο μεγαλύτερα νευρωνικά δίκτυα εκπαιδεύουμε με όλο και περισσότερα δεδομένα τόσο η απόδοσή τους συνεχίζει να αυξάνεται. Αυτή είναι η κύρια διαφορά με άλλες τεχνικές Μηχανικής Μάθησης οι οποίες φτάνουν σε ένα όριο απόδοσης το οποίο δεν μπορούν να ξεπεράσουν.

Τέλος, ο Andrew Ng αναφέρει ότι μέχρι τώρα τα επιτεύγματα της Βαθιάς Μάθησης προέρχονται από μάθηση με επίβλεψη και ότι στο κοντινό μέλλον θα υπάρξει τεράστια πρόοδος και στο τομέα της μάθησης χωρίς επίβλεψη μιας και υπάρχει πληθώρα δεδομένων τα οποία δεν έχουν και ούτε είναι πρακτικό να αποκτήσουν ετικέτες.

Επιπρόσθετα της κλιμακωσιμότητας, δηλαδή της δυνατότητας ενός συστήματος να διαχειρίζεται όλο και μεγαλύτερο φόρτο εργασίας/δεδομένων, ένα ακόμα μεγάλο πλεονέκτημα της Βαθιάς Μάθησης είναι η ικανότητα της αυτόματης εξαγωγής χαρακτηριστικών από ανεπεξέργαστα δεδομένα. Αυτή η ιδιότητα ονομάζεται Μάθηση Χαρακτηριστικών (feature learning). Ένας ακόμα διακεκριμένος επιστήμονας στον τομέα της Βαθιάς Μάθησης ο Yoshua Bengio επισημαίνει την ικανότητα των αλγορίθμων Βαθιάς Μάθησης να ανακαλύπτουν και να μαθαίνουν καλές αναπαραστάσεις των δεδομένων χρησιμοποιώντας Μάθηση χαρακτηριστικών. Όπως ο ίδιος αναφέρει " Η Βαθιά Μάθηση ψάχνει τον τρόπο να εκμεταλλευτεί την απροσδιόριστη δομή της κατανομής των δεδομένων εισόδου με σκοπό την αναπαράστασή τους σε πολλαπλά επίπεδα στα οποία τα



Σχήμα 3.1: Βαθιά Μάθηση σε σύγκριση με άλλες τεχνικές μάθησης
Slide by Andrew Ng, all rights reserved.

υψηλότερου επιπέδου προσδιορισμένα χαρακτηριστικά ορίζονται με βάση αυτά των χαμηλότερων επιπέδων".

Σε αυτό το σημείο είναι κρίσιμο να τονιστεί η έννοια της ιεραρχίας των χαρακτηριστικών στη Βαθιά Μάθηση, κατά την οποία τα υψηλότερου επιπέδου χαρακτηριστικά προκύπτουν από την σύνθεση χαρακτηριστικών που βρίσκονται σε χαμηλότερο επίπεδο.

Τέλος, η αυτόματη μάθηση χαρακτηριστικών μέσω των επιπέδων των βαθιών νευρωνικών δικτύων επιτρέπει στο σύστημα να μαθαίνει πολύπλοκες συναρτήσεις που αντιστοιχίζουν την είσοδο με την έξοδο απευθείας, χωρίς αποκλειστικά τη χρήση χαρακτηριστικών που έχουν δημιουργηθεί από κάποιον άνθρωπο. Deep learning definition

3.1.4 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι σχεδιασμένα με βάση την εσωτερική λειτουργία του ανθρώπινου εγκεφάλου. Πιο συγκεκριμένα, προσομοιώνουν τις λειτουργίες των διασυνδεδεμένων νευρώνων περνώντας τα χαρακτηριστικά της εισόδου σε πολλαπλά επίπεδα τεχνητών νευρώνων που ονομάζονται perceptrons. Κάθε ένα από αυτά τα επίπεδα των perceptrons μετασχηματίζουν την είσοδο μέσω ενός συνόλου συναρτήσεων. Παρακάτω αναλύεται η δομή και τα συστατικά στοιχεία από τα οποία απαρτίζεται ένα perceptron το οποίο αποτελεί και τη βάση για τη δημιουργία των σύγχρονων νευρωνικών δικτύων. Neural Network Elements

- **Είσοδος** (Input signals): Η είσοδος είναι οι μετρήσεις των χαρακτηριστικών, η πληροφορία δηλαδή που θα δώσουμε στο νευρώνα (perceptron). Συνήθως, πρόκειται για ένα πίνακα της μορφής [59, 34, 21]

- **Βάρη** (Synaptic Weights): Τα βάρη αναπαριστούν βαθμωτούς πολλαπλασιασμούς. Ο ρόλος τους είναι να εκτιμήσουν πόσο σημαντική είναι η κάθε είσοδος αλλά και πως η αύξηση ή η μείωση της εκάστοτε εισόδου επηρεάζει την έξοδο.

- **Αθροιστής (Summing Function)**. Ο αθροιστής διαφέρει από τα άλλα στοιχεία γιατί λαμβάνει πολλές εισόδους. Η δουλειά του είναι να συνδυάζει όλες τις εισόδους και να τις μετατρέπει σε μία και μόνο έξοδο έτσι ώστε να μπορεί η συνάρτηση ενεργοποίησης να εφαρμοστεί σε αυτή την έξοδο.

- **Συνάρτηση ενεργοποίησης (Activation function)**. Η συνάρτηση ενεργοποίησης λαμβάνει την τιμή της εξόδου του αθροιστή και αποφασίζει εάν ο νευρώνας θα ενεργοποιηθεί ή όχι. Στόχος είναι η εισαγωγή μη-γραμμικότητας στην έξοδο του νευρώνα.

- **Πόλωση (bias)**. Η πόλωση είναι ουσιαστικά μια επιπλέον είσοδος με τιμή 1 ή -1 ίδια για όλα τα perceptrons. Αυτή πολλαπλασιάζεται με ένα βάρος όπως και οι άλλες είσοδοι και ο σκοπός της είναι να επιτρέψει στην τιμή πριν εισαχθεί στην συνάρτηση ενεργοποίησης να μετατοπιστεί πάνω ή κάτω ανεξάρτητα από την επίδραση των άλλων εισόδων.

3.1.5 Συναρτήσεις Ενεργοποίησης

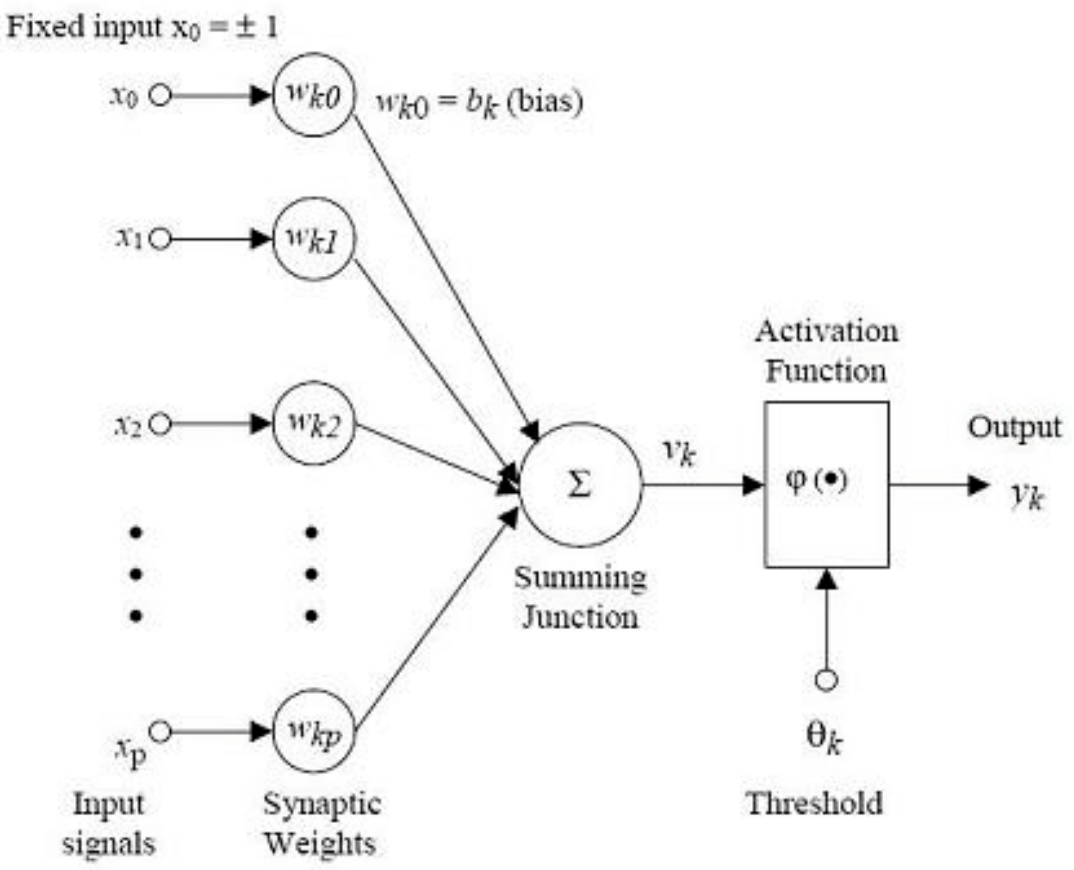
Όπως αναφέρθηκε προηγουμένως η συνάρτηση ενεργοποίησης είναι βασικό στοιχείο ενός νευρώνα. Πρόκειται για μη-γραμμικές συναρτήσεις απαραίτητες ώστε το νευρωνικό δίκτυο να αποκωδικοποιήσει πολύπλοκες αναπαραστάσεις των δεδομένων εισόδου και να βγάλει εύστοχα συμπεράσματα. Παρακάτω παρουσιάζονται μερικές από τις πιο γνωστές συναρτήσεις ενεργοποίησης.

- **Σιγμοειδής συνάρτηση**. Ο ρόλος της σιγμοειδούς συνάρτησης είναι να περιορίσει την έξοδο σε ένα εύρος τιμών μεταξύ 0 και 1, κάνοντας έτσι αυτό το είδος των συναρτήσεων χρήσιμο για πρόβλεψη πιθανοτήτων. Η σιγμοειδής συνάρτηση εκφράζεται από το μαθηματικό τύπο

$$f(x) = \frac{1}{1-e^{-x}}$$

- **Συνάρτηση ενεργοποίησης** $Softmax$. Η συνάρτηση αυτή εφαρμόζεται στο επίπεδο εξόδου σε πολλά είδη νευρωνικών δικτύων ειδικότερα όταν οι πιθανές κλάσεις κατηγοριοποίησης των δεδομένων εισόδου είναι παραπάνω από δύο. Σκοπός είναι η κανονικοποίηση των τιμών της εξόδου στο διάστημα $[0, 1]$ και την χρησιμοποιούμε όταν έχουμε πολλές εξόδους. Η συνάρτηση softmax δίνεται από τη σχέση:

$$Softmax(y_i) = \frac{e^{y_i}}{\sum_{i=0}^{N} e^{y_i}}$$



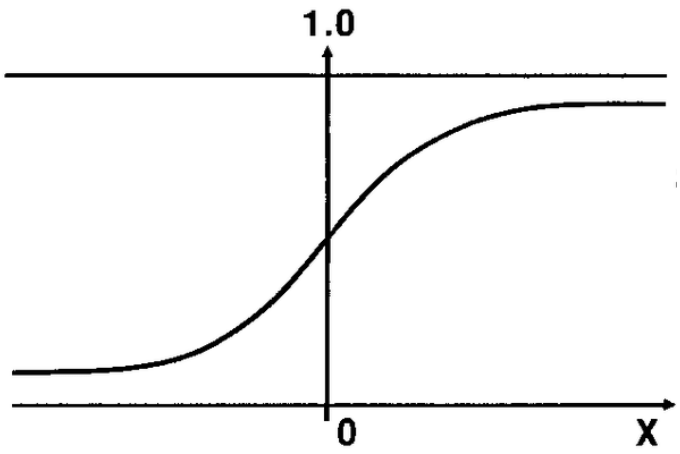
Σχήμα 3.2: Συστατικά μέρη ενός τεχνητού νευρώνα
ResearcgGate Slide by Les P. Beard

3.1.6 Πολυεπίπεδα νευρωνικά Δίκτυα

Για την επίλυση μη-γραμμικών προβλημάτων χρησιμοποιούνται πολλοί νευρώνες, όπως αυτός που αναλύσαμε στο Σχήμα 3.1, οι οποίοι χωρίζονται στα εξής επίπεδα:

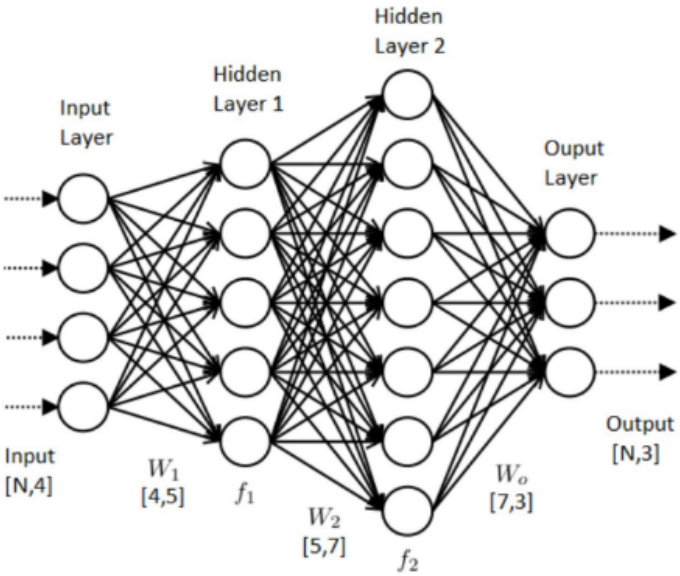
- Επίπεδο Εισόδου (Input Layer)
- Κρυφό Επίπεδο (Hidden Layer)
- Επίπεδο Εξόδου (Output Layer)

Το επίπεδο εισόδου λαμβάνει τα δεδομένα εισόδου και τα μεταφέρει στο κρυφό επίπεδο. Έπειτα το κρυφό επίπεδο επεξεργάζεται αυτά τα δεδομένα και είναι υπεύθυνο για να εξάγει τα απαραίτητα χαρακτηριστικά από αυτά. Μπορούν να υπάρχουν περισσότερα από ένα κρυφά επίπεδα σε ένα νευρωνικό δίκτυο. Το πλήθος τους έχει να κάνει με την πολυπλοκότητα των δεδομένων, με την έννοια ότι εάν βάλουμε υπερβολικά μεγάλο αριθμό κρυφών επιπέδων τότε το δίκτυο θα υπερεκπαιδευτεί και θα αντιγράψει τη μορφή των δεδομένων εισόδου χάνοντας την δυνατότητα πρόβλεψης. Δεν υπάρχει ακριβής φόρμουλα για την εύρεση του βέλτιστου αριθμού κρυφών επιπέδων. Αυτός καθορίζεται μέσω πειραμάτων και από την ευρύτερη εμπειρία του προγραμματιστή. Το επίπεδο



Σχήμα 3.3: Σιγμοειδής συνάρτηση
ResearcgGate Slide by Knut Kvaal

εξόδου συγκεντρώνει και μεταδίδει την πληροφορία στη μορφή που έχει καθοριστεί από το χρήστη. Ο αριθμός των νευρώνων εξόδου εξαρτάται από τη δουλειά που είναι προορισμένος να κάνει. Για παράδειγμα, εάν το δίκτυο λαμβάνει φωτογραφίες που αναπαριστούν γάτες ή αυτοκίνητα (κλάσεις), τότε ο αριθμός των νευρώνων εξόδου θα είναι ίδιος με το πλήθος των κλάσεων. Αξίζει να σημειωθεί ότι όλοι οι κύκλοι στο Σχήαμ 3.4 είναι νευρώνες (perceptrons).



Σχήμα 3.4: Πολυεπίπεδο Νευρωνικό Δίκτυο
Data Science Central Slide by Jason Brownlee

3.1.7 Συναρτήσεις Κόστους (Cost functions)

Αυτές οι συναρτήσεις χρησιμοποιούνται για την μέτρηση της απόδοσης του μοντέλου Μάθησης που χρησιμοποιούμε. Μέσω αυτών ποσοτικοποιούμε το σφάλμα μεταξύ προβλεπόμενων και πραγματικών τιμών σε έναν και μόνο αριθμό. Όσο αυτός ο αριθμός τείνει στο κάτω φράγμα της συνάρτησης κόστους τόσο καλύτερες είναι και οι προβλέψεις του μοντέλου. Μια ευρέως χρησιμοποιούμενη συνάρτηση κόστους η οποία αποτελεί και μέρος του μοντέλου της παρούσας εργασίας

είναι η κοινή εντροπία (Cross Entropy loss Function). Το σφάλμα εντροπίας (cross entropy loss) ή αλλιώς το λογαριθμικό σφάλμα (log loss) μετράει τη απόδοση ενός μοντέλου ταξινόμησης του οποίου η έξοδος είναι μια τιμή πιθανότητας μεταξύ 0 και 1. Το σφάλμα εντροπίας αυξάνει όσο η προβλεπόμενη τιμή πιθανότητας αποκλίνει από την πραγματική κλάση. Έτσι προβλέποντας μια πιθανότητα π.χ 0.12 όταν η πραγματική πιθανότητα της κλάσης είναι 1, το σφάλμα εντροπίας θα είναι μεγάλο. Ένα τέλειο μοντέλο θα είχε λογαριθμικό σφάλμα ίσο με το μηδέν. Στο ίδιο παράδειγμα, αν η προβλεπόμενη πιθανότητα (0.12) μεγάλωνε τότε το λογαριθμικό σφάλμα θα μειωνόταν. Αν η προβλεπόμενη πιθανότητα μειωνόταν, απομακρυνόταν δηλαδή από την επιθυμητή τιμή (1) τότε το λογαριθμικό σφάλμα θα αυξανόταν δραματικά. Η κοινή εντροπία και το λογαριθμικό σφάλμα είναι ελαφρώς διαφορετικές έννοιες αλλά στη μηχανική μάθηση όταν υπολογίζουμε σφάλματα, οι δύο έννοιες συγκλίνουν. Ο μαθηματικός τύπος για περισσότερες από δύο κλάσεις ($M > 2$) είναι ο εξής :

$$\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{3.1}$$

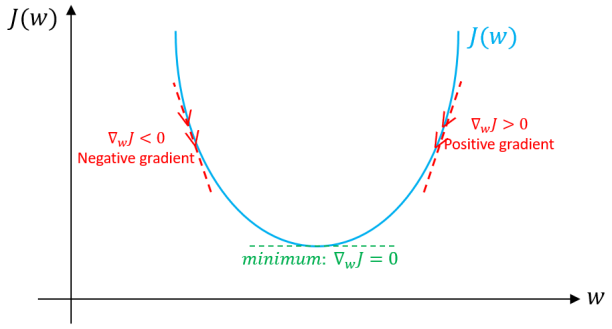
όπου

- M είναι το πλήθος των κλάσεων
- log ο φυσικός λογάριθμος
- y δυαδικός δείκτης (0 ή 1) αν η ετικέτα της κλάσης c είναι η σωστή για την παρατήρηση o
- p προβλεπόμενη πιθανότητα της παρατήρησης o για τον αν ανήκει στην κλάση c

3.1.8 Αλγόριθμος Κατάβασης Κλίσης (Gradient Descent Algorithm)

Ο αλγόριθμος κατάβασης κλίσης στόχο έχει την ελαχιστοποίηση της συνάρτησης κόστους J(w) όπου w είναι το διάνυσμα που αντιπροσωπεύει τις παραμέτρους του δικτύου. Ο αλγόριθμος κατάβασης κλίσης είναι ένας ευρέως γνωστός αλγόριθμος βελτιστοποίησης και με μεγάλη διαφορά ο πιο κοινός τρόπος για τη βελτιστοποίηση νευρωνικών δικτύων. Πρόκειται για μια επαναληπτική μέθοδο η οποία χωρίζεται σε βήματα. Για να βρει το ελάχιστο της συνάρτησης κόστους πρέπει να αποφασίσει σε ποια κατεύθυνση πρέπει να κινηθεί και πόσο μεγάλο θα είναι το βήμα με το οποίο θα προχωρήσει. Για την κατεύθυνση ο αλγόριθμος χρησιμοποιεί τις μερικές παραγώγους της συνάρτησης, ενώ το πλάτος του βήματος δίνεται ως υπερπαράμετρος στο μοντέλο με το όνομα ρυθμός μάθησης (learning rate). Αν και το μέγεθος του ρυθμού μάθησης καθορίζει πόσο γρήγορα θα φτάσουμε στο ελάχιστο, θα πρέπει να ληφθεί υπόψιν ότι εάν το βήμα είναι πολύ μεγάλο τότε υπάρχει πιθανότητα να ξεπεράσουμε το σημείο του ελαχίστου χωρίς αυτό να γίνει αντιληπτό. Από την άλλη μεριά εάν το βήμα είναι πολύ μικρό θα χρειαστούμε μεγάλο αριθμό επαναλήψεων. Συνεπώς, ο καθορισμός της τιμής αυτής της παραμέτρου στηρίζεται κυρίως σε πειράματα και στην εμπειρία του εκάστοτε κατασκευαστή του μοντέλου μηχανικής μάθησης.

Ο αλγόριθμος της οπίσθιας διάδοσης σφάλματος αν και διατυπώθηκε πρώτη φορά τη δεκαετία του 70, η χρησιμότητα δεν είχε γίνει πλήρως αντιληπτή από την επιστημονική κοινότητα μέχρι τη διάσημη δημοσίευση Rumelhart and Williams ((1986). Αυτή η δημοσίευση περιγράφει πολλά νευρωνικά δίκτυα όπου η μέθοδος οπίσθιας διάδοσης δουλεύει πολύ γρηγορότερα από κάθε άλλη



Σχήμα 3.5: Αλγόριθμος πίσω Διάδοσης Σφάλματος (Backpropagation algorithm)

μέθοδο που είχε ήδη δοκιμαστεί. Στις μέρες μας είναι σχεδόν η αποκλειστική μέθοδος η οποία χρησιμοποιείται για τη μάθηση ενός νευρωνικού δικτύου. Στον πυρήνα για την κατανόηση της μεθόδου βρίσκεται η μερική παράγωγος $\frac{\partial C}{\partial w}$ της συνάρτησης κόστους C ως προς κάθε βάρος w του δίκτυο. Υπολογίζοντας αναλυτικά όλες αυτές τις μερικές παραγώγους των σφαλμάτων ορίζει την μέθοδο οπίσθιας διάδοσης (backpropagation). Αυτός ο τίτλος δόθηκε από τη λειτουργία της μεθόδου στα πολυεπίπεδα νευρωνικά δίκτυα. Πιο συγκεκριμένα έστω $y = f(w, x)$ η εξίσωση για την ταξινόμηση του δικτύου, όπου x και y είναι η είσοδος και η έξοδος του δικτύου και w οι παράμετροι (βάρη). Οι υπολογισμοί για την εύρεση της συνάρτησης $f(w, x$ "ρέουν εμπρόσθια" και "ρέουν οπίσθια" για την εύρεση για τον υπολογισμό της παραγώγου $\frac{df}{dt}$. Μόλις, αποκτήσει κανείς τους υπολογισμούς των παραγώγων των βαρών πρέπει να τα επανακαθορίσει ελαφρώς αυτά τα βάρη προς την σωστή κατεύθυνση μέσω του αλγορίθμου κατάβασης κλήσης.

Άρα ανακεφαλαιώνοντας

- Αρχικά πρέπει να ορίσουμε τους όρους της εξίσωσης $y = f(w, x)$ όπου y το διάνυσμα εξόδου (κλάσεις) x το διάνυσμα εισόδου (πίνακας χαρακτηριστικών) και w ο πίνακας των παραμέτρων που πρέπει να μάθει το μοντέλο

- Για την μέτρηση απόδοσης της μάθησης του μοντέλου ορίζουμε τη συνάρτηση κόστους C

- Βρίσκουμε την παράγωγο του σφάλματος για κάθε βάρος w_i αναλυτικά

- Επαναπροσδιορίζουμε την τιμή αυτού του βάρους κατά τη φάση της εκπαίδευσης μέσω του αλγορίθμου κατάβασης κλίσης

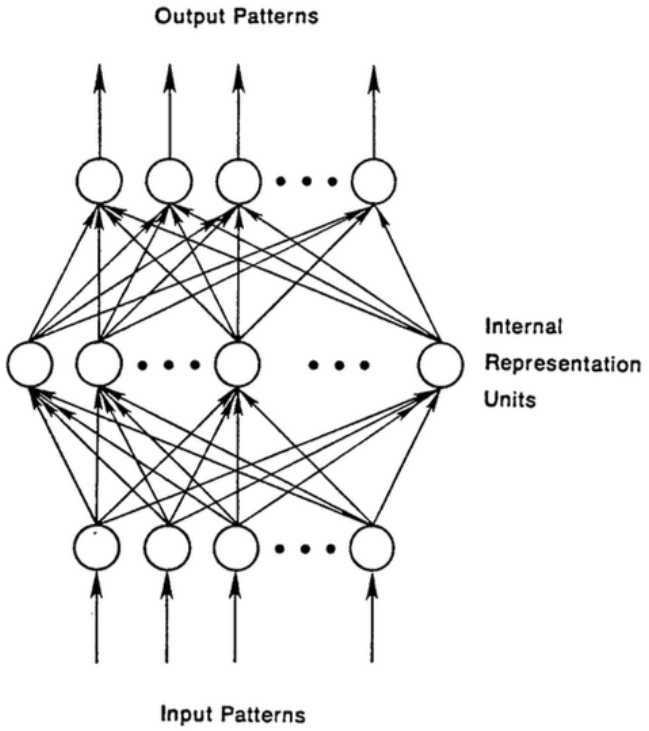
3.2 Επαναλαμβανόμενα νευρωνικά Δίκτυα (Recurrent Neural Networks)

Τα επαναλαμβανόμενα νευρωνικά δίκτυα (reccurent neural nets) Lipton et al. (2015) είναι ένα είδος τεχνητού νευρωνικού δικτύου σχεδιασμένο να αναγνωρίζει μοτίβα σε ακολουθίες δεδομένων όπως χρηματιστηριακές τιμές, κείμενα αλλά και ηχητικά δεδομένα. Το πιο γνωστό από αυτά τα δίκτυα ονομάζεται Δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short Memory Networks).

Αυτό που διαφοροποιεί τα επαναλαμβανόμενα (RNNs) από τα υπόλοιπα δίκτυα είναι ότι λαμβάνουν υπόψιν το χρόνο και την ακολουθία ως προσωρινές διαστάσεις (μνήμη). Λόγω αυτής της μνημονικής ιδιότητας που υπάρχει σε κάποια μορφή στα RNNs, θα γίνει αρκετές φορές αναφορά σε αναλογίες της πραγματικής μνήμης (ανθρώπινης) με την ιδιάζουσα μνήμη των τεχνητών αυτών δικτύων.

3.2.1 Περιληπτική Αναφορά στα Δίκτυα Εμπρόσθιας τροφοδότησης (Feedforward Neural Nets))

Αμφότερα τα δύο είδη νευρωνικών δικτύων (FNN RNN) παίρνουν το όνομά τους από τον τρόπο με τον οποίο περνούν την πληροφορία μέσω μαθηματικών πράξεων σε κάθε κόμβο του δικτύου. Από τη μία τα εμπρόσθια (FNNs) περνούν την πληροφορία απευθείας (χωρίς να αγγίζουν τον ίδιο κόμβο δεύτερη φορά) ενώ τα RNNs δημιουργούν ένα κύκλο μέσω ανάδρασης ο οποίος επιτρέπει στην πληροφορία να διατηρηθεί σε ένα κόμβο. Στην περίπτωση των εμπρόσθιων (FNNs) δικτύων, τα δεδομένα εισόδου εισέρχονται στο δίκτυο και μετασχηματίζονται σε δεδομένα εξόδου. Αν η μάθηση γίνεται με επίβλεψη η έξοδος θα είναι μια ετικέτα η οποία αναφέρεται σε μια κατηγορία των δεδομένων εισόδου. Αντιστοιχίζονται δηλαδή δεδομένα σε κατηγορίες αναγνωρίζοντας πρότυπα (μοτίβα).

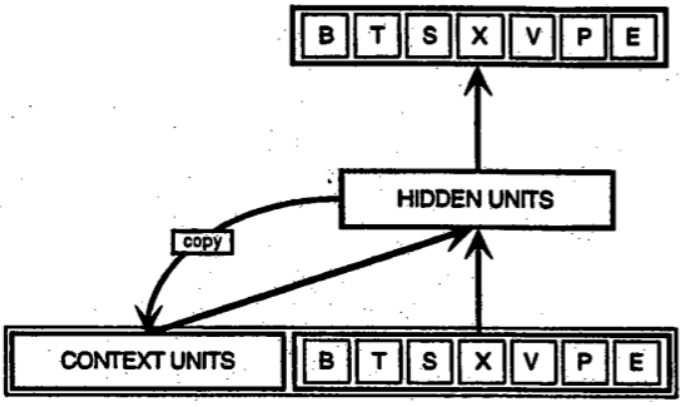


Σχήμα 3.6: Δίκτυα Εμπρόσθιας τροφοδότησης

Ένα εμπρόσθιο δίκτυο (FNN) εκπαιδεύεται σε κατηγοριοποιημένα δεδομένα ελαχιστοποιώντας το σφάλμα που προκύπτει όταν προβλέπει (μαντεύει) τις κατηγορίες των δεδομένων. Έπειτα, μόλις τελειώσει η διαδικασία της εκπαίδευσης και έχουν υπολογιστεί όλες οι παράμετροι του δικτύου (έχει βρει δηλαδή ότι πρότυπα και μοτίβα ήταν δυνατόν να ανιχνευθούν από τα δεδομένα εισόδου) το δίκτυο λαμβάνει ως είσοδο δεδομένα τα οποία δεν έχει ξαναεπεξεργαστεί ούτε γνωρί-

ζει σε ποια κατηγορία ανήκουν. Για παράδειγμα, μια πιθανή εφαρμογή θα ήταν να "τραφεί" αρχικά το δίκτυο με φωτογραφίες οι οποίες αναπαριστούν μια γάτα ή έναν ελέφαντα, να εκπαιδευτεί το δίκτυο σε ένα πλήθος φωτογραφιών για το οποίο η κατηγορία είναι γνωστή και αφού ολοκληρωθεί αυτό το βήμα της εκπαίδευσης (έχουν καθοριστεί τα βάρη w της εξίσωσης $y = f(w, x)$) το δίκτυο θα "τραφεί" με ένα σύνολο από άγνωστες για αυτό φωτογραφίες και θα πρέπει να αποφανθεί αν σε αυτές περιέχεται γάτα (πρώτη κατηγορία) ή ελέφαντας (δεύτερη κατηγορία). Όπως εύκολα διακρίνει κανείς, αυτός ο μηχανισμός του δικτύου εμπρόσθιας τροφοδότησης δεν περιλαμβάνει καμία αίσθηση της χρονικής διάστασης και το μόνο που λαμβάνει υπόψιν είναι το δεδομένο (φωτογραφία στο προαναφερθέν παράδειγμα) την κάθε χρονική στιγμή σειριακά.

Συνεπώς, αγνοώντας τελείως το πρόσφατο παρελθόν αυτά τα δίκτυα χαρακτηρίζονται αμνησιακά. Από την άλλη μεριά τα επαναλαμβανόμενα νευρωνικά δίκτυα παίρνουν ως είσοδο αυτό που τους δίνεται την εκάστοτε χρονική στιγμή αλλά και αυτό που τους δόθηκε προηγουμένως στο χρόνο. Παρατίθεται ένα διάγραμμα Servan-Schreiber et al. ((1991) όπου το BTSXPE στο κάτω μέρος αποτελεί το παράδειγμα εισόδου την τωρινή στιγμή και το CONTEXT UNITS αποτελεί το παράδειγμα της εξόδου προγενέστερων χρονικών στιγμών.



Σχήμα 3.7: Δομή Επαναλαβανόμενου Νευρωνικού Δικτύου

Η απόφαση που πήρε ένα επαναλαμβανόμενο δίκτυο τη χρονική στιγμή $t - 1$ επηρεάζει την απόφαση που θα πάρει την επόμενη χρονική στιγμή t. Συνεπώς, τα επαναλαμβανόμενα δίκτυα έχουν δύο πηγές εισόδου το παρόν και το πρόσφατο παρελθόν. Συνδυάζοντας αυτές αποφασίζει τι πληροφορία δίνουν τα καινούργια δεδομένα. Αυτό που τα διαφοροποιεί ουσιαστικά από τα εμπρόσθια δίκτυα (FNN) είναι ο βρόχος που συνδέεται με τις παλαιότερες αποφάσεις, λαμβάνοντας τις εξόδους από στιγμή σε στιγμή ως είσοδο, προσθέτοντας μνήμη τελικά στο δίκτυο. Αυτό δεν γίνεται τυχαία καθώς στην ίδια την ακολουθία (στη δομή της) υπάρχει πληροφορία που μπορεί να αξιοποιηθεί από τα RNNs. Η ακολουθιακή πληροφορία διατηρείται στο κρυφό επίπεδο (hidden state). Εκεί το μοντέλο αναζητά συσχετίσεις γεγονότων (δεδομένων) που έχουν συμβεί σε ξεχωριστές χρονικές στιγμές οι οποίες μπορεί να είναι αρκετά απομακρυσμένες (χρονικά) μεταξύ τους (long-term dependencies).

Περιγράφουμε μαθηματικά τη διαδικασία της μετάδοσης της μνήμης ως εξής :

$$h_t = f_h(W_h x_t + U_h h_{t-1} + b_h) \tag{3.2}$$

$$y_t = f_y(W_y h_t + b_y) \tag{3.3}$$

Εκτός από το διάνυσμα βαρών W υπάρχει ένα επιπλέον διάνυσμα βαρών U το οποίο επιδρά πάνω στην έξοδο του προηγούμενου νευρώνα και συμβάλλει με τη σειρά του στον υπολογισμό της επόμενης κατάστασης. Στις παραπάνω εξισώσεις y_t η έξοδος του δικτύου τη χρονική στιγμή t, W_h W_t U_h τα βάρη του δικτύου, x_t η είσοδος του δικτύου τη χρονική στιγμή t, h_t η κρυφή έξοδος του δικτύου τη χρονική στιγμή t και τέλος b_h b_y οι τιμές πόλωσης για τις κρυφές καταστάσεις και την έξοδο του δικτύου.

Το σφάλμα γυρνάει μέσω του αλγορίθμου οπίσθιας διάδοσης προσαρμόζοντας τα βάρη μέχρι το σφάλμα να γίνει ελάχιστο. Το άθροισμα, λοιπόν των βαρών συμπιέζεται από τη συνάρτηση ενεργοποίησης f (μη γραμμική). Αξίζει να αναφερθεί ότι επειδή ο βρόχος ανάδρασης λειτουργεί κάθε χρονική στιγμή της ακολουθίας, κάθε κρυφή κατάσταση (hidden state) εμπεριέχει 'ίχνη' όχι μόνο από την αμέσως προηγούμενη χρονική στιγμή αλλά από όλες όσες προηγήθηκαν (με όριο όσα χωράει η μνήμη). Recurrent Neural Networks

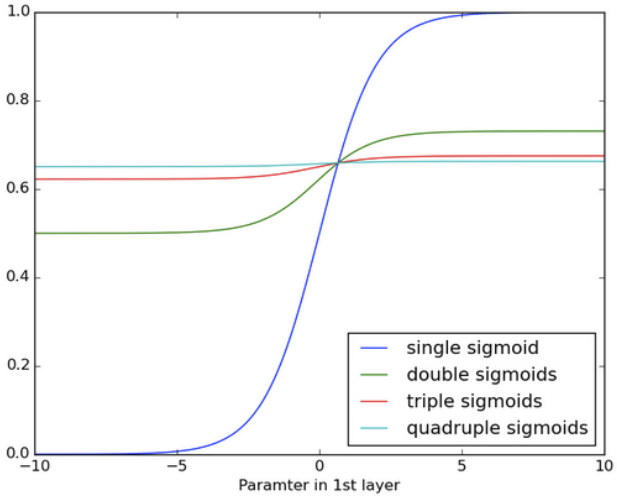
3.2.2 Οπίσθια Διάδοση στο Χρόνο

Όπως έχει ήδη αναφερθεί, σκοπός ενός επαναλαμβανόμενου νευρωνικού δικτύου είναι να κατηγοριοποιήσει επιτυχώς τα δεδομένα εισόδου που βρίσκονται σε ακολουθιακή μορφή. Αυτό επιτυγχάνεται όπως και σε όλα τα άλλα νευρωνικά δίκτυα με τις μεθόδους της οπίσθιας διάδοσης (Back Propagation) και κατάβασης κλίσης (Gradient Descent). Η λειτουργία της οπίσθιας διάδοσης έχει ήδη αναφερθεί εκτενώς για τα FNNs. Συνοπτικά, η λειτουργία είναι η εξής: αφού υπολογίστηκε το σφάλμα του δικτύου πηγαίνουμε προς τα πίσω στις εξόδους, στα βάρη και τέλος στις εισόδους του κάθε κρυφού επιπέδου. Ο αλγόριθμος της οπίσθιας διάδοσης δίνει το ποσοστό που αναλογεί σε κάθε βάρος για το τελικό σφάλμα. Ο αλγόριθμος υπολογίζει την ευθύνη που αναλογεί σε κάθε βάρος (μέσω των μερικών παραγώγων) ως προς το ολικό σφάλμα E. Αυτές οι παράγωγοι μετά χρησιμοποιούνται από τον κανόνα μάθησης (learning rule) και από τον αλγόριθμο κατάβασης κλίσης (gradient descent) για να προσαρμόσουν τα βάρη είτε προς τα πάνω είτε προς τα κάτω ανάλογα με το ποια κατεύθυνση μειώνει το σφάλμα. Τα επαναλαμβανόμενα νευρωνικά δίκτυα χρησιμοποιούν μια επέκταση του αλγορίθμου της οπίσθιας διάδοσης, την οπίσθια διάδοση στο χρόνο. Ο χρόνος σε αυτή την περίπτωση εκφράζεται ως μια καλά ορισμένη ακολουθία υπολογισμών η οποία συνδέει το ένα χρονικό βήμα με το άλλο. Όλα τα νευρωνικά δίκτυα, επαναλαμβανόμενα ή μη, μεταφράζονται μαθηματικά ως εμφωλευμένες συναρτήσεις της μορφής f(g(h(x))). Προσθέτοντας το στοιχείο του χρόνου απλά επεκτείνουμε την ακολουθία των συναρτήσεων και υπολογίζουμε τις μερικές παραγώγους με τον κανόνα της αλυσίδας. Backpropagation Through Time (BPTT)

3.2.3 Το πρόβλημα του (Vanishing (and Explonding) Gradients)

Η ιδέα των επαναλαμβανόμενων δικτύων είναι αρκετά παλιά. Ήδη από τις αρχές της δεκαετίας του ενενήντα είχαν δημιουργηθεί μοντέλα με επαναλαμβανόμενα δίκτυα τα οποία όμως είχαν ένα σημαντικό μειονέκτημα που λειτουργούσε ως εμπόδιο στην απόδοσή τους, αυτό της εκρηκτικής μείωσης ή αύξησης της κλίσης (παραγώγου) (vanishing or exploding gradient). Η κλίση (gradient) εκφράζει την αλλαγή που προκύπτει σε όλα τα βάρη σε σχέση με την αλλαγή στο σφάλμα. Αν

δεν γνωρίζουμε την κλίση (gradient) δεν μπορούμε να καθορίσουμε σε ποια κατεύθυνση πρέπει να μεταβάλουμε τα βάρη έτσι ώστε να μειώσουμε το σφάλμα (error). Τα επαναλαμβανόμενα δίκτυα προσπαθούν να συνδέσουν την έξοδο με δεδομένα που συνέβησαν πολλά χρονικά βήματα πίσω και να εκτιμήσουν το ποσοστό που επηρεάζουν αυτά τα απομακρυσμένα δεδομένα την έξοδο. Αυτό είναι ιδιαιτέρως δύσκολο και ένας από τους λόγους είναι ότι η πληροφορία στα νευρωνικά δίκτυα περνάει από πολλά στάδια με πολλαπλασιασμούς.Vanishing (and Exploding) Gradients



Σχήμα 3.8: Πολλαπλή εφαρμογή της σιγμοειδούς συνάρτησης στο πρώτο επίπεδο ενός δικτύου

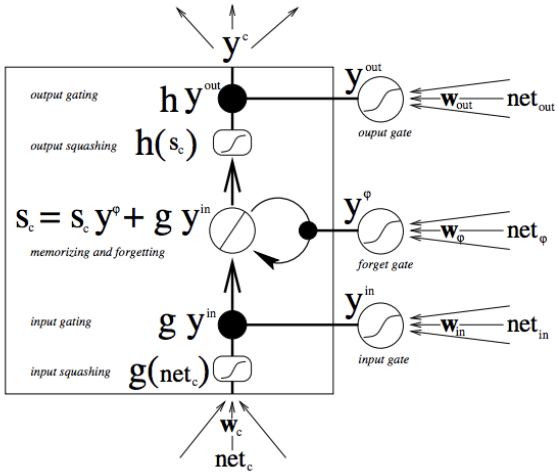
Όταν πολλαπλασιάζεις μία ποσότητα με ένα αριθμό μεγαλύτερο της μονάδας όσο λίγο μεγαλύτερος κι αν είναι αυτός από τη μονάδα άμα πολλαπλασιαστεί πάρα πολλές φορές τότε η ποσότητα θα γίνει τεράστια. Ομοίως όταν πολλαπλασιάσεις πολλές φορές με μια ποσότητα μικρότερη της μονάδας αυτή κάποια στιγμή θα γίνει σχεδόν μηδέν. Από τα παραπάνω γίνεται πλέον προφανές ότι επειδή τα επίπεδα (layers) του νευρωνικού δικτύου συνδέονται με σχέσεις πολλαπλασιασμού, οι παράγωγοι καταλήγουν είτε να εξαλείφονται (vanishing) είτε να τείνουν προς το άπειρο (explode). Στο Σχήμα 3.8 φαίνονται τα αποτελέσματα εφαρμογής της σιγμοειδούς συνάρτησης επαναλαμβανόμενες φορές. Τα δεδομένα τείνουν στην ευθεία (επιπεδοποιούνται) και χάνουν οποιαδήποτε καμπύλη στην οποία μπορούν να υπολογιστούν οι κλίσεις.

3.3 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long short term Memory Networks)

Η λύση στο πρόβλημα εκρηκτικής μείωσης κλίσης (Vanishing Problem) έρχεται να δώσει ένα είδος επαναλαμβανόμενου δικτύου που ονομάζεται δίκτυο Μακράς Βραχυπρόθεσμης μνήμης (Long Short Term Memory) το οποίο προτάθηκε για πρώτη φορά το 1997 Hochreiter and Schmidhuber (1997) από τους ερευνητές Sepp Hochreiter και Juergen Schmidhuber. Διατηρώντας ένα πιο σταθερό σφάλμα, επιτρέπουν στα επαναλαμβανόμενα δίκτυα να "μαθαίνουν" ακόμα και μετά το πέρας πολλών χρονικών βημάτων (πάνω από χίλια) μέσω ενός καναλιού το οποίο συνδέει απομακρυσμένα αίτια και αιτιατά. Τα LSTMs περιέχουν την πληροφορία σε διαφορετική μορφή από τα άλλα επαναλαμβανόμενα δίκτυα. Για την αποθήκευση, την εγγραφή ή το διάβασμα χρησιμοποιείται ένα κύτταρο πύλης (gated cell) παρόμοιο με αυτό της μνήμης ενός υπολογιστή. Το κύτταρο

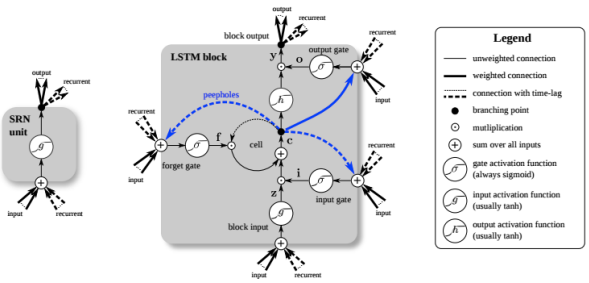
αυτό παίρνει τις αποφάσεις όσον αφορά το τι θα αποθηκευτεί, αν θα επιτραπεί το διάβασμα η εγγραφή ή η διαγραφή ενός δεδομένου μέσω πυλών οι οποίες ανάλογα με την ενέργεια που θέλουμε να πραγματοποιηθεί ανοίγουν ή κλείνουν.

Σε αντίθεση με τα κύτταρα μνήμης ενός ηλεκτρονικού υπολογιστή αυτές οι πύλες είναι αναλογικές και υλοποιούνται με πολλαπλασιασμούς (ανά στοιχείο) σιγμοειδών συναρτήσεων με εύρος $[0, 1]$. Το πλεονέκτημα των αναλογικών πυλών είναι η παραγωγισιμότητά τους σε αντίθεση με τις ψηφιακές, κάτι που τις καθιστά κατάλληλες για εφαρμογή του αλγορίθμου οπίσθιας διάδοσης (back propagation). Αυτές οι πύλες δρουν πάνω στα σήματα που δέχονται και παρόμοια με τους κόμβους του νευρωνικού δικτύου είτε αποκλείουν είτε επιτρέπουν την διέλευση της πληροφορίας αποφασίζοντας βασιζόμενες σε δικές τους συστοιχίες βαρών. Τα βάρη αυτά όπως ακριβώς και τα άλλα βάρη του νευρωνικού δικτύου αναπροσαρμόζονται από τη μέθοδο μάθησης του δικτύου. Η διαδικασία δηλαδή είναι η πύλη να 'μαντέψει' ποια είσοδο θα 'κόψει 'η θα 'αφήσει να περάσει μέσω των βαρών και έπειτα ο αλγόριθμος της οπίσθιας διάδοσης και της κατάβασης κλίσης (gradient descent) θα αναπροσαρμόσουν τα βάρη ώστε το σφάλμα να μειωθεί. Το Σχήμα 3.9 απεικονίζει τη ροή των δεδομένων σε ένα τέτοιο κύτταρο μνήμης μέσω των πυλών.



Σχήμα 3.9: Ροή πληροφορίας σε ένα κύτταρο μνήμης LSTM δικτύου

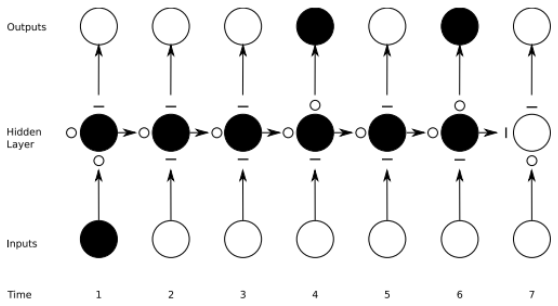
Ξεκινώντας από το κάτω μέρος του διαγράμματος τα τριπλά βέλη δείχνουν τη ροή της πληροφορίας στο κύτταρο. Τον συνδυασμό της εισόδου την τωρινή χρονική στιγμή και της κατάστασης του κυττάρου την προηγούμενη χρονική στιγμή, όπως φαίνεται στο σχήμα, δεν τον λαμβάνει μόνο το ίδιο το κύτταρο αλλά και η κάθε μία από τις τρεις πύλες του ξεχωριστά.



Σχήμα 3.10: Σύγκριση κυττάρων απλού επαναλαμβανόμενου δικτύου (αριστερά) και LSTM (δεξιά))

Οι μαύρες κουκίδες είναι οι πύλες, S_c είναι η τωρινή κατάσταση του κυττάρου μνήμης και g_{y_in} είναι η είσοδος που δέχεται το κύτταρο. Τα μεγάλα και έντονα γράμματα στο διάγραμμα αναπαριστούν τα αποτελέσματα του κάθε σταδίου λειτουργίας του κυττάρου. Παρατίθεται ακόμα ένα διάγραμμα 3.10 το οποίο δείχνει ένα απλό στοιχείο νευρωνικού δικτύου (στα αριστερά) και ένα κύτταρο δικτύου μακράς βραχυπρόθεσμης μνήμης (στα δεξιά). Είναι σημαντικό να τονίσουμε ότι τα κύτταρα μνήμης του LSTM δίνουν διαφορετικούς ρόλους στις αθροίσεις και τους πολλαπλασιασμούς που δημιουργούν τον μετασχηματισμό της εισόδου. Το θετικό πρόσημο που βρίσκεται στο κέντρο του κύτταρου αποτελεί το μυστικό της επιτυχίας του LSTM. Όσο απλό κι αν φαίνεται, αυτή η βασική αλλαγή βοηθά το κύτταρο να διατηρήσει ένα σταθερό σφάλμα όταν αυτό πρέπει να διαδοθεί προς τα πίσω (back propagated) σε μεγάλο βάθος. Πιο αναλυτικά, αντί να υπολογίσει τη μεταγενέστερη κατάσταση του κυττάρου, πολλαπλασιάζοντας την τωρινή κατάσταση με την είσοδο, απλά τις προσθέτει.

Η πύλη αμνησίας (forget gate) παρόλα αυτά συνεχίζει να βασίζεται στον πολλαπλασιασμό. Η πύλη αυτή έχει την ιδιότητα να "ξεχνά" κάτι που είναι ιδιαιτέρως χρήσιμο ακόμα και σε δίκτυα που ο κύριος στόχος τους είναι να διατηρούν χρονικά απομακρυσμένες πληροφορίες που όμως σχετίζονται με κάποιον τρόπο μεταξύ τους. Για παράδειγμα, κατά την ανάλυση κειμένων θα ήταν παράλογο να θεωρήσει κανείς πως μόλις τελειώσει το ένα κείμενο και συνεχίσει στο επόμενο θα υπάρχει κάποια συσχέτιση μεταξύ των δύο. Μόνο μεταξύ των προτάσεων του ιδίου κειμένου θα μπορούσε να υπάρχει εξάρτηση. Στο Σχήμα 3.11 διακρίνονται οι πύλες σε λειτουργία, με τις κάθετες γραμμές να αναπαριστούν κλειστές πύλες και τους λευκούς κύκλους ανοιχτές. Οι γραμμές και οι κύκλοι με οριζόντια κατεύθυνση είναι οι πύλες αμνησίας (forget gates).

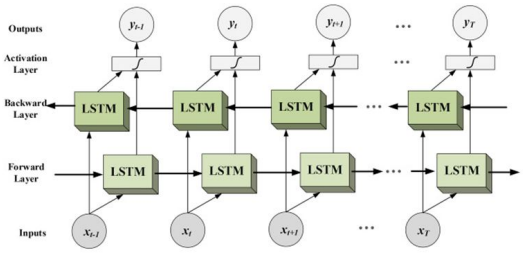


Σχήμα 3.11: Αναπαράσταση των πυλών ενός δικτύου LSTM)

Πρέπει να τονιστεί ότι ενώ στα δίκτυα εμπρόσθιας τροφοδότησης (Feed forward neural nets) μια είσοδος αντιστοιχίζεται σε μία έξοδο, στα επαναλαμβανόμενα δίκτυα μπορούμε να αντιστοιχίσουμε μια σε πολλές (one to many) όπως στο παραπάνω σχήμα, πολλές σε πολλές (many to many) ή πολλές σε μια (many to one), εισόδους σε εξόδους.

3.3.1 Bidirectional LSTMs

Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης διπλής κατεύθυνσης είναι μια επέκταση των τυπικών LSTM τα οποία έχουν σημαντικά καλύτερη απόδοση σε προβλήματα κατηγοριοποίησης ακολουθιών.



Σχήμα 3.12: Αμφίδρομο Επαναλαμβανόμενο Νευρωνικό Δίκτυο

Αυτά τα δίκτυα εκπαιδεύουν συγχρόνως δυο LSTM δίκτυα στην ακολουθία εισόδου, το ένα στην ακολουθία στην αρχική της μορφή και το δεύτερο σε ένα αντίγραφο της ακολουθίας που έχει αντίστροφη φορά. Αυτό το καινούργιο περιεχόμενο εισάγεται στο δίκτυο δημιουργώντας καλύτερα και γρηγορότερα αποτελέσματα.

3.4 Επεξεργασία κείμενου

Στο τομέα της τεχνητής νοημοσύνης και ειδικότερα στο κομμάτι της Επεξεργασίας φυσικής γλώσσας, τα δεδομένα πρέπει να υποστούν ειδική επεξεργασία και να μετασχηματιστούν σε μορφή "κατανοητή" από ένα ηλεκτρονικό υπολογιστή. Η μαθηματική αναπαράσταση των λέξεων είναι ένα καθοριστικό αλλά συγχρόνως απαιτητικό πρόβλημα. Στη συνέχεια παρουσιάζονται προσεγγίσεις σε αυτό το θέμα.

3.4.1 Σύνολα από λέξεις (Bag of words)

Το μοντέλο Συνόλων από λέξεις (Bag-of-words) Zhang et al. (2010) είναι ένας τρόπος αναπαράστασης κειμένων ο οποίος έχει χρησιμοποιηθεί με επιτυχία σε πολλά προβλήματα όπως μοντελοποίηση γλώσσας (language modeling) και κατηγοριοποίηση κειμένων (document classification). Ένα μοντέλο συνόλου από λέξεις (bag-of words) είναι ένας τρόπος να εξάγουμε χαρακτηριστικά από ένα κείμενο με σκοπό τη χρήση τους σε αλγορίθμους μηχανικής μάθησης. Πρόκειται για μία αναπαράσταση που περιγράφει πόσες φορές εμφανίζεται κάθε λέξη σε ένα κείμενο. Για το λόγο αυτό απαιτείται ένα δεδομένο λεξιλόγιο και ένας τρόπος να μετριούνται τα στιγμιότυπα μιας γνωστής λέξης στο κείμενο. Το όνομα (bag) αυτής της μεθόδου οφείλεται στο γεγονός ότι εκτός από το πλήθος των εμφανίσεων δεν λαμβάνεται υπόψιν καμία πληροφορία ως προς τη δομή και τη σειρά των λέξεων στο κείμενο. Με άλλα λόγια αυτή η μέθοδος ασχολείται μόνο αν υπάρχει μια λέξη στο κείμενο, όχι σε ποιο σημείο του κειμένου αυτή βρίσκεται. Το αισθητήριο πίσω από αυτή την ιδέα είναι ότι παρόμοια κείμενα έχουν μεγάλο αριθμό κοινών λέξεων και ότι αυτές οι λέξεις δίνουν πληροφορίες για το περιεχόμενο των κειμένων. Παρακάτω δίνεται ένα παράδειγμα με τα βήματα που ακολουθεί ο αλγόριθμος; (bag-og-words).

- **Βήμα** 1: Συλλογή Δεδομένων

  Έστω ότι εξάγουμε τα ακόλουθα δεδομένα κείμενου " Ήταν οι καλύτερες των εποχών, Ήταν οι χειρότερες των εποχών, Ήταν η στιγμή της σοφίας." Σε αυτό το δείγμα θεωρούμε τη κάθε γραμμή ως ξεχωριστό "κείμενο" και τις τρεις γραμμές συνολικά ως τη συλλογή των κειμένων(corpus).

- **Βήμα** 2: Σχεδιασμός του λεξιλογίου

  Δημιουργείται μια λίστα με τις ξεχωριστές λέξεις που περιέχει η συλλογή. Αυτό είναι ένα λεξιλόγιο 10 λέξεων από μία συλλογή κειμένων που περιέχει 15 λέξεις.

- **Βήμα** 3: Δημιουργία Διανυσμάτων κειμένου (Document vectors)

  Ο στόχος αυτού του βήματος είναι η μετατροπή του κειμένου σε ένα διάνυσμα το οποίο θα δοθεί ως είσοδος ή έξοδος σε ένα μοντέλο μηχανικής μάθησης. Στο παράδειγμά μας, αφού το μέγεθος του λεξικού είναι 10, το διάνυσμα αναπαράστασης θα έχει κι αυτό 10 θέσεις, με κάθε μια να αντιστοιχεί σε μια δυαδική τιμή (0 αν δεν υπάρχει η λέξη στο συγκεκριμένο κείμενο και 1 αν υπάρχει). Για το πρώτο κείμενο (πρώτη γραμμή) θα είναι:

  "Ήταν" = 1 "οι" = 1 "καλύτερες" = 1 "των" = 1 "εποχών" = 1 "χειρότερες" = 0 "η" = 0 "στιγμή" = 0 "της" = 0 "σοφίας" = 0 Και ως διάνυσμα πλέον εκφράζεται με τον ακόλουθο τρόπο : $[1, 1, 1, 1, 1, 0, 0, 0, 0, 0]$. Ομοίως εκφράζονται και τα υπόλοιπα κείμενα (γραμμές).

3.4.2 Bag of n-grams

Ένα από τα κυριότερα προβλήματα της προηγούμενης μεθόδου παρουσιάζεται όταν η συλλογή των κειμένων είναι πολύ μεγάλη με συνέπεια το λεξιλόγιο και κατ' επέκταση η διάσταση των διανυσμάτων να είναι τεράστια. Αυτό οδηγεί σε αραιούς πίνακες, που περιέχουν σχεδόν αποκλειστικά μηδενικά και ελάχιστες μονάδες. Η διαχείριση αυτών των διανυσμάτων είναι ακριβή με την έννοια ότι απαιτεί πόρους όπως υπολογιστική ισχύ και μνήμη. Μια από τις προσεγγίσεις επίλυσης αυτού του προβλήματος είναι η ομαδοποίηση των λέξεων σε φράσεις μεγέθους δυο, τριών ή περισσότερων (n) λέξεων (gram).

3.4.3 Term Frequency-Inverse Term Frequency (TF-IDF)

Η τεχνική TF-IDF Qaiser and Ali (2018) σκοπό έχει να αντιμετωπίσει το πρόβλημα των λέξεων που εμφανίζονται σε πάρα πολλά κείμενα με μεγάλη συχνότητα (όπως τα άρθρα), χωρίς να προσδίδει η παρουσία τους κάποια πληροφορία για το περιεχόμενο. Ο αλγόριθμος χωρίζεται σε δύο μέρη το Term Frequency (TF) όπου υπολογίζεται η συχνότητα της λέξης στο κείμενο και το Inverse Document Frequency(IDF) όπου υπολογίζεται πόσο σπάνια εμφανίζεται η λέξη σε όλη τη συλλογή των κειμένων (corpus). Αυτοί οι υπολογισμοί διαμορφώνουν ένα βάρος για το κατά πόσο σημαντική είναι η κάθε λέξη.

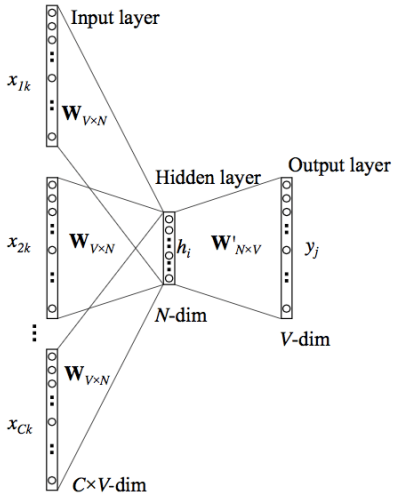
3.4.4 Word2vec

Ένας από τους πιο διαδεδομένους τρόπους αναπαράστασης λέξεων είναι ο αλγόριθμος Word2Vec Mikolov et al. (2013). Το κύριο πλεονέκτημά του σε σχέση με τις προαναφερθείσες μεθόδους είναι ότι μπορεί να "καταλάβει" τα συμφραζόμενα (context) μιας λέξης, την συντακτική και σημασιολογική της σχέση με άλλες λέξεις. Το ερώτημα είναι με ποια διαδικασία τα διανύσματα των λέξεων του αλγορίθμου Word2Vec δημιουργούνται. Ας υποθέσουμε ότι έχουμε με τις προηγούμενες μεθόδους βρει μια αναπαράσταση κάποιων λέξεων σύμφωνα με τους εξής πίνακες: $[1, 0, 0, 0, 0]$, $[0, 1, 0, 0, 0]$, $[0, 0, 1, 0, 0]$, $[0, 0, 0, 1, 0]$, $[0, 0, 0, 0, 1]$. Από τη γραμμική άλγεβρα ξέρουμε

ότι αυτά τα διανύσματα αποτελούν σημεία με συγκεκριμένη θέση στο χώρο. Συνεπώς ζητούμενο είναι λέξεις με παρόμοιο νόημα όπως "καλός" και "τέλειος" να βρίσκονται (το διάνυσμα τους) σε πιο κοντινές θέσεις από ότι λέξεις με καμία σημασιολογική σύνδεση όπως η "μέρα" και το "έχω". Μία μαθηματική μετρική για τη μέτρηση πόσο κοντά είναι δύο λέξεις είναι η ομοιότητα συνημιτόνου (cosine similarity). Όσο τείνει προς τη μονάδα τόσο πιο κοντά βρίσκονται οι λέξεις.

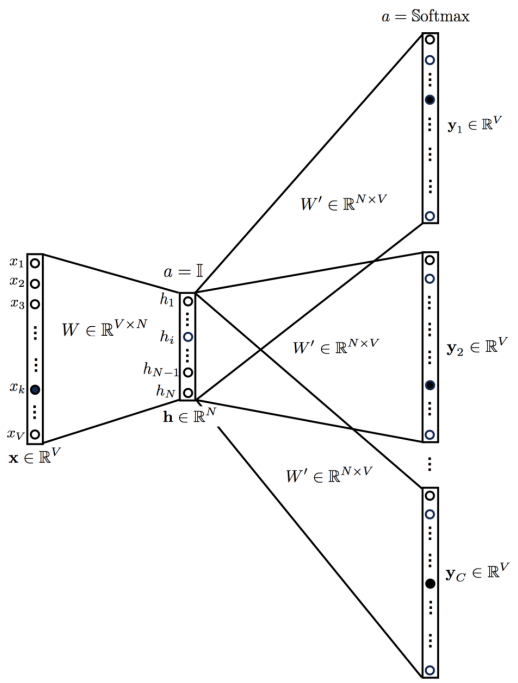
3.4.5 Λειτουργία και μηχανισμοί του αλγορίθμου Word2Vec

Το Word2Vec αποτελεί ένα νευρωνικό δίκτυο με δύο κρυφά επίπεδα το οποίο επεξεργάζεται ένα μεγάλο κείμενο για να εξάγει διανύσματα χαρακτηριστικών για τις λέξεις του κειμένου. Μια βασική διαφορά του Word2Vec με τους προγενέστερους αλγορίθμους αποτελεί η χρήση γραμμικών συναρτήσεων ενεργοποίησης στους νευρώνες του δικτύου. Λαμβάνοντας υπόψιν τα συμφραζόμενα (context) και χωρίς ανθρώπινη παρέμβαση ο αλγόριθμος Word2Vec εκπαιδεύεται από τις γειτονικές σχέσεις λέξεων, υπολογίζοντας τη συχνότητα συνεμφάνισης (co-occurance) τους και δημιουργώντας αριθμητικές αναπαραστάσεις τις οποίες προσθέτει σε ένα συμπαγές λεξιλογικό διάνυσμα. Το διάνυσμα αυτό περιέχει αριθμητικές τιμές με την μορφή πιθανοτήτων, με χρήση της συνάρτησης $softmax$, για την ομοιότητα και τη συσχέτιση μεταξύ των λέξεων. Το βασικό πλεονέκτημα του Word2Vec έναντι των άλλων αλγορίθμων για τη δημιουργία των λεξικών διανυσμαων αποτελεί η ανάλυση σχέσεων μεταξύ των λέξεων. Οι μετρικές που κυρίως χρησιμοποιούνται είναι η ομοιότητα συνιμητόνου (Cosine Similarity) αλλά και η Ευκλείδια Απόσταση. Ο αλγόριθμος χρησιμοποιεί δύο βασικές τεχνικές της Επεξεργασίας Φυσικής Γλώσσας για την δημιουργία κατανεμημένων αναπαραστάσεων των λέξεων κειμένου. Ένα για την εύρεση της λέξης των συμφραζομένων (Continuous Bag of Words (CBoW)) και ένα για την εύρεση της λέξης των συμφραζομένων με βάση τη λέξη (Skip-gram) όπως φαίνεται στο Σχήμα 3.13 και στο Σχήμα 3.14 αντίστοιχα.



Σχήμα 3.13: Μέθοδος Continuous Bag of Words (CBoW)

Ο αλγόριθμος skip-gram ουσιαστικά είναι αντίστροφος από τον CBOW με την έννοια ότι από μια λέξη προσπαθούμε να καταλάβουμε τα συμφραζόμενα. Ο αλγόριθμος δέχεται ως είσοδο μία λέξη και βγάζει ως έξοδο κατανομές πιθανότητας που περιέχουν πιθανότητες για κάθε λέξη του λεξιλογίου. Πόσο πιθανό για κάθε λέξη του λεξιλογίου να αποτελεί συμφραζόμενα της λέξης εισόδου.



Σχήμα 3.14: Μέθοδος Skip-gram

Αυτές οι δύο μέθοδοι έχουν διαφορετικά πλεονεκτήματα και μειονεκτήματα. Σύμφωνα με τον εμπνευστή της μεθόδου του Word2Vec ο αλγόριθμος Skip-gram παρουσιάζει καλύτερα αποτελέσματα σε μικρής έκτασης δεδομένα και είναι ιδιαίτερα αποδοτικός στην αναπαράσταση σπάνιων λέξεων. Από την άλλη μεριά ο αλγόριθμος CBOW είναι πιο γρήγορος και αναπαριστά καλύτερα τις πιο συχνά εμφανιζόμενες λέξεις.

3.4.6 Μέθοδος Δημιουργίας Διανυσμάτων Λέξεων GloVe (Global Vectors)

Η μέθοδος word2vec που παρουσιάστηκε σε προηγούμενο κεφάλαιο είναι πολύ αποδοτική και σε ορισμένα προβλήματα έχει δείξει εντυπωσιακά αποτελέσματα. Παρόλα αυτά, υπάρχει ένα μειονέκτημα σε σχέση με την προσέγγισή της ως προς την εύρεση των διανυσμάτων των λέξεων. Ο υπολογισμός του διανύσματος λαμβάνει υπόψιν μόνο κοντά στη ζητούμενη λέξη συμφραζόμενα. Η μέθοδος GloVe Pennington et al. (2014) μπορεί να αντιμετωπίσει αυτό το πρόβλημα καθώς είναι σχεδιασμένη κατά τέτοιο τρόπο ώστε να μπορεί να υπολογίζει ευρύτερες στατιστικές πληροφορίες για τις λέξεις όπως την συνεμφάνιση λέξεων (co-occurance) μέσα σε μια μεγάλη συλλογή κειμένων (corpus). Η βασική ιδέα για την λειτουργία αυτής της μεθόδου είναι ότι μπορείς να αντλήσεις σημασιολογική σχέση μεταξύ λέξεων από τον πίνακα συνεμφάνισης (co-occurance matrix). Πιο συγκεκριμένα, δεδομένης μιας συλλογής κειμένου η οποία διαθέτει συνολικά V λέξεις, ο πίνακας συνεμφάνισης (co-occurance matrix) θα είναι διαστάσεων $V * V$, όπου το στοιχείο X στην i γραμμή και στη j στήλη δηλώνει πόσες φορές η λέξη i εμφανίστηκε μαζί με τη λέξη j . Ένα παράδειγμα φαίνεται στο Σχήμα 3.15 όπου η πρόταση είναι η εξής “the cat sat on the mat”.

Εδώ τίθεται το εξής ερώτημα: Πώς λαμβάνουμε τη μετρική που δείχνει τη σημασιολογική ομοιότητα μέσω αυτής της διαδικασίας; Γι αυτό θα χρειαστούμε ένα παράδειγμα.

Ας ορίσουμε την οντότητα $\frac{P_{i,k}}{P_{j,k}}$ όπου $P_{i,k} = \frac{X_{i,k}}{X_i}$. Η $P_{i,k}$ δηλώνει την πιθανότητα να εμφανιστεί η λέξη i και k μαζί, και υπολογίζεται διαιρώντας τον αριθμό $X_{i,k}$ με τον συνολικό αριθμό

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Σχήμα 3.15: Παράδειγμα πίνακα συνεμφάνισης (co-occurance matrix

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Σχήμα 3.16: Παράδειγμα που δείχνει τη μέθοδο εύρεσης της σημασιολογικής ομοιότητας

που εμφανίζεται η λέξη i σε ολόκληρη τη συλλογή των κειμένων (X_i). Όπως φαίνεται από τον παραπάνω πίνακα, δεδομένων δύο λέξεων π.χ ice και $steam$, αν η τρίτη λέξη k είναι:

- αρκετά παρόμοια με τη λέξη ice αλλά άσχετη με το $steam(k = solid)$, η αναλογία $\frac{P_{i,k}}{P_{j,k}}$ θα είναι πολύ υψηλή (> 1),
- αρκετά παρόμοια με το $steam$ αλλά άσχετη με το $ice(k = gas)$, η αναλογία $\frac{P_{i,k}}{P_{j,k}}$ θα είναι πολύ μικρή (<1),
- παρόμοια ή άσχετη και με τις δύο λέξεις, τότε $\frac{P_{i,k}}{P_{j,k}}$ θα προσεγγίζει τη μονάδα

Για να δημιουργήσουμε τα διανύσματα των λέξεων πρέπει να λύσουμε τρία προβλήματα. Πρώτον, δεν γνωρίζουμε την εξίσωση παρά μόνο μία έκφραση της μορφής $F(i, j, k) = \frac{P_{i,k}}{P_{j,k}}$. Δεύτερον, τα διανύσματα των λέξεων πρέπει να έχουν αρκετές διαστάσεις ενώ η μετρική $\frac{P_{i,k}}{P_{j,k}}$ είναι βαθμωτό μέγεθος. Άρα, υπάρχει αναντιστοιχία διαστάσεων. Τρίτον, υπάρχουν τρεις οντότητες που εμπλέκονται (i, j, k) οι οποίες πρέπει να μειωθούν σε δύο για να μπορούν να είναι διαχειρίσιμες από τη συνάρτηση σφάλματος (loss function). Θα χρησιμοποιήσουμε την ακόλουθη σημειογραφία:

- w, u Δύο ξεχωριστά επίπεδα διανυσμάτων (embedding layers)
- $w*$ Ανάστροφο διάνυσμα του w
- X πίνακας συνεμφάνισης (co-occurance matrix)
- b_w και b_u πόλωση (bias) του w και του u αντίστοιχα

Για την επίλυση του προβλήματος απλά θα υποθέσουμε ότι υπάρχει μια συνάρτηση F η οποία λαμβάνει τα διανύσματα των λέξεων i j και k δίνοντας σαν έξοδο την αναλογία που ζητάμε $F(w_i, w_j, u_k) = \frac{P_{i,k}}{P_{j,k}}$. Σύμφωνα με τη δημοσίευση η δημιουργία δυο επιπέδων διανυσμάτων w και u βοηθάει το μοντέλο στο να μειώσει το πρόβλημα της υπερεκπαίδευσης (overfitting). Τα διανύσματα των λέξεων είναι γραμμικά συστήματα. Για παράδειγμα, μπορεί να πραγματοποιηθεί

η ακόλουθη πράξη w_{king}–$w_{male} + w_{female} = w_{queen}$. Γι' αυτό, θα αλλάξουμε τη μορφή της συνάρτησης ως εξής: $F(w_i$–$w_j, u_k) = \frac{P_{i,k}}{P_{j,k}}$. Αυτό το κάνουμε επειδή στις περισσότερες περιπτώσεις τα διανύσματα λέξεων χρησιμοποιούνται για τον υπολογισμό αποστάσεων .

Για την επίλυση του δεύτερου προβλήματος (αναντιστοιχία διαστάσεων) θα εισάγουμε την μαθηματική μέθοδο της αναστροφής ενός πίνακα αλλά και το γινόμενο πινάκων (dot product) . $F(w_i$–$w_j) * .u_k) = \frac{P_{i,k}}{P_{j,k}}$. Πιο αναλυτικά, αν υποθέσουμε ότι το διάνυσμα λέξης είναι διαστάσεων $D * 1$ τότε ο πίνακας $(w_i$–$w_j)*$ θα είναι $1xD$ και το γινόμενο με το u_k θα δώσει βαθμωτό μέγεθος. Ας υποθέσουμε τώρα ότι η συνάρτηση F έχει μια συγκεκριμένη ιδιότητα όπως ο ομομορφισμός (homomorphism) μεταξύ πρόσθεσης και πολλαπλασιασμού τότε η εξίσωση γίνεται $F(w_i * u_k - w_j * u_k) = \frac{F(w_i * u_k)}{F(w_j * u_k)} = \frac{P_{i,k}}{P_{j,k}}$. Μέσω της οποίας προκύπτει η σχέση $F(w_i * u_k) = P_{i,k}$. Θέτοντας την εκθετική συνάρτηση Exp ως τη συνάρτηση F έχουμε $Exp(w_i * u_k) = P_i k = \frac{X_{i,k}}{X_i}$ και άρα $w_i * u_k = \log(X_{i,k})$– $\log(X_i)$. Προσθέτοντας τις πολώσεις b_w b_u από το νευρωνικό δίκτυο έχουμε την τελική μορφή της εξίσωσης. Σε ιδανική περίπτωση τέλειων διανυσμάτων λέξεων η εξίσωση αυτή θα ήταν μηδέν. Τώρα ο στόχος είναι η ελαχιστοποίηση αυτής μέσω της συνάρτησης κόστους $J(w_i, w_j) = (w_i * u_j + bw_i + bu_j$–$log(X_i j))^2$.

3.4.7 Διανύσματα λέξεων από Γλωσσικά Μοντέλα (ELMo: Embeddings from Language Models)

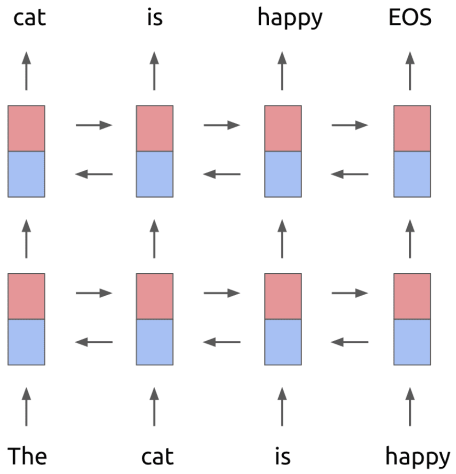
Οι προεκπαιδευμένες λεξικές αναπαραστάσεις είναι ένα στοιχείο καθοριστικής σημασίας σε πολλά μοντέλα κατανόησης γλωσσικού περιεχομένου που χρησιμοποιούν νευρωνικά δίκτυα. Γι αυτό δίνεται ιδιαίτερη έμφαση στην ποιότητα αυτών των αναπαραστάσεων. Ιδανικά, αυτές οι αναπαραστάσεις θα πρέπει να μοντελοποιούν πολύπλοκα χαρακτηριστικά μιας λέξης όπως η σύνταξη, η σημασιολογία κ.α., και πως αυτά τα χαρακτηριστικά διαφέρουν όταν αυτή η λέξη χρησιμοποιείται με διαφορετική σημασία (πολυσημία). Σε αυτήν την ενότητα θα περιγράψουμε μια μέθοδο βαθιάς αναπαράστασης που λαμβάνει υπόψιν τα συμφραζόμενα, με το όνομα ELMo (Embeddings from Language Models).

Η κύρια διαφορά αυτού του μοντέλου από τις κλασικές μεθόδους (Word2Vec, GloVe) είναι ότι κάθε λέξη που εισάγεται σε αυτό είναι μια συνάρτηση ολόκληρης της πρότασης από την οποία η λέξη περιέχεται. Τα διανύσματα προέρχονται από αμφίδρομο LSTM το οποίο έχει εκπαιδευτεί μέσω γλωσσικού μοντέλου σε τεράστια δεδομένα κειμένου. Σε αντίθεση με προηγούμενες προσεγγίσεις διανυσμάτων με συμφραζόμενα, το μοντέλο ELMo είναι βαθύ με την έννοια ότι οι αναπαραστάσεις είναι συναρτήσεις όλων των επιπέδων του γλωσσικού μοντέλου.

Πιο αναλυτικά, το ELMo μαθαίνει ένα γραμμικό συνδυασμό των διανυσμάτων της κάθε λέξης που δίνεται σαν είσοδος κάτι που βελτιστοποιεί την απόδοση σε σχέση με τη χρήση μόνο του τελευταίου επιπέδου του LSTM. Συνδυάζοντας έτσι τις εσωτερικές καταστάσεις με τον παραπάνω τρόπο, το μοντέλο EMLo δημιουργεί πολύ πλούσιες λεξικές αναπαραστάσεις. Επίσης, όπως αναφέρεται και στην επιστημονική εργασία που εισάγει αυτό το μοντέλο Peters et al. (2018), τα ανώτερα επίπεδα του LSTM συγκρατούν πληροφορία σε σχέση με συμφραζόμενα, ενώ χαμηλότερα επίπεδα μοντελοποιούν συντακτικά στοιχεία των λέξεων. Αυτή η διαπίστωση, η οποία έχει τεκμηριωθεί μέσω πειραμάτων είναι ιδιαιτέρως χρήσιμη γιατί επιτρέπει στα ήδη προεκπαιδευμένα μοντέλα να εστιάσουν στα επίπεδα του μοντέλου για το ειδικό πρόβλημα που πρέπει να επιλύσουν την εκάστοτε φορά.

Παρακάτω γίνεται εκτενής αναφορά στη λειτουργία και τα βασικά πλεονεκτήματα του μοντέλου ELMo.

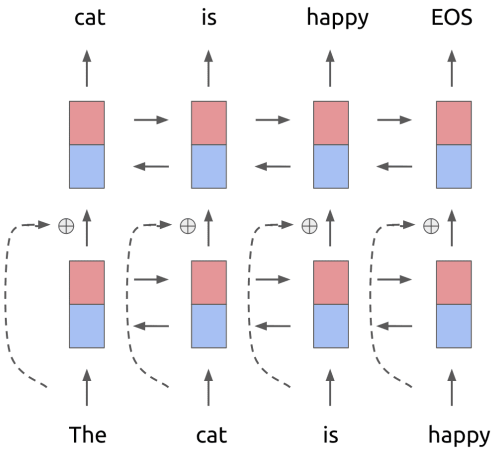
Η αρχιτεκτονική του μοντέλου ELMo ξεκινάει με την εκπαίδευση ενός αρκετά εκλεπτυσμένου νευρωνικού δικτύου γλωσσικού μοντέλου, το οποίο έχει εμπνευστεί από προηγούμενες επιστημονικές εργασίες σε μεγάλης κλίμακας γλωσσικά μοντέλα. Με τον όρο γλωσσικό μοντέλο εννοούμε μια κατανομή πιθανότητας πάνω σε μια ακολουθία λέξεων. Δοσμένης αυτής της ακολουθίας, έστω μήκους m, το γλωσσικό μοντέλο δίνει μια πιθανότητα $P(w_1, ..., w_m)$ σε όλη την ακολουθία λαμβάνοντας υπόψιν λέξεις που προηγήθηκαν. Τέτοια μοντέλα επιτρέπουν να καθοριστεί αν για παράδειγμα η φράση $I\ am\ going\ to\ write\ with\ a$, η λέξη $pencil$ είναι πιο πιθανή από ότι η λέξη $frog$. Για τους σκοπούς του ELMo το γλωσσικό μοντέλο που χρησιμοποιήθηκε ξεκίναγε με ένα δύο επιπέδων αμφίδρομο LSTM δίκτυο το οποίο φαίνεται στο Σχήμα 3.17.



Σχήμα 3.17: Ένα αμφίδρομο LSTM δύο επιπέδων. Τα κόκκινα τετράγωνα αναπαριστούν το εμπρόσθιο επαναλαμβανόμενο τμήμα και τα γαλάζια το οπίσθιο επαναλαμβανόμενο τμήμα.

Έπειτα, σε αυτά το δύο επιπέδων δίκτυο προστίθεται μια υπολειπόμμενη (residulal) He et al. (2015) σύνδεση μεταξύ πρώτου και δεύτερου επιπέδου. Το κίνητρο για αυτή την προσθήκη έγκειται στο γεγονός ότι αυτού του τύπου οι συνδέσεις βοηθούν τα μοντέλα βαθιών νευρωνικών δικτύων να εκπαιδευτούν πιο αποτελεσματικά. Το γλωσσικό μοντέλο πλέον προσομοιώνεται στο Σχήμα 3.18

Στα παραδοσιακά νευρωνικά γλωσσικά μοντέλα, κάθε λέξη στο αρχικό επίπεδο εισόδου σε αυτή την περίπτωση στο Σχήμα 3.18 $The\ cat\ is\ happy$ θα μετασχηματιζόταν σε ένα προδιαγεγραμένου μήκους διάνυσμα λέξης πριν περαστεί στο επαναλαμβανόμενο τμήμα (του LSTM). Αυτό θα γινόταν είτε αρχικοποιώντας ένα διάνυσμα λέξης διαστάσεων (μήκος λεξικού) επί (Διάσταση του διανύσματος λέξης) είτε με τη χρήση προεκπαιδευμένων διανυσμάτων όπως το GloVe. Παρόλα αυτά,για το γλωσσικό μοντέλο του Elmo γίνεται κάτι πιο πολύπλοκο. Αντί να αναζητούμε απλά ένα διάνυσμα λέξης από τον ολικό πίνακα των διανυσμάτων, μετασχηματίζεται αρχικά κάθε λέξη σε μια αναπαράσταση χρησιμοποιώντας διανύσματα χαρακτήρων (character embeddings). Αυτά τα διανύσματα χαρακτήρων περνούν από ένα συνελικτικό επίπεδο Krizhevsky et al. (2012) χρησιμοποιώντας ένα αριθμό από φίλτρα και ένα max-pool επίπεδο. Τέλος, αυτή η αναπαράσταση διοχετεύεται σε ένα highway δίκτυο Srivastava et al. (2015) και έπειτα παρέχεται στο επίπεδο του LSTM. Η διαδικασία αυτή ακολουθείται για κάθε λέξη βλ. Σχήμα 3.19



Σχήμα 3.18: Η προσθήκη της υπολλειπόμμενης (residual) σύνδεσης μεταξύ πρώτου και δεύτερου επιπέδου του LSTM. Η είσοδος του πρώτου επιπέδου προστίθεται απευθείας στην έξοδό του πριν περαστεί σαν είσοδος στο δεύτερο επίπεδο.

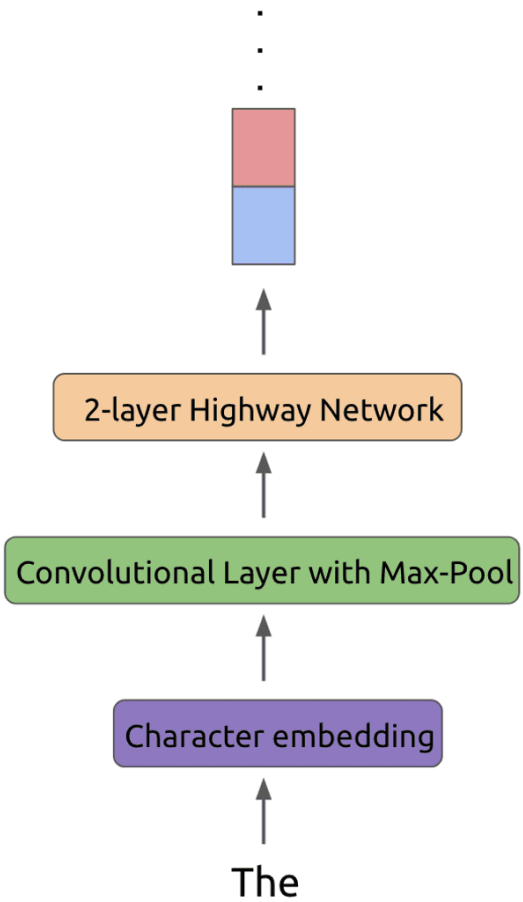
Αυτοί οι μετασχηματισμοί που χρησιμοποιούνται σε κάθε λέξη πριν αυτή εισέλθει στο LSTM παρέχουν μια σειρά από πλεονεκτήματα. Πρώτα από όλα, η χρήση των διανυσμάτων χαρακτήρων επιτρέπει να αναλυθούν μορφολογικά χαρακτηριστικά των λέξεων κάτι που διανύσματα σε επίπεδο λέξεων αδυνατούν να εντοπίσουν. Επίσης, η χρήση των διανυσμάτων χαρακτήρων εξασφαλίζει ότι θα δημιουργηθούν ορθές αναπαραστάσεις λέξεων που δεν περιέχονται στο λεξιλόγιο των προεκπαιδευμένων πινάκων. Τα συνελικτικά φίλτρα από την άλλη μεριά, βοηθούν στο να εντοπιστούν χαρακτηριστικά από το μοντέλο των n-gram, χτίζοντας έτσι πιο ισχυρές αναπαραστάσεις. Το highway δίκτυο επιτρέπει ομαλότερη μεταφορά της πληροφορίας στην είσοδο. Αυτές είναι οι κύριες διαφορές και τα σημεία κλειδιά που ξεχωρίζουν το μοντέλο ELMo από τα άλλα γλωσσικά μοντέλα.

Αξίζει επίσης να αναφέρουμε μεγάλο πλεονέκτημα αυτής της μεθόδου είναι στο πως το χρησιμοποιούμε τελικά αφού έχει εκπαιδευτεί. Πιο συγκεκριμένα, ας υποθέσουμε ότι κοιτάμε στην k-οστή λέξη. Χρησιμοποιώντας το εκπαιδευμένο 2 επιπέδων γλωσσικό μοντέλο, παίρνουμε τη λεξική αναπαράσταση x_k καθώς και την αμφίδρομη αναπαράσταση των κρυφών επιπέδων h_1, k και h_2, k και τις συνδυάζουμε σε μια νέα (προσαρμοσμένη με τα επιμέρους βάρη) αναπαράσταση όπως φαίνεται στο Σχήμα 3.20. Σε αυτό το Σχήμα δεν φαίνεται το πολύπλοκο δίκτυο για τη δημιουργία της λεξικής αναπαράστασης h_k το οποίο αναλύσαμε εκτενώς προηγουμένως.

Για να είμαστε πιο ακριβείς από μαθηματική άποψη, η συνάρτηση f που δημιουργεί τα βάρη για τη λέξη k και φαίνεται στο Σχήμα 3.20 περιγράφεται ως εξής:

$$ELMo_k = \gamma(s_0 * x_k + s_1 * h_{1,k} + s_2 * h_{2,k}) \tag{3.4}$$

Όπου το s_i προσδιορίζει τα βάρη των αναπαραστάσεων των κρυφών επιπέδων του γλωσσικού μοντέλου και το γ_k προσδιορίζει μια παράμετρο κλίμακας που καθορίζεται για κάθε συγκεκριμένο πρόβλημα στο οποίο εφαρμόζεται το ELMo. Αξίζει επίσης να σημειωθεί ότι για κατηγορία προβλημάτων (π.χ ερώτηση-απάντηση, σημασιολογική ανάλυση, αναγνώριση οντοτήτων) χρησιμοποιείται διαφορετική αναπαράσταση του ELMo. Πιο συγκεκριμένα, για να χρησιμοποιήσουμε σε ένα πρόβλημα, πρέπει πρώτα να "παγώσουμε" τα βάρη του εκπαιδευμένου γλωσσικού μοντέλου



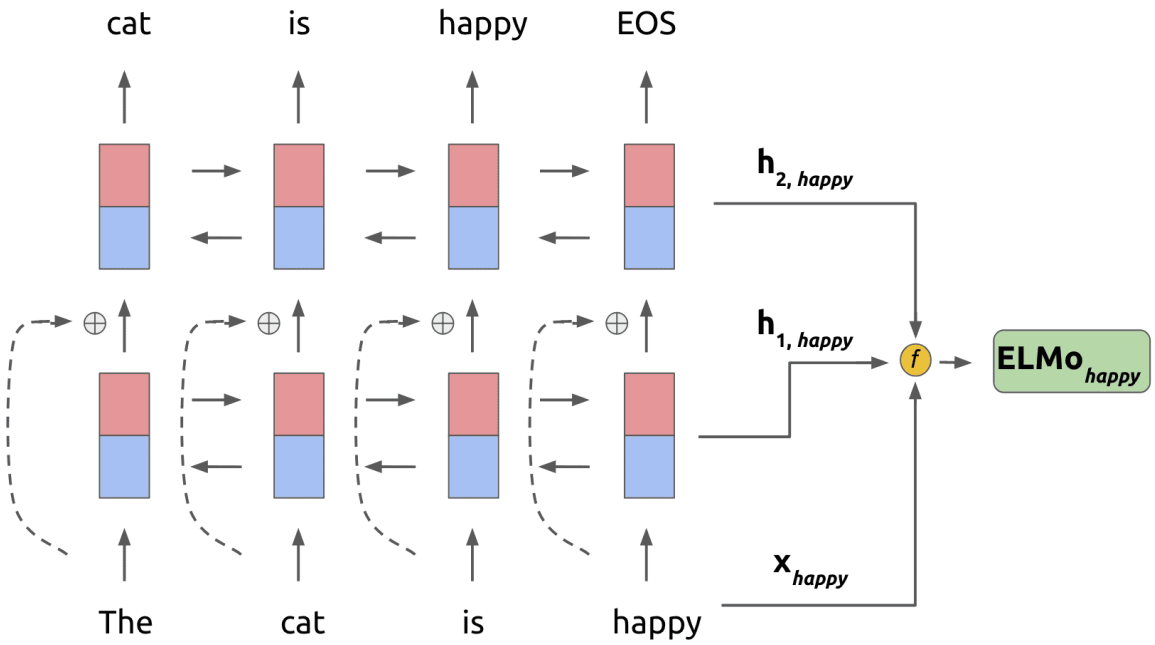
Σχήμα 3.19: Μετασχηματισμός που εφαρμόζεται σε κάθε λέξη πριν αυτή διοχετευθεί σαν είσοδος στο πρώτο επίπεδο LSTM. Εδώ φαίνεται ενδεικτικά η διαδικασία για τη λέξη the.

Σύνολο δεδομένων/Κατηγορία προβλήματος	Pre ELMo	After ELMo
SQuAD (question/answering)	84.4	85.8
SNLI (textual entailment)	88.6	88.7
Semantic Role Labelling	81.7	84.6
Coref Resolution	67.2	70.4
NER	91.93	92.22
SST-5 (sentiment analysis)	53.7	54.7

Πίνακας 3.1: Αποτελέσματα χρήσης του μοντέλου Elmo σε διαφορετικές κατηγορίες προβλημάτων στον τομέα Επεξεργασίας Φυσικής Γλώσσας

και μετά να συνενώσουμε τα αποτελέσματα $ELMo_k$ της κάθε λέξης με την αναπαράσταση της εισόδου του εκάστοτε συγκεκριμένου προβλήματος (task specific). Οι παράμετροι γ_k και s_i μαθαίνονται κατά τη διάρκεια της εκπαίδευσης επίσης του εκάστοτε συγκεκριμένου προβλήματος.

Τα πειραματικά αποτελέσματα του πίνακα 3.1 αποδεικνύουν πόσο χρήσιμο εργαλείο είναι το μοντέλο ELMo σχεδόν στο σύνολο των βασικών κατηγοριών προβλημάτων στον τομέα Επεξεργασίας Φυσικής Γλώσσας.



Σχήμα 3.20: Ένα παράδειγμα συνδυασμού των αναπαραστασέων των κρυφών επιπέδων και της λεξικής αναπαράστασης έτσι ώστε η λέξη "happy" να αποκτήσει την αναπαράστασή της στο μοντέλο ELMo.

Κεφάλαιο 4

Αρχιτεκτονικές Μοντέλων Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας

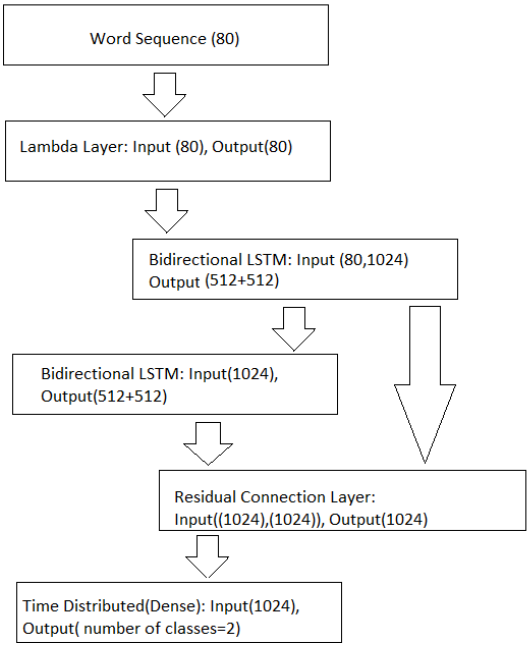
4.1 Ανίχνευση Προπαγανδιστικών Κομματιών

4.1.1 Λογιστική Παλινδρόμηση

Κατά τη διάρκεια της παρούσας διπλωματικής εργασίας υλοποιήθηκαν τρεις κύριες αρχιτεκτονικές για την αντιμετώπιση του προβλήματος της αυτοματοποιημένης ανίχνευσης προπαγανδιστικών σχολίων σε ειδησεογραφικά άρθρα το οποίο για συντομία θα το αναφέρουμε ως ΠΤΔ (Πρόβλημα Ταυτοποίησης Διαστήματος). Αρχικά, η πρώτη προσέγγιση στο πρόβλημα αποτελείται από τον ευρέως διαδεδομένο αλγόριθμο μηχανικής μάθησης Λογιστική Παλινδρόμηση. Για το μοντέλο αυτό χρησιμοποιούμε στην είσοδο προ-εκπαιδευμένα διανύσματα λέξεων (Word Embeddings) από το μοντέλο Word2Vec τα οποία τροφοδοτούν το μοντέλο της Λογιστικής Παλινδρόμησης Logistic Regression) με τη διανυσματική αναπαράσταση των λέξεων των κειμένων. Αυτό το μοντέλο χρησιμοποιήθηκε σαν βάση σύγκρισης για τα επόμενα μοντέλα που περιέχουν πιο σύγχρονες μεθόδους όπως βαθιά νευρωνικά δίκτυα.

4.1.2 Αμφίδρομο LSTM

Το πρώτο μοντέλο όπως αναφέρθηκε αποτελείται από μία αρκετά απλή αρχιτεκτονική η οποία έχει αρκετά μειονεκτήματα που δεν επιτρέπουν την επίτευξη ικανοποιητικών αποτελεσμάτων. Ένα κύριο μειονέκτημα είναι η απώλεια πολλών λέξεων που δεν αντιστοιχίζονται σε κάποιο από τα προ-εκπαιδευμένα διανύσματα λόγω της μη-εκπαίδευσης των διανυσμάτων λέξεων στα δεδομένα του προβλήματός μας. Επίσης, η ακολουθιακή φύση των δεδομένων κειμένου συστήνει την χρήση Ακολουθιακών Νευρωνικών Δικτύων στο πρόβλημα. Λαμβάνοντας υπόψιν αυτά τα μειονεκτήματα δημιουργήσαμε μια δεύτερη αρχιτεκτονική η οποία αποτελείται από αμφίδρομα LSTM νευρωνικά δίκτυα και προ-εκπαιδευμένα διανύσματα λέξεων από το μοντέλο GloVe. Η επιλογή του συστήματος GloVe για την αναπαράσταση των λέξεων στο διανυσματικό χώρο έναντι του μοντέλου Word2Vec έγκειται στο γεγονός ότι το GloVe, όπως αναφέρεται και στην ενότητα 6.7 του κεφαλαίου 3, δημιουργεί αναπαραστάσεις των λέξεων από τα συμφραζόμενα σε μεγαλύτερο εύρος του κειμένου από ότι το μοντέλο Word2Vec. Ακόμα, αυτά τα διανύσματα της βιβλιοθήκης GloVe εκπαιδεύτηκαν στη συστοιχία των κειμένων που μας δόθηκαν, έτσι ώστε να αποφευχθεί η απώλεια λέξεων. Όπως αναμενόταν αυτή η αρχιτεκτονική ξεπέρασε σε απόδοση το βασικό πρώτο μοντέλο ακόμα και με όχι ιδιαίτερα διεξοδική ρύθμιση των υπερπαραμέτρων του δικτύου.



Σχήμα 4.1: Σχηματική αναπαράσταση του τελικού μοντέλου για το υπο-πρόβλημα της Ανίχνευσης που περιλαμβάνει αμφίδρομο residual LSTM και το μοντέλο ELMo

4.1.3 Αμφίδρομο residual LSTM

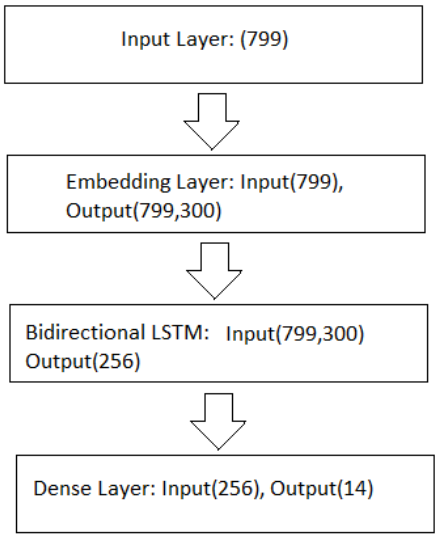
Το τελικό μοντέλο αποτελείτε από δύο επίπεδα αμφίδρομων LSTM και από διανύσματα λέξεων από το γλωσσικό μοντέλο ELMo. Οι αναπαραστάσεις των λέξεων που προκύπτουν από τον ElMo αλγόριθμο έχουν ένα πολύ σημαντικό πλεονέκτημα σε σχέση με αυτές από το Word2Vec και από το GloVe γιατί λαμβάνει υπόψιν τα μορφολογικά χαρακτηριστικά της κάθε λέξης αλλά και τα συμφραζόμενα (context). Αυτές οι αναπαραστάσεις είναι συναρτήσεις όχι μόνο της ίδιας της λέξης αλλά και του πλαισίου μέσα στο οποίο αυτή χρησιμοποιείται, επιτρέποντας έτσι την αποσαφήνιση των λέξεων σε διαφορετικές σημασιολογικές αναπαραστάσεις.

Για παράδειγμα, το ELMo μοντέλο για τη λέξη milk δημιουργεί διαφορετικό διάνυσμα αναπαράστασης για αυτή τη λέξη όταν χρησιμοποιείται με τη σημασία γάλα και άλλη αναπαράσταση για την ίδια λέξη όταν χρησιμοποιείται ως ρήμα με την έννοια ότι κάποιος εξαπατά και παίρνει χρήματα από κάποιον άλλο. Το Σχήμα 4.1 αναπαριστά την αρχιτεκτονική του μοντέλου το οποίο αποτελείται από δύο επίπεδα με Αμφίδρομα LSTM που στόχο έχουν να κάνουν την πρόβλεψη (για το αν η κάθε λέξη της δοσμένης ακολουθίας συνιστά προπαγανδιστικό σχόλιο ή όχι) χρησιμοποιώντας για κάθε λέξη τις προηγούμενες λέξεις αλλά και τις επόμενες λέξεις ως πληροφορία (context information).

Το Αμφίδρομο LSTM είναι συνδυασμός δύο LSTMs όπου το ένα ξεκινά να επεξεργάζεται προγενέστερα τμήματα της ακολουθίας και το άλλο μεταγενέστερα, διασχίζουν δηλαδή την ακολουθία ταυτόχρονα με αντίθετη φορά. Σημαντική επίσης είναι η προσθήκη σε σχέση με την προηγούμενη αρχιτεκτονική της υπολειπόμενης (residual) σύνδεσης μεταξύ του πρώτου αμφίδρομου LSTM και της εξόδου του δεύτερου που επιτρέπει την προσπέραση της εκπαίδευσης των επιπέδων που δεν είναι χρήσιμα και δεν προσθέτουν στην καλύτερη απόδοση του μοντέλου. Το σημείο κλειδί της τελικής αρχιτεκτονικής είναι η προσθήκη του ELMo μοντέλου. Χάρη σε αυτό δεν χρειάζεται να δημιουργήσουμε χαρακτηριστικά από μόνοι μας παρά μόνο να εισάγουμε τις ακολουθίες με τις

ετικέτες (προπαγάνδα, μη-προπαγάνδα) σε επίπεδο λέξεων στο μοντέλο και να το εκπαιδεύσουμε στα δεδομένα κειμένου αξιοποιώντας παράλληλα το γεγονός ότι έχει γίνει ήδη εκπαίδευση του ELMo σε τεράστιου όγκου δεδομένα. Η μεταφορά αυτή του προ-εκπαιδευμένου μοντέλου (Transfer Learning) είναι καταλυτική στην απόδοση αλλά και στη πρακτικότητα της εφαρμογής αφού χωρίς αυτή θα χρειαζόταν τεράστια υπολογιστική ισχύς για την αποτελεσματική εκπαίδευση του μοντέλου από μηδενική βάση.

Τέλος, το ELMo επεξεργάζεται την πληροφορία των λέξεων σε επίπεδο χαρακτήρων κάτι που επιτρέπει στο νευρωνικό δίκτυο να εκμεταλλευτεί μορφολογικά χαρακτηριστικά και αναπαραστάσεις των λέξεων.



Σχήμα 4.2: Σχηματική αναπαράσταση του μοντέλου Κατηγοριοποίησης με χρήση αμφίδρομου LSTM και διανύσματα λέξεων από το μοντέλο GloVe

4.2 Κατηγοριοποίηση προπαγανδιστικών τεχνικών

Το δεύτερο πρόβλημα που καλούμαστε να λύσουμε στην παρούσα πτυχιακή εργασία είναι η κατηγοριοποίηση προπαγανδιστικών τεχνικών. Πιο αναλυτικά, δοθέντων αποσπασμάτων των κειμένων που έχουν ήδη χαρακτηριστεί ως προπαγανδιστικά πρέπει αυτοματοποιημένα να αποφανθούμε σε ποια κατηγορία προπαγάνδας ανήκει το κάθε απόσπασμα. Για παράδειγμα δύο από τις κατηγορίες προπαγάνδας είναι η φορτισμένη γλώσσα και η επίκληση στο φόβο. Παρόμοια με το πρόβλημα ανίχνευσης προπαγάνδας δημιουργήσαμε σταδιακά τρία μοντέλα.

Η πρώτη προσέγγιση αποτελείται από τον αλγόριθμο της Λογιστικής παλινδρόμησης και διανύσματα λέξεων από το μοντέλο Wor2Vec. Αξίζει να επισημανθεί ότι σε αυτό το πρόβλημα είχε δοθεί ένας βασικός αλγόριθμος από τους διοργανωτές του διαγωνισμού τον οποίο έπρεπε να ξεπεράσουμε ως προς την απόδοση. Το μοντέλο της Λογιστικής παλινδρόμησης κατάφερε να έχει υπερδιπλάσια ποσοστά από αυτό το βασικό αλγόριθμο. Τα αποτελέσματα και των τριών μοντέ-

λων που δημιουργήσαμε για αυτό το πρόβλημα αναφέρονται στον Πίνακα 5.3. Για να βελτιώσουμε περαιτέρω τα αποτελέσματα της κατηγοριοποίησης χρησιμοποιήσαμε έναν από τους πιο διαδεδομένους, για το συγκεκριμένο είδος προβλήματος (multiclass classification), αλγόριθμο Μηχανικής Μάθησης το Multilayer Perceptron (MLP) . Ο πίνακας χαρακτηριστικών για το MLP προκύπτει και πάλι από τη αντιστοίχηση των λέξεων των ακολουθιών στο διανυσματικό χώρο μέσω των προ-εκπαιδευμένων διανυσμάτων του Word2Vec. Τέλος, για να εκμεταλλευτούμε την ακολουθιακή φύση των δεδομένων χρησιμοποιήσαμε για το τρίτο και τελικό μοντέλο ένα αμφίδρομο LSTM και εκπαιδεύσαμε εκ νέου τα διανύσματα από τη βιβλιοθήκη του Glove μέσω του επιπέδου διανυσμάτων (Embedding Layer) του νευρωνικού μας δικτύου. Η αρχιτεκτονική του και τα επίπεδα που αυτό περιέχει αναπαριστώνται στο Σχήμα 4.2.

4.3 Παρατηρήσεις ως προς την Υλοποίηση

Οι αρχιτεκτονικές που παρουσιάστηκαν στις παραπάνω ενότητες αποτελούν τους βασικούς αλγορίθμους για την αντιμετώπιση του προβλήματος της Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας. Η απλοποιημένη σχηματική αναπαράστασή τους καθώς και η σύντομη περιγραφή της αρχιτεκτονικής τους δεν διατυπώνει πλήρως κάποια χαρακτηριστικά της υλοποίησης που δεν αναπτύχθηκαν εκτενώς. Παρακάτω γίνεται μια προσπάθεια αποσαφήνισης ορισμένων από αυτά τα στοιχεία.

Για το Πρόβλημα Ταυτοποίησης Διαστήματος (τρίτο και τελικό μοντέλο):

- Χρησιμοποιήθηκε 10% dropout σε όλα τα κρυφά επίπεδα
- Και τα δύο επίπεδα Αμφίδρομων LSTM περιέχουν 512 στοιχεία
- Το πλήθος των χαρακτηριστικών για την αναπαράσταση των λέξεων από το ELMo είναι 1024 (προεπιλογή)
- Για τη βελτιστοποίηση χρησιμοποιείται ο Adam optimizer με τις προεπιλεγμένες παραμέτρους και η sparse categorical crossentropy ως συνάρτηση κόστους (Loss funcion).
- Το μέγεθος του κομματιού (Batch) επιλέχθηκε να είναι 32 και η εκπαίδευση κράτησε 8 εποχές.

Για το Πρόβλημα Κατηγοριοποίησης Τεχνικών:

- Χρησιμοποιήθηκε 10% dropout σε όλα τα κρυφά επίπεδα
- Το επίπεδο Αμφίδρομου LSTM περιέχει 128 στοιχεία
- Το πλήθος των χαρακτηριστικών για την αναπαράσταση των λέξεων από το μοντέλο GloVe είναι 300
- Το μέγεθος του κομματιού (Batch) επιλέχθηκε να είναι 32 και η εκπαίδευση κράτησε 4 εποχές.

Κεφάλαιο 5

Επισκόπηση του Συστήματος Ανίχνευσης και Κατηγοριοποίησης Προπαγάνδας

5.1 Δεδομένα

Το πρόβλημα της ανίχνευσης ψευδών ειδήσεων, εξτρεμιστικών σχολίων και προπαγάνδας αποτελεί ένα ιδιαίτερα δύσκολο πρόβλημα με το οποίο όπως αναφέρθηκε και στο κεφάλαιο 2 η επιστημονική κοινότητα έχει ασχοληθεί αρκετά τα προηγούμενα χρόνια. Σε αυτό το διαγωνισμό γίνεται πρώτη φορά η προσπάθεια να ανιχνευθεί η προπαγάνδα όχι σε επίπεδο κειμένου, δηλαδή να αποφανεί αν το κείμενο περιέχει προπαγανδιστικά σχόλια, αλλά να εντοπιστούν τα σημεία αυτά του κειμένου (αν υπάρχουν) και να κατηγοριοποιηθούν. Για το σκοπό αυτό οι διοργανωτές δημιούργησαν στο πλαίσιο του Propaganda Analysis Project ένα σύνολο 451 κειμένων από 48 διαφορετικές δημοσιογραφικές πηγές. Λόγω του μεγάλου όγκου αυτών των δεδομένων η εύρεση των ζητούμενων προπαγανδιστικών αποσπασμάτων χρειαζόταν αφενός ιδιαίτερα πολύ χρόνο και αφετέρου κατανόηση όλων των μορφών προπαγάνδας.

Έτσι αυτή η διαδικασία ανατέθηκε στην εταιρία Data Pro η οποία έχει επαγγελματική εμπειρία σε αυτό τον τομέα. Ως προς τα δεδομένα των κειμένων παρακάτω αναφέρουμε αρκετά χρήσιμα στατιστικά. Αρχικά, ο συνολικός αριθμός των προπαγανδιστικών τεχνικών που βρέθηκαν στα άρθρα είναι $7,485$ σε $21,230$ προτάσεις (35.2%). Ο Πίνακας 5.1 δίνει στατιστικά για όλες τις μορφές προπαγάνδας ξεχωριστά. Ακόμα, κατά μέσο όρο τα προπαγανδιστικά κομμάτια έχουν μήκος 47 χαρακτήρες και ο μέσος όρος του μήκους των προτάσεων είναι 112.5 χαρακτήρες, δηλαδή το προπαγανδιστικό κομμάτι καταλαμβάνει περίπου τη μισή πρόταση. Η πιο συχνή μορφή προπαγάνδας είναι Φορτισμένη Γλώσσα (Loaded Language) και Επίκληση σε Όνομα (Name Calling, Labelling) καλύπτοντας $2,547$ και $1,294$ περιπτώσεις αντίστοιχα. Εμφανίζονται 6.7 και 4.7 φορές ανά άρθρο ενώ καμία άλλη τεχνική δεν εμφανίζεται πάνω από δύο φορές κατά μέσο όρο. Παρακάτω θα εισάγουμε τους ορισμούς των 18 μορφών προπαγάνδας τις οποίες το μοντέλο μας καλείται να ταξινομήσει. Για περισσότερες πληροφορίες σχετικά με τις κατηγορίες παραθέτουμε την δημοσίευση Da San Martino et al. (2019e):

- **Φορτισμένη Γλώσσα (Loaded Language)** Χρήση λέξεων/φράσεων με μεγάλες συναισθηματικές επιπτώσεις (θετικές ή αρνητικές) με σκοπό να επηρεάσει ένα πλήθος ανθρώπων Weston (2018)

- **Επίκληση σε Όνομα (Name Calling or Labelling)** Αναφορά στο αντικείμενο της προπαγάνδας ως κάτι που το κοινό φοβάται, μισεί ή αγαπάει, επαινεί. Miller (1939)

- **Επανάληψη (Repetition** Συνεχής επανάληψη του ίδιου μηνύματος, με σκοπό το κοινό τελικά να το αποδεχθεί Miller (1939); Torok (2015)

- **Υπερβολή ή Ελαχιστοποίηση (Exaggeration or minimization)** Είτε παρουσιάζεις κάτι σε υπερβολική μορφή κάνοντας το να φαίνεται μεγαλύτερο, καλύτερο, χειρότερο είτε προσπαθείς να δείξεις ότι κάτι έχει λιγότερη σημασία π.χ λέγοντας ότι η προσβολή ήταν μικρής σημασίας. Jowett and O'Donnell (2012)

- **Αμφιβολία (Doubt)** Προκύπτει αμφισβητώντας τη φερεγγυότητα κάποιου ανθρώπου ή γεγονότος

- **Επίκληση στο φόβο/προκατάληψη (Appeal to fear/prejudice)** Προσπάθεια να αποκτήσεις υποστήριξη για μια ιδέα μέσω ενστάλαξης συναισθημάτων πανικού/ανησυχίας για οποιαδήποτε άλλη εναλλακτική

- **Flag-waving** Δόλια εκμετάλλευση των εθνικών ευαισθησιών (ή σε σχέση με συγκεκριμένη ομάδα πχ ευαισθησίες περί φυλής, πολιτικής ιδεολογίας) προκειμένου να δικαιολογηθεί ή να προωθηθεί κάποια πράξη ή ιδέα. Hobbs (2014)

- **Κοινή υπεραπλούστευση (Casual oversimplification)** Υποθέτοντας ότι υπάρχει μία αιτία όταν υπάρχουν πολλές πίσω από ένα θέμα. Επίσης, εδώ συμπεριλαμβάνεται η προσπάθεια να πέσει το φταίξιμο για ένα γεγονός σε ένα άτομο ή μια ομάδα ανθρώπων χωρίς να έχει διερευνηθεί αν αυτό είναι αλήθεια.

- **Συνθήματα (Slogans)** Χρήση μικρών και αναγνωρίσιμων φράσεων που συχνά περιέχουν ονόματα και στερεότυπα. Dan (2015)

- **Επίκληση Σε Αυθεντία (Appeal to authority)** Η αυτόματη αποδοχή ότι κάτι ισχύει επειδή το είπε κάποιος (υποτιθέμενα) ειδικός χωρίς όμως να το τεκμηρίωσει επαρκώς. Goodwin and McKerrow (2011); Jowett and O'Donnell (2012)

- **Άσπρο ή Μαύρο (Black-and-white fallacy, dictatorship)** Παραθέτοντας δύο εναλλακτικές ως τις μόνες πιθανές επιλογές, όταν όμως υπάρχουν κι άλλες. Σε ακραία περίπτωση, η επιβολή στο κοινό μιας πρότασης, χωρίς περιθώρια επιλογής (δικτατορία). Torok (2015)

- **Κλισέ (Thought-terminating cliché)** Αποτελούνται από λέξεις ή φράσεις που αποθαρρύνουν την κριτική σκέψη και την εποικοδομητική συζήτηση σε ένα θέμα. Hunter (2015)

- **Whataboutism** Δημιουργία κακών εντυπώσεων για τη θέση ενός αντιπάλου κατηγορώντας τον απλά ως υποκριτή, χωρίς αντιπαράθεση με επιχειρήματα. Richter (2017)

- **Reductio ad Hitlerum** Προσπάθεια να πειστεί ένα συγκεκριμένο κοινό ότι μια πράξη ή ιδέα δεν είναι ορθή επειδή είναι δημοφιλής από μια άλλη μισητή ομάδα ανθρώπων. Teninbaum (2009)

- **Αποπροσανατολισμός (Red herring)** Εισαγωγή άσχετου με το θέμα υλικού έτσι ώστε να αποπροσανατολιστεί η προσοχή του αναγνώστη. Teninbaum (2009); Weston (2018)

- **Bandwagon** Προσπάθεια να πειστεί η ομάδα-στόχος να πάρει μέρος επειδή και "όλοι οι άλλοι κάνουν το ίδιο". Hobbs (2014)

- **Δόλια ασάφεια (Obfuscation, intentional vagueness, confusion)** Δόλια χρήση ασαφών λέξεων, προκειμένου το κοινό να κάνει δικιά του ερμηνεία. Για παράδειγμα, όταν μια ασαφής φράση με πολλαπλές πιθανές ερμηνείες χρησιμοποιείται σε ένα επιχείρημα και κατά συνέπεια δεν μπορεί να υποστηρίξει στα αλήθεια το συμπέρασμα. Suprabandari (2007)

- **Straw men** Όταν η άποψη του αντιπάλου αντικαθίσταται με μια παρόμοια η οποία μετά διαψεύδεται. Walton (1996)

Propaganda Technique	#Train	(%)	F1(Dev)	F1(Test)
Bandwagon, Reductio ad Hitlerum	72	1.17	0.44	0.09
Whataboutism, Straw Men, Red Herring	108	1.76	0.08	0.04
Causal Oversimplification	209	3.41	0.17	0.00
Doubt	493	8.04	0.51	0.48
Appeal to authority	144	2.34	0.06	0.36
Black-and-white Fallacy, Dictatorship	107	1.74	0.07	0.00
Name calling or labeling	1058	17.26	0.59	0.54
Loaded Language	2123	34.63	0.69	0.65
Exaggeration or Minimisation	466	7.60	0.43	0.29
Flag-waving	229	3.73	0.73	0.46
Appeal to fear/prejudice	294	4.79	0.21	0.28
Slogans	129	2.10	0.27	0.13
Thought-terminating cliché	76	1.24	0.00	0.11
Repetition	621	10.13	0.27	0.18
overall(our team:NTUAAILS)	6129		0.53	0.46
overall (best team:ApplicaAI)			0.70	0.62

Πίνακας 5.1: Στατιστικά στοιχεία που αφορούν τις προπαγανδιστικές τεχνικές και τα αποτελέσματα στα σύνολα test και dev για το πρόβλημα Κατηγοριοποίησης

5.2 Μετρικές Αξιολόγησης

5.2.1 Ανίχνευση Προπαγανδιστικών Τμημάτων (SI task)

Το SI task αποτελεί το κομμάτι της αναγνώρισης των σημείων του κειμένου. Η συνάρτηση αξιολόγησης λαμβάνει υπόψιν μερική επικάλυψη ανάμεσα σε δύο προπαγανδιστικές μεθόδους. Με λίγα λόγια, το βάρος που δίνεται σε αυτή τη μερική επικάλυψη είναι ανάλογο του εύρους στο οποίο επικαλύπτονται. Αυτό το βάρος κανονικοποιείται επίσης, σύμφωνα με το μήκος τους. Παρακάτω γίνεται εκτενής περιγραφή αυτού του τρόπου αξιολόγησης.

Έστω ότι έχουμε ένα κείμενο d το οποίο αναπαριστάται ως μια ακολουθία χαρακτήρων. Το i-στό προπαγανδιστικό κομμάτι αναπαριστάται επίσης ως μια ακολουθία συνεχόμενων χαρακτήρων $t \subseteq d$. Ένα κείμενο περιέχει ένα σύνολο (πιθανώς επικαλυπτόμενων) προπαγανδιστικών κομματιών T. Ένας αλγόριθμος μάθησης παράγει ένα σύνολο S με κομμάτια $s \subseteq d$. Η συνάρτηση που δίνει τις ετικέτες $l(x) \in 1, ..., 14$ αντιστοιχίζει $t \in T, s \in S$ σε μία από τις 14 κατηγορίες προπαγάνδας. Ένα παράδειγμα δίνεται στο Σχήμα 5.1 με $t1$ σημειώνουμε τις λέξεις "stupid and petty" με

την τεχνική Loaded Language. Έπειτα, ορίζουμε την ακόλουθη συνάρτηση για να διαχειριστούμε μερικώς επικαλυπτόμενα κομμάτια με τις ίδιες ετικέτες.

$$C(s,t,h) = \frac{|s \cap t|}{h}\delta(l(s), l(t)), \tag{5.1}$$

όπου το h είναι ένας συντελεστής κανονικοποίησης και $\delta(\alpha, b) = 1$ αν $\alpha = b$, και 0 σε άλλη περίπτωση. Για παράδειγμα, αναφερόμενοι πάλι στο Σχήμα 5.1, $C(t_1, s_1, |t_1|) = \frac{6}{16}$ και $C(t_1, s_2, |t_1|) = 0$. Δοσμένης της εξίσωσης 5.1, δημιουργούμε παραλλαγές των γνωστών μετρικών precision και recall ικανών να λάβουν υπόψιν την ανισορροπία (imbalance) του συνόλου δεδομένων :

$$P(S,T) = \frac{1}{|S|}\sum_{s \in S, t \in T} C(s,t,|s|), \tag{5.2}$$

$$R(S,T) = \frac{1}{|T|}\sum_{s \in S, t \in T} C(s,t,|t|), \tag{5.3}$$

Θέτουμε την εξίσωση 5.2 ίση με το μηδέν αν $|S| = 0$ και την εξίσωση 5.3 ίση με το μηδέν αν $|T| = 0$. Ακολουθώντας τη μεθοδολογία του , στις εξισώσεις 5.2 και 5.3 δίνουμε ποινή σε συστήματα που προβλέπουν πάρα πολλά ή πολύ λίγα συμβάντα διαιρώντας με $|S|$ και $|T|$ αντίστοιχα. Για παράδειγμα, στο Σχήμα 5.1 $R(s_3, s_4, s_5, t_1) = \frac{7}{24} < R(s_1, t1) = \frac{9}{24} < R(t1, t1) = 1$.

Τέλος, συνδυάζοντας τις εξισώσεις 5.2 και 5.3 σε μια μια μετρική της μορφής F1, προκύπτει ο αρμονικός μέσος όρος του precision και recall:

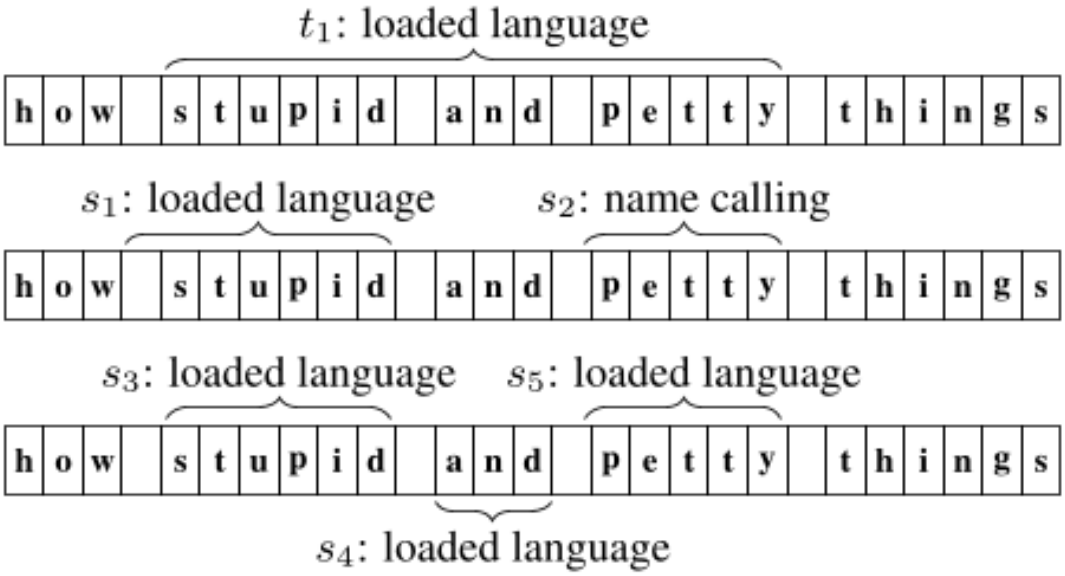
$$F_1(S,T) = 2\frac{P(S,T)R(S,T)}{P(S,T) + R(S,T)} \tag{5.4}$$

Η εξίσωση 5.4 μπορεί να υπολογιστεί σε σχέση μόνο με μία προπαγανδιστική τεχνική, αντικαθιστώντας τη συνάρτηση δ στην εξίσωση 5.1 με τη $\delta_L = 1$ αν $\alpha = b = L$, όπου L είναι η επιλεγμένη προπαγανδιστική τεχνική.

Η προηγούμενη ανάλυση των μετρικών πραγματεύεται την γενική περίπτωση όπου μπορεί να αξιολογηθεί η εύρεση πολλών προπαγανδιστικών τεχνικών. Το SI task είναι μια ειδική περίπτωση της προηγούμενης μεθοδολογίας όπου $\forall \alpha, b, \delta(\alpha, b) = 1$. Επίσης οι εξισώσεις 5.2 και 5.3 για να έχουν τιμές μικρότερες ή ίσες της μονάδας, όλες οι επικαλυπτόμενες επισημειώσεις, ανεξάρτητα της τεχνικής που χρησιμοποιείται, πρώτα συγχωνεύονται. Για παράδειγμα, στο Σχήμα 5.2 μετά την συγχώνευση οι τρεις επισημειώσεις $(t_1, s_1 + s_2)$ και $(s_3 + s_4)$ γίνονται ισοδύναμες μεταξύ τους.

5.2.2 Κατηγοριοποίηση Τεχνικών (TC task)

Αυτό το υπο-πρόβλημα συγκαταλέγεται στην κατηγορία προβλημάτων ταξινόμησης σε πολλαπλές κλάσεις. Αυτές οι κλάσεις, όπως έχουμε ήδη αναφέρει είναι ιδιαίτερα μη ισορροπημένες ως προς το πλήθος των στοιχείων τους. Για παράδειγμα, οι κλάσεις Loaded Language και Nmae Calling, Labelling περιέχουν 3.841 στοιχεία ενώ άλλες κλάσεις όπως η Appeal to Authority περιέχει λιγότερα από 180 στοιχεία. Είναι προφανές λοιπόν ότι πιο συχνά εμφανιζόμενες κλάσεις θα είναι πιο εύκολο να εντοπιστούν από ότι αυτές με λίγα στοιχεία για το λόγο ότι ο αλγόριθμος θα έχει περισσότερα δεδομένα για να εκπαιδευτεί και άρα να μάθει καλύτερα από ότι στις λιγότερο συχνές κλάσεις. Συνεπώς, η επίσημη μετρική αξιολόγησης που επιλέχθηκε είναι η micro-averaged F1



Σχήμα 5.1: Παράδειγμα επισημειώσεων (άνω μέρος) και των προβλέψεων ενός μοντέλου μάθησης με επίβλεψη (κάτω μέρος) σε ένα κείμενο που αναπαριστάται ως μια ακολουθία χαρακτήρων.

measure. Αυτή η μετρική είναι γνωστή επίσης με το όνομα balanced F-score (ισορροπημένη F-score) και υπολογίζεται από το σταθμισμένο μέσο όρο του precision και του recall. Η μέγιστη (καλύτερη) τιμή της είναι η μονάδα και η μικρότερη το 0. Η σχετική συνεισφορά του precision και του recall είναι ισοδύναμη στο τελικό F1-score. Η φόρμουλα με την οποία αυτό υπολογίζεται είναι η εξής:

$$F1 = average_parmeter \frac{(precision * recall)}{precision + recall} \tag{5.5}$$

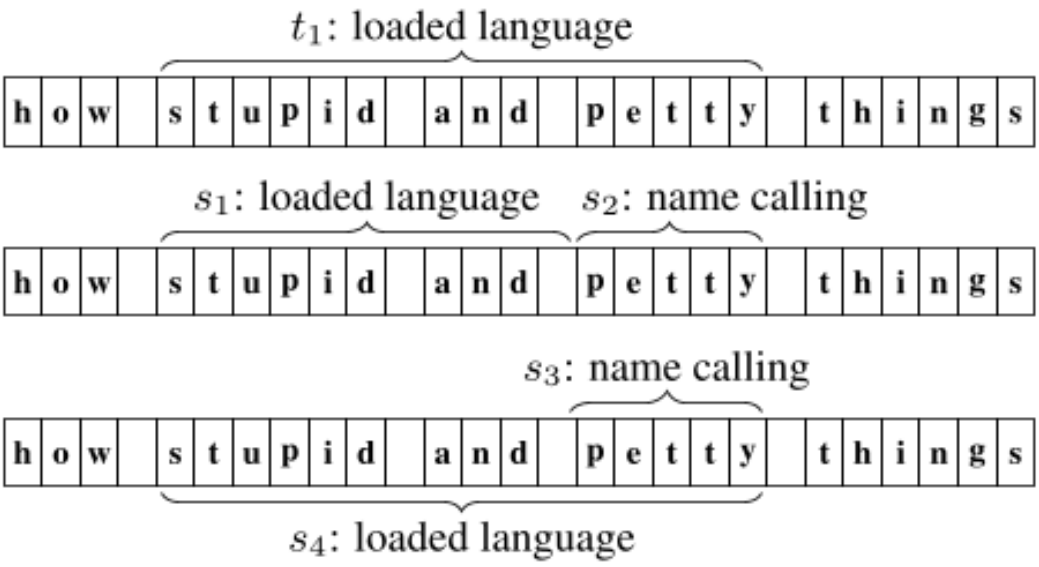
Σε περίπτωση με πολλαπλές κλάσεις, αυτή η average_parameter ισούται με το πλήθος των κλάσεων.

Μοντέλο	F1	Precision	Recall
Logistic Regression Word2Vec	0.16	0.09	0.47
BiLSTM + GloVe	0.30	0.25	0.37
BiLSTM + ELMo	0.32	0.27	0.38

Πίνακας 5.2: Αποτελέσματα για το πρόβλημα της Ανίχνευσης στο σύνολο δεδομένων του dev φακέλου

5.3 Προ-επεξεργασία Δεδομένων

Τα δεδομένα όπως έχει ήδη ειπωθεί είναι σε μορφή κειμένων. Σε πρώτη φάση οι συμμετέχοντες του διαγωνισμού έλαβαν δύο φακέλους, ο ένας περιείχε τα άρθρα με τα οποία θα γινόταν η εκπαίδευση του αλγορίθμου και ο δεύτερος με τον οποίο θα γινόταν η πρώτη φάση της αξιολόγησης των αλγορίθμων. Τα κομμάτια των κειμένων και οι ετικέτες των μορφών προπαγάνδας δόθηκαν μόνο για τα άρθρα εκπαίδευσης και όχι για τα άρθρα υλοποίησης (δεύτερος φάκελος (dev)) προκειμένου



Σχήμα 5.2: Παράδειγμα επισημειώσεων

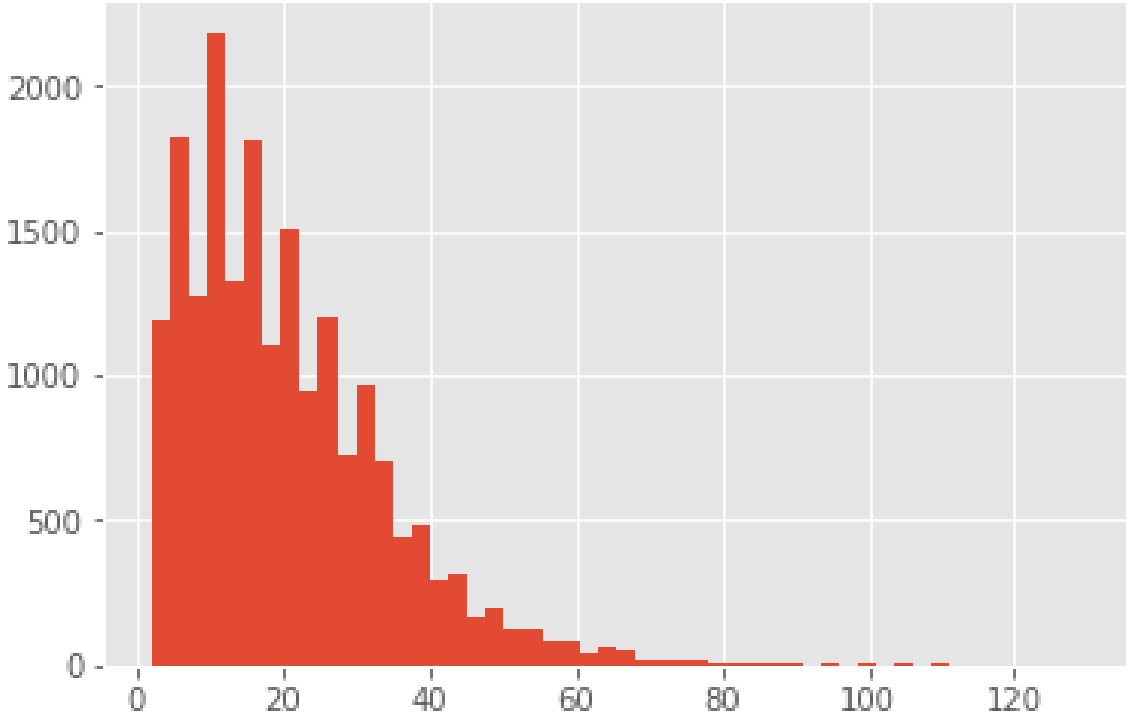
η αξιολόγηση να γίνει αδιάβλητα χωρίς να υπάρχει η δυνατότητα να εκμεταλλευτεί κάποια ομάδα τις απαντήσεις. Παρόλα αυτά, λόγω της μορφής του προβλήματος τα δεδομένα εισόδου του δεύτερου υπο-προβλήματος (TC task) περιείχαν τις απαντήσεις για το πρώτο υπο-πρόβλημα(SI task). Για το λόγο αυτό, στη δεύτερη και τελευταία φάση του διαγωνισμού δόθηκε ένας τρίτος φάκελος (test-set) ξεχωριστά για το κάθε υπο-πρόβλημα με διαφορά μιας εβδομάδας για να μην υπάρχει καμία περίπτωση κακόβουλης χρήσης των ετικετών. Σε κάθε κείμενο οι προτάσεις χωρίστηκαν αυτόματα με τη βιβλιοθήκη NLTK sentence splitter. Ακόμα, αφαιρέσαμε όλα τα σημεία στίξης και τους ακόλουθους χαρακτήρες !()-[];:'",<>./?@ , τα οποία δεν προσδίδουν πληροφορία που μπορεί να αξιοποιηθεί από κάποιο από τα μοντέλα μας. Έπειτα, χωρίσαμε την κάθε πρόταση στις επιμέρους λέξεις της. Μέχρι σε αυτό το σημείο ακολουθήσαμε την ίδια μεθοδολογία και στα δυο υπο-προβλήματα του διαγωνισμού, αφού τα δεδομένα είχαν την ίδια μορφή κειμένου. Η μεθοδολογία διαφοροποιείται στο παρακάτω κεφάλαιο της δημιουργίας χαρακτηριστικών, λόγω του πολύ μεγάλου όγκου δεδομένων για το πρόβλημα ανίχνευσης.

Model	micro-averaged F1 measure
Logistic Regression + Word2Vec	0.41
MLP + Word2Vec	0.44
BiLSTM + GloVE	0.53

Πίνακας 5.3: Αποτελέσματα για το πρόβλημα της κατηγοριοποίησης τεχνικών στα δεδομένα του φακέλου dev

5.4 Δημιουργία Χαρακτηριστικών

Για κάθε λέξη δημιουργείται ένας πίνακας χαρακτηριστικών ο οποίος διαμορφώνεται ανάλογα με το ποια προ-εκπαιδευμένα διανύσματα λέξεων (word embeddings) χρησιμοποιούνται. Πιο



Σχήμα 5.3: Κατανομή μήκους των προτάσεων των κειμένων

αναλυτικά, για τα μοντέλα που χρησιμοποιούν τη βιβλιοθήκη Wor2Vec ο πίνακας χαρακτηριστικών περιέχει 100 διαστάσεις για κάθε λέξη. Μετά από αρκετά πειράματα καταλήξαμε στο συμπέρασμα ότι από τις τέσσερις διαφορετικές διαστάσεις πινάκων που είναι διαθέσιμες από τη βιβλιοθήκη GloVe, αυτή με τα καλύτερα αποτελέσματα ήταν με τις 300 διαστάσεις για κάθε στοιχείο (λέξη). Όσον αφορά τον τελικό αλγόριθμο του υπο-προβλήματος ανίχνευσης χρησιμοποιήθηκε η βιβλιοθήκη ELMo για τη δημιουργία των χαρακτηριστικών.Τα πολλαπλά πλεονεκτήματα αυτού του μοντέλου έχουν αναλυθεί εκτενώς στο κεφάλαιο 3 στην ενότητα 3.4.7. Λόγω του μεγάλου όγκου των δεδομένων του πρώτου υπο-προβλήματος (πλήθος προτάσεων: 19.428, πλήθος λέξεων: 401.573, πλήθος διαφορετικών λέξεων: 26.393) σε συνδυασμό με το πίνακα χαρακτηριστικών από το μοντέλο ELMo 1.024 διαστάσεων για κάθε λέξη, αναγκαστήκαμε να περικόψουμε τις προτάσεις που περιείχαν πάνω από 80 λέξεις. **Όπως φαίνεται και στο Σχήμα 5.3 πολύ λίγες προτάσεις έχουν μήκος άνω των 80 λέξεων και με αυτό τον τρόπο πετύχαμε χάνοντας πρακτικά ελάχιστη πληροφορία να κερδίσουμε επιτάχυνση της εκπαίδευσης του μοντέλου και κατά συνέπεια περισσότερο διαθέσιμο χρόνο για πειράματα με διαφορετικές παραμέτρους. Αυτό είναι ένα γνωστό φαινόμενο σε πολλά φυσικά συστήματα μεταξύ των οποίων και η γλώσσα που ακολουθεί την κατανομή μακράς ουράς (long-tail distribution) ή αλλιώς τον κανόνα 80-20 (το 20% των μοναδικών λέξεων εμφανίζεται στο 80% των συνολικών εμφανίσεων).**

Όσον αφορά το δεύτερο υπο-πρόβλημα, δεν χρειάστηκε να προβούμε στις παραπάνω ενέργειες, γιατί τα δεδομένα εισόδου ήταν μόνο τα κομμάτια του κειμένου που είχαν ανιχνευθεί ως προπαγανδιστικά. Επίσης, σε αυτή την περίπτωση λόγω αυτής της μικρής ποσότητας των δεδομένων δεν ενδείκνυται η χρήση γλωσσικών μοντέλων όπως το ELMo που δημιουργούν χαρακτηριστικά πολλών διαστάσεων, αλλά πιο απλά μοντέλα όπως το GloVe.

Team	F1	Precision	Recall
Hitachi (best)	0.51	0.56	0.47
NTUAAILS	0.33	0.46	0.26

Πίνακας 5.4: Τελικά αποτελέσματα για το πρόβλημα της Ανίχνευσης σε σύγκριση με την νικήτρια ομάδα του διαγωνισμού

5.5 Αξιολόγηση Αλγορίθμων

Μετά τη δημιουργία των απαραίτητων χαρακτηριστικών και την προεπεξεργασία των δεδομένων χρησιμοποιήσαμε για κάθε ένα από αυτά διάφορους αλγορίθμους ταξινόμησης, όπου και αξιολογήσαμε τα αποτελέσματά τους. Επίσης, συγκρίναμε την απόδοση αυτών των αρχιτεκτονικών με άλλες παρεμφερείς εργασίες αλλά και με τις πρώτες στην κατάταξη ομάδες του διαγωνισμού.

5.5.1 Ανίχνευση Προπαγάνδας

Μετά από πολλαπλά πειράματα και δοκιμές προσπαθήσαμε να μεγιστοποιήσουμε την απόδοση των αλγορίθμων μας και στα δυο υπο-προβλήματα του διαγωνισμού. Ο Πίνακας 5.2 αποτυπώνει την απόδοση των τριών κύριων μεθόδων που χρησιμοποιήσαμε για την ανίχνευση προπαγάνδας. Όπως ήταν αναμενόμενο τα μοντέλα που περιέχουν αλγορίθμους βαθιάς νευρωνικής Μάθησης υπερισχύουν κατά πολύ σε απόδοση έναντι των αλγορίθμων Μηχανικής Μάθησης.

Πιο συγκεκριμένα, με την προσθήκη του αμφίδρομου LSTM υπήρξε δραματική αύξηση στη μετρική Precision και αξιοσημείωτη απόδοση στη μετρική Recall. Ο πρώτος αλγόριθμος αποτελούμενος από Logistic Regression και Word2Vec πέτυχε Precision μόλις 0.09 και σχετικά μεγάλο Recall 0.47. Η κακή απόδοση στη μετρική Precision κυρίως οφείλεται στο γεγονός ότι αυτός ο αλγόριθμος κατηγοριοποιεί μία λέξη τη φορά, μη λαμβάνοντας υπόψιν την πληροφορία από τα συμφραζόμενα. Από την άλλη μεριά, με την προσθήκη του αμφίδρομου LSTM στο δεύτερο αλγόριθμο αυτό το μειονέκτημα αντιμετωπίζεται αποτελεσματικά και παρατηρείται αύξηση στο Precision ακόμα και χωρίς ιδιαίτερη ρύθμιση των υπερπαραμέτρων του δικτύου. Η αύξηση στη μετρική του precision είναι τουλάχιστον $+0.10$. Συγχρόνως όμως παρατηρείται μια ελαφρά μείωση στο Recall από 0.47 στο 0.37. Παρόλα αυτή την αντιστρόφως ανάλογη σχέση, το F1 που αποτελεί και την επίσημη μετρική αξιολόγησης, αυξάνεται με την προσθήκη του LSTM. Μετά από δοκιμές και ρύθμιση των υπερπαραμέτρων, ο δεύτερος αλγόριθμος πετυχαίνει 0.30, 0.25, 0.37 σε F1, Precision, Recall αντίστοιχα.

Τέλος το τρίτο μοντέλο το οποίο περιλαμβάνει το ELMo και το διπλό αμφίδρομο LSTM με την υπολειπόμενη (residual) σύνδεση έχει την καλύτερη απόδοση που πέτυχε η ομάδα μας στον διαγωνισμό στο development φάκελο, για αυτό το λόγο το χρησιμοποιήσαμε και για την τελική μας υποβολή. Παρατηρήθηκε απόκλιση στα αποτελέσματα μεταξύ των δεδομένων από τον test φάκελο (τελική υποβολή) και development φάκελο στην πλειονότητα των διαγωνιζόμενων. Αυτό κατά πάσα πιθανότητα οφείλεται στο μικρό μέγεθος του συνόλου δεδομένων και την ευαισθησία του κατά συνέπεια σε δεδομένα με ακραίες τιμές.

Η ομάδα μας έλαβε 0.33, 0.46, 0.26 **στις μετρικές F1, Precision και Recall αντίστοιχα στην τελική υποβολή και κατατάχθηκε 28η στο πρόβλημα της ανίχνευσης προπαγάνδας**. Ο Πίνακας 5.4 απεικονίζει τα αποτελέσματα σε σύγκριση με την πρώτη (νικήτρια) ομάδα.

5.5.2 Κατηγοριοποίηση Τεχνικών

Σε αυτό το υπό-πρόβλημα ο στόχος είναι να προσδιορίσουμε τον τύπο της προπαγάνδας του κάθε τμήματος κειμένου που δίνεται. Αρχικά, το σύνολο δεδομένων περιείχε 18 κλάσεις για ταξινόμηση. Κατά τη διάρκεια του διαγωνισμού παρατηρήθηκε ότι κλάσεις με πολύ μικρό αριθμό δειγμάτων ήταν σχεδόν αδύνατο να εντοπιστούν. Συνεπώς, τέσσερις κλάσεις συμπτύχθηκαν με παρεμφερείς τους κλάσεις. Παρόλα αυτά, η κατανομή των δειγμάτων των κλάσεων παρέμεινε ιδιαίτερα ανισορροπημένη και το σύνολο δεδομένων αρκετά μικρό. Αυτός μας οδήγησε στο να χρησιμοποιήσουμε μόνο ένα επίπεδο αμφίδρομου biLSTM νευρωνικού δικτύου στο τελικό μας μοντέλο. Χαρακτηριστικό στοιχείο της ανισορροπίας και της ελλιπούς πληροφορίας που προσφέρουν τα δεδομένα είναι ότι τα αποτελέσματα σχεδόν όλων των ομάδων (συμπεριλαμβανομένης και της δικής μας) είχαν πολύ μεγάλη απόκλιση μεταξύ της αξιολόγησης στα δεδομένα των test και dev φακέλων. Για παράδειγμα η νικήτρια ομάδα έλαβε 0.70 στη μετρική F1 στο dev φάκελο και 0.62 στο test φάκελο. Η δική μας ομάδα έλαβε 0.53 στο dev φάκελο και 0.46 στο test φάκελο. Μετά από επισκόπηση στα επιμέρους αποτελέσματα των αλγορίθμων σε κάθε μία από τις 14 κλάσεις αυτή η απόκλιση οφείλεται στην αδυναμία αξιόπιστης πρόβλεψης των κλάσεων με πολύ μικρό πλήθος δειγμάτων στα δεδομένα εκπαίδευσης. Ως επακόλουθο των ελλιπών δεδομένων είναι ιδιαίτερα δύσκολο ένας αλγόριθμος να δημιουργήσει τα σωστά πρότυπα για την εύρεση αυτών των κλάσεων.

Κεφάλαιο 6

Συμπεράσματα

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία αντιμετωπίζουμε το πρόβλημα της αυτοματοποιημένης ανίχνευσης προπαγάνδας σε ειδησιογραφικά άρθρα και της κατηγοριοποίησης σε μια από τις 14 μορφές αυτής. Αυτό το πρόβλημα συμπεριλαμβάνεται στο διεθνές συνέδριο SemEval, ένα από τα πιο γνωστά και αναγνωρισμένα συνέδρια στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Precessing).

Για πρώτη φορά στον ευρύτερο κλάδο της ανίχνευσης ψευδών ειδήσεων, εξτρεμιστικών και ρατσιστικών σχολίων και προπαγάνδας αποπειράθηκε η εύρεση συγκεκριμένων κομματιών του κειμένου που χρησιμοποιούνται τέτοιες τεχνικές. Προηγούμενες επιστημονικές εργασίες σε αυτό τον κλάδο επιχειρούσαν να ταξινομήσουν σε επίπεδο κειμένου αν τα άρθρα περιείχαν σε σημαντικό βαθμό αυτές τις κακόβουλες τεχνικές και όχι να βρουν συγκεκριμένα σημεία όπου αυτές χρησιμοποιούνται.

Για το σκοπό αυτό οι διοργανωτές ανέθεσαν στην εταιρία Data Pro τη δημιουργία ένος συνόλου δεδομένων από άρθρα προερχόμενα από 48 διαφορετικές ειδησεογραφικές πηγές, στα οποία επισημειώθηκαν τα συγκεκριμένα σημεία του κειμένου και η αντίστοιχη μορφή προπαγάνδας σε ένα όγκο συνολικά 451 κειμένων. Αυτά μετέπειτα χωρίστηκαν στα άρθρα για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης και στα άρθρα για την αξιολόγηση αυτών.

Από την μεριά μας ως διαγωνιζόμενοι έπρεπε να προβούμε σε προεπεξεργασία των δεδομένων που μας δόθηκαν προκειμένου αλγόριθμοι μηχανικής μάθησης και βαθιάς νευρωνικής μάθησης να μπορέσουν να αξιοποιήσουν τα δοσμένα κείμενα. Η βασική ιδέα ήταν να αντιμετωπίσουμε το πρόβλημα σαν μία παραλλαγή της κατηγορίας προβλημάτων NER (Name Entity Recognition). Αυτό είχε σαν συνέπεια να χωρίσουμε τα κείμενα σε προτάσεις και έπειτα σε λέξεις και να μεσχηματίσουμε το πρόβλημα από ταξινόμηση σε επίπεδο χαρακτήρων σε επίπεδο λέξεων. Λόγω αυτής της μετατροπής προέκυψε μια μικρή απώλεια πληροφορίας αλλά εν τέλει αυτός ο μετασχηματισμός μας βοήθησε να επιτύχουμε αξιόλογα αποτελέσματα, ειδικά στο κύριο υπο-πρόβλημα της ανίχνευσης της προπαγάνδας.

Έπειτα όπως αναφέρουμε διεξοδικά στο κεφάλαιο 4 δημιουργήσαμε συνολικά τρεις αρχιτεκτονικές για το υπο-πρόβλημα της ανίχνευσης προπαγάνδας και τρεις αρχιτεκτονικές για το υπο-πρόβλημα της κατηοριοποίησης των τεχνικών. Τα τελικά μοντέλα παρατίθενται και σχηματικά στα Σχηματα 4.1 και 4.2.

Τα αποτελέσματα αυτών των αρχιτεκτονικών για τα δύο υπο-προβλήματα παρατίθενται στους Πίνακες 5.4 και 5.3 για την Ανίχνευση και την Κατηγοριοποίηση της προπαγάνδας αντίστοιχα. Πιο

αναλυτικά, για το πρόβλημα της Ανίχνευσης Προπαγάνδας (SI task) χρησιμοποιώντας δυο επίπεδα αμφίδρομων LSTM νευρωνικών δικτύων τα οποία τροφοδοτήθηκαν με διανυσματικές αναπαραστάσεις λέξεων από το γλωσσικό μοντέλο ELMo λάβαμε ποσοστό 0.33, 0.46 και 0.26 στις μετρικές $F1$m, $Precision$ και $Recall$ αντίστοιχα. Έτσι, τερματίσαμε 27οι στην κατάταξη με την πρώτη ομάδα να πετυχαίνει 0.51, 0.56 και 0.47 στις μετρικές $F1$, $Precision$ και $Recall$ αντίστοιχα.

Όσον αφορά το πρόβλημα της Κατηγοριοποίησης η ομάδα μας έλαβε την 25η θέση χρησιμοποιώντας ένα νευρωνικό δίκτυο από αμφίδρομο LSTM εμπλουτισμένου με διανυσματικές αναπαραστάσεις λέξεων από το μοντέλο GloVe. Αυτό ο αλγόριθμος έλαβε ποσοστό 0.463 στη μετρική $F1$, ενώ η πρώτη στην κατάταξη ομάδα έλαβε ποσοστό 0.620. Όπως αναφέρθηκε στο Κεφάλαιο 5 λόγω της ανισορροπίας των κλάσεων και του μικρού αριθμού δειγμάτων σε πολλες από αυτές τις κλάσεις σχεδόν όλες οι ομάδες που συμμετείχαν στον διαγωνισμό παρουσίασαν σημαντική μείωση της απόδοσης των αλγορίθμων τους ανάμεσα στις υποβολές στα σύνολα dev και στο (τελικό) σύνολο δεδομένων test. Αυτό καταδεικνύει την ανάγκη για περαιτέρω διεύρυνση του συνόλου των δεδομένων.

6.2 Προτάσεις για Μελλοντική Έρευνα

Στα πλαίσια αυτής της εργασίας μελετήθηκαν διαφορετικές προσεγγίσεις για την επίλυση του προβλήματος της Ανίχνευσης Προπαγανδιστικών Σχολίων. Παρακάτω παραθέτουμε επιπλέον προτάσεις για μελλοντική έρευνα τόσο στο συγκεκριμένο πρόβλημα με στόχο την επίτευξη καλύτερων αποτελεσμάτων όσο και στο ευρύτερο πεδίο της ανίχνευσης κακόβουλων σχολίων.

Αρχικά, όπως εκτεταμένα αναφέρθηκε μία από τις κύριες προκλήσεις του διαγωνισμού και ειδικά στο υπο-πρόβλημα της Κατηγοριοποίησης ήταν η μικρή ποσότητα των δεδομένων. Αυτό το πρόβλημα θα μπορούσε να αντιμετωπιστεί με χρήση τεχνικών όπως της υπεδειγματοληψίας (oversampling) αλλά χρησιμοποιώντας επίσης παρόμοιου τύπου σύνολα δεδομένων. Πέρα από αυτό, σκοπεύουμε να εμπλουτίσουμε την αρχιτεκτονική μας (κυρίως όσον αφορά το SI task) με άλλες μεθόδους όπως το μοντέλο Conditional Random Fields Lafferty et al. (2001) το οποίο έχει δείξει εντυπωσιακά αποτελέσματα στον τομέα της εξόρυξης πληροφορίας (information extraction). Ενδεικτικά, οι πρώτες ομάδες σε κατάταξη στον διαγωνισμό χρησιμοποίησαν τα Conditional Random Fields με Μετασχηματιστές (Transformers) όπως ο BERT Devlin et al. (2019) και ο ROBERTA.

Τέλος, στοχεύουμε να χρησιμοποιήσουμε αυτά τα μοντέλα και σε άλλα παρεμφερή προβλήματα του τομέα Επεξεργασίας Φυσικής Γλώσσας όπως η ορθότητα πολιτικών δεδομένων (political fact checking), ένα ανοικτό πρόβλημα που τίθεται κάθε χρόνο στο διεθνές συνέδριο Conference and Labs of the Evaluation Forum (CLEF).

Βιβλιογραφία

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.

American Psychological Association. *Publications Manual*. American Psychological Association, Washington, DC, 1983.

Anastasios Arsenos and Georgios Siolas. Ntuaails at semeval-2020 task 11: Propaganda detection and classification with bilstms and elmo. International Workshop on Semantic Evaluation, 2020.

Ipek Baris, Lukas Schmelzeisen, and Steffen Staab. Clearumor at semeval-2019 task 7:convolving elmo against rumors. In *Proceedings of SemEval 2019 Conference*, Institute for Web Science and Technologies (WeST), University of Koblenz-Landau, Germany2Web and Internet Science Group (WAIS), University of Southampton, United Kingdom, 2019.

Benjamin Börschinger and Mark Johnson. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia, December 2011. URL https://www.aclweb.org/anthology/U11-1004.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981. doi: 10.1145/322234.322243.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Israa Jaradat, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864, 2019a.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Israa Jaradat, and Preslav Nakov. "proppy: A system to unmask propaganda in online news". *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19)*, page 9847–9848, 2019b.

Giovanni Da San Martino, Albertono Barrón-Cedeno, and Preslav Nakov. Evaluation of propaganda detection tasks. *Shared Task at SemEval 2020 Task 11:"Detection Of Propaganda Techniques In News Articles"*, (5), 2019c.

Giovanni Da San Martino, Albertono Barrón-Cedeno, and Preslav Nakov. Findings of the nlp4if-2019 shared taskon fine-grained propaganda detection. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 201*, (5), 2019d.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November 2019e.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain, December 2020.

Lavinia Dan. Techniques for the translation of advertising slogan. In *Proceedings of the International Conference Literature, Discourse and Multi-cultural Dialogue, LDMD '15,*, pages 13–23, Mures,Romania, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Adam Ek and Mehdi Ghanimifard. Synthetic propaganda embeddings to train a linear projection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 155–161, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5023. URL https://www.aclweb.org/anthology/D19-5023.

Association for Computing Machinery. *Computing Reviews*, 24(11):503–512, 1983.

Jean Goodwin and Raymie E. McKerrow. Accounting for the force of the appeal to authority. 2011.

Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK, 1997.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

Renee Hobbs. Teaching about propaganda: An examination of the historical roots of media literacy. *Journal of Media Literacy Education*, 6:56–67, 11 2014. doi: 10.23860/JMLE-2016-06-02-5.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

John Hunter. Brainwashing in a large groupawareness training? the classical conditioning hypothesis of brainwashing. Uni-versity of Kwazulu-Natal, Pietermaritzburg, SouthAfrica, 2015.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. Team bertha von suttner at semeval-2019 task 4: Hyperpartisan news detection using elmo sentence representation convolutional network. In *Proceedings of SemEval 2019 Conference*, University of Sheffield Sheffield , UK, 2019.

Garth S. Jowett and Victoria O'Donnell. *What is propaganda, and how does it differ from persuasion?* Sage Publishing, 2012.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. *In 12th International Workshop on Semantic Evaluation*, (5), 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. 2015.

Ping Liu, Wen Li, and Liang Zou. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2011. URL https://www.aclweb.org/anthology/S19-2011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Clyde R. Miller. *The Techniques of Propaganda.* an address given at Town Hall. The Center for learning, 1939.

Miller, C. R. *THE PUBLICATIONS OF THE INSTITUTE FOR PROPAGANDA ANALYSIS, INC.*, volume I. INSTITUTE FOR PROPAGANDA ANALYSIS, INC., 1937-1938.

Tom M. Mitchell. Machine learnig. *Machine Learning*, 1997.

Alexander Oberstrass, Julia Romberg, Anke Stoll, and Sefan Conrad. Hhu at semeval-2019 task 6: Context does matter - tackling offensivelanguage identification and categorization with elmo. In *Proceedings of SemEval 2019 Conference*, Institute of Computer Science, Heinrich Heine University D¨usseldorf, Germany, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL http://arxiv.org/abs/1802.05365. cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready.

Shahzad Qaiser and Ramsha Ali. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 07 2018. doi: 10.5120/ijca2018917395.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2931–2937, 2017.

Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. URL http://arxiv.org/abs/1503.06733. version 2.

Monika L Richter. The kremlin's platform for useful idiots' in the west: An overview of rt's editorial strategy and evidence of impact. Technical report, Kremlin Watch, 2017.

G. Rumelhart, D.and Hinton and R Williams. Learning representations by back-propagating errors. page 533–536, (1986. doi: https://doi.org/10.1038/323533a0.

Alessandro Seganti, Helena Sobol, Iryna Orlova, Hannam Kim, Jakub Staniszewski, Tymoteusz Krumholc, and Krystian Koziel. NLPR@SRPOL at SemEval-2019 task 6 and task 5: Linguistically enhanced deep learning offensive sentence classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 712–721, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2126. URL https://www.aclweb.org/anthology/S19-2126.

Servan-Schreiber, A. Cleeremans, and J. L McClelland. Graded state machines: The representation of temporal contingencies in simple recurrent networks. machine learning. page 161–193, (1991.

Rupesh Kumar Srivastava, Klaus Greff, and J̈urgen Schmidhuber. Highway networks. 2015.

Francisca Niken Vitri Suprabandari. Ameri-can propaganda in john steinbeck's the moon isdown. Sanata Dharma University,Yogyakarta, Indonesia, 2007.

Gabriel H Teninbaum. Reductio ad hitlerum:trumping the judicial nazi card. page 541. Michigan StateLaw Review, 2009.

Robyn Torok. Symbiotic radicalisation strate-gies: Propaganda tools and neuro linguistic programming. In *InProceedings of the Australian Se-curity and Intelligence Conference*, page 58–65, Perth, Australia, 2015.

Douglas Walton. The straw man fallacy. Royal, Netherlands Academy of Arts and Sciences, 1996.

Anthony Weston. A rulebook for arguments. 2018.

Shehel Yoosuf and Yin Yang. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5011. URL https://www.aclweb.org/anthology/D19-5011.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1:43–52, 12 2010. doi: 10.1007/s13042-010-0001-0.

Δημοσιεύσεις

Paper title: NTUAAILS at SemEval-2020 Task 11: Propaganda Detection and Classification with biLSTMs and ELMo.

Authors: Anastasios Arsenos and Georgios Siolas

Accepted to the SemEval-2020 workshop on semantic evaluation, 28th International Conference on Computational Lingustics (COLING-2020) December 8-13, 2020