ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εκτίμηση και Επικύρωση Περιεχομένου με Χρήση Μεθόδων Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γρηγόριος Χ. Παπανικολάου

Επιβλέπων : **Ευστάθιος Δ. Συκάς**
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Εκτίμηση και Επικύρωση Περιεχομένου με Χρήση Μεθόδων Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γρηγόριος Χ. Παπανικολάου

Επιβλέπων : **Ευστάθιος Δ. Συκάς**
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 31η Οκτωβρίου 2023.

.....	.....	.....
Ευστάθιος Δ. Συκάς	Νικόλαος Μήτρου	Ιωάννα Ρουσσάκη
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπληρώτρια Καθηγήτρια Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

.....

Γρηγόριος Χ. Παπανικολάου

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

Περίληψη

Η συγκεκριμένη εργασία, προτείνει μια εναλλακτική μέθοδο για την επίλυση του προβλήματος επικύρωσης εγκυρότητας περιεχομένου. Η έλλειψη αξιόπιστων συνόλων δεδομένων, συνδυαστικά με τους περιορισμούς των μεθόδων αναπαράστασης κειμένου, αποτελούσαν για πολλά χρόνια πρόκληση για τους ερευνητές που αξιοποιούσαν τεχνικές μηχανικής μάθησης στην ταξινόμηση των δειγμάτων. Ο σκοπός της προτεινόμενης μεθοδολογίας, είναι η βελτιστοποίηση προσεγγίσεων μηχανικής μάθησης που χρησιμοποιούνται για την ανίχνευση των fake news. Στην παρούσα έρευνα, πραγματοποιείται σύγκριση και μελέτη πολλαπλών συνόλων δεδομένων, συνδυαστικά με μια σύγχρονη μέθοδο αναπαράστασης κειμένου που αξιοποιεί το πλαίσιο χρήσης των λέξεων, βασισμένη σε μηχανισμούς προσοχής. Τα δείγματα των συνόλων δεδομένων ISOT, LIAR, FakeNewsNet, PHEME και FakeNewsChallenge κωδικοποιούνται με χρήση του μοντέλου DistilBERT, το οποίο βασίζεται σε δομές μετασχηματιστών (transformers). Στη συνέχεια, οι αναπαραστάσεις των δειγμάτων τροφοδοτούν αρχιτεκτονικές νευρωνικών δικτύων που αναπτύχθηκαν στο πλαίσιο της εργασίας και έπειτα επιβεβαιώνεται η αποτελεσματικότητα της εκάστοτε αρχιτεκτονικής, μέσω μιας διαδικασίας διασταυρωμένης επικύρωσης. Η μεθοδολογία, οδήγησε σε βελτιωμένη απόδοση συγκριτικά με προσεγγίσεις που αξιοποιούν συμβατικές τεχνικές αναπαράστασης κειμένου. Το συμπέρασμα που προκύπτει από τα πειράματα, είναι πως η χρήση εμφυτευμάτων που βασίζονται στο πλαίσιο χρήσης των λέξεων προσφέρουν πολύ πιο ικανοποιητικά αποτελέσματα στην ταξινόμηση περιεχομένου σε όλα τα σύνολα δεδομένων.

Λέξεις Κλειδιά : εγκυρότητα περιεχομένου, μηχανική μάθηση, βαθιά μάθηση, μετασχηματιστές, μηχανισμοί προσοχής, DistilBERT

ABSTRACT

This work proposes an alternative method for addressing the problem of content validity verification. The lack of reliable datasets coupled with limitations in text representation methods, has been a challenge for researchers utilizing machine learning methods in classifying samples, particularly in the context of fake news. The purpose of the proposed methodology is to optimize machine learning approaches used for fake news classification. The study involves comparison and analysis of multiple datasets in conjunction with a modern text representation method based on attention mechanisms. Samples from datasets such as ISOT, LIAR, FakeNewsNet, PHEME, and FakeNewsChallenge are encoded using the DistilBERT model which is built upon transformer units. Subsequently, the representations are fed into neural network architectures followed by a cross-validation process to confirm the effectiveness of each architecture. The methodology resulted in improved performance compared to approaches utilizing conventional text representation techniques. The conclusion drawn from the experimental results is that using embeddings based on the context leads to enhanced performance across all datasets.

Keywords: BERT, DistilBERT, contextual embeddings, transformers, ISOT, PHEME, LIAR, FakeNewsNet, FakeNewsChallenge

Περιεχόμενα

Σχήματα **11**

Πίνακες **12**

1 Εισαγωγή **14**

1.1 Ορισμός των fake news **14**

1.2 Παραδείγματα και πιθανές συνέπειες **16**

1.3 Μορφή των fake news και του περιεχομένου τους **18**

1.4 Μέσα διάδοσης των fake news **21**

1.5 Κατηγορίες των fake news **22**

1.6 Μέθοδοι προσέγγισης του προβλήματος **24**

1.7 Περιγραφή του προβλήματος και προτεινόμενη λύση **27**

2 Μέθοδοι κωδικοποίησης κειμένου **30**

2.1 Προγενέστερες διαδικασίες κωδικοποίησης κειμένου **30**

2.1.1 Επεξεργασία κειμένου πριν την κωδικοποίηση 30

2.1.2 Term Frequency – Inverse Document Frequency (TF-IDF) 31

2.1.3 Word2Vec 32

2.1.4 GloVe 34

2.2 Μέθοδοι κωδικοποίησης βασισμένες στο πλαίσιο χρήσης των λέξεων **35**

2.2.1 Transformers 35

2.2.2 Embeddings from Language Models (ELMo) 36

2.2.3 Generative Pretrained Transformer (GPT) 38

2.2.4 Bidirectional Encoder Representations from Transformers (BERT) 39

2.3 DistilBERT **41**

3 Ανάλυση συνόλων δεδομένων **48**

3.1 Σημασία εύρεσης κατάλληλων συνόλων δεδομένων **48**

3.2 Παρουσίαση και διαδικασία ανάλυσης συνόλων δεδομένων **49**

3.2.1 ISOT 49

3.2.2 PHEME 53

3.2.3 FakeNewsNet 56

3.2.4 LIAR 59

3.2.5 FakeNewsChallenge 61

4 Περιγραφή μεθοδολογίας **64**

4.1 Παρουσίαση χρήσιμων βιβλιοθηκών **64**

4.1.1 Google Colab 64

4.1.2 NumPy 65

4.1.3 Pandas 66

4.1.4 Seaborn 66

4.1.5 TensorFlow 67

4.1.6 Transformers 68

4.2 Παρουσίαση προηγούμενων μεθοδολογιών **69**

4.3 Περιγραφή πειραματικής μεθοδολογίας **71**

4.4 Περιγραφή αρχιτεκτονικών **77**
4.4.1 Περιγραφή λειτουργίας δομικών στοιχείων των αρχιτεκτονικών 77
4.4.2 CNN 79
4.4.3 Bidirectional LSTM 80
4.4.4 FakeBERT 81
4.4.5 CNN-L2 Regularization 82
4.4.6 LSTM 83
4.4.7 CNN Light 83
4.4.8 Title – Text 84

5 Ανάλυση αποτελεσμάτων **85**

5.1 ISOT **85**
5.1.1 Τίτλοι 85
5.1.2 Άρθρα 86
5.1.3 Συμπυκνωμένα άρθρα 86

5.2 FakeNewsNet **87**

5.3 FakeNewsChallenge **87**
5.3.1 Άρθρα 88
5.3.2 Συμπυκνωμένα άρθρα 88

5.4 LIAR **89**

5.5 PHEME **89**

5.6 Σύγκριση βέλτιστων αποτελεσμάτων ανά σύνολο δεδομένων **90**

6 Σύγκριση αποτελεσμάτων και σχολιασμός **92**

6.1 Σύγκριση με προσεγγίσεις συμβατικών τεχνικών κωδικοποίησης **92**
6.1.1 ISOT 92
6.1.2 FakeNewsNet 93
6.1.3 FakeNewsChallenge 94
6.1.4 LIAR 96
6.1.5 PHEME 97

6.2 Συμπεράσματα **98**

6.3 Μελλοντικές προεκτάσεις **99**

Βιβλιογραφία **100**

Παράρτημα Α **104**

Παράρτημα Β **109**

Σχήματα

Εικόνα 1 - Μελλοντικά και ενεργά πεδία έρευνας για την ανίχνευση fake news στα social media (Shu, Wang, Silva, Tang, & Liu, 2017)..... 27
Εικόνα 2- Παρουσίαση των αρχιτεκτονικών CBOW και Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013)..... 33
Εικόνα 3 - Βασική αρχιτεκτονική κωδικοποίησης - αποκωδικοποίησης των transformers (Vaswani, et al., 2017)..... 36
Εικόνα 4 - Περιγραφή της δημιουργίας των εμφυτευμάτων και βελτιστοποίηση ανά διαφορετικό σκοπό. (Radford, Narasimhan, Salimans, & Sutskever, 2018)..... 39
Εικόνα 5-Συνδυασμός των επιμέρους εμφυτευμάτων πριν την διαδικασία της αρχικής μη επιβλεπόμενης εκπαίδευσης (Devlin, Chang, Lee, & Toutanova, 2019)..... 40
Εικόνα 6 - Περιγραφή των δύο διαδικασιών εκπαίδευσης του BERT (Devlin, Chang, Lee, & Toutanova, 2019)..... 41
Εικόνα 7 - Παρουσίαση του αυξανόμενου αριθμού παραμέτρων με την πάροδο του χρόνου (Sanh, Debut, Chaumond, & Wolf, 2020). 42
Εικόνα 8 - Παρουσίαση μεθόδων κωδικοποίησης κειμένου (Goswami, Kaliyar, & Narang, 2021)..... 43
Εικόνα 9 - Κατανομή των άρθρων σε κάθε σύνολο (Fake/True) ως προς τη θεματολογία τους. 50
Εικόνα 10 - Παρουσίαση κατανομών των μηκών των ψευδών τίτλων και άρθρων. 51
Εικόνα 11 - Παρουσίαση κατανομών των μηκών των αληθών τίτλων και άρθρων. 51
Εικόνα 12 - Word cloud των τίτλων των ψευδών δειγμάτων του ISOT. 52
Εικόνα 13- Word cloud των άρθρων των ψευδών δειγμάτων του ISOT. 52
Εικόνα 14 - Κατανομή των δειγμάτων ως προς τον χαρακτηρισμό τους. 54
Εικόνα 15 - Κατανομή των δειγμάτων σε κατηγορίες συνδυαστικά με το κύρος του δημιουργού. 54
Εικόνα 16 - Κατανομή των δειγμάτων ως προς το πλαίσιο στο οποίο αναφέρονται. 55
Εικόνα 17 - Word cloud των δημοσιεύσεων του συνόλου δεδομένων PHEME. 55
Εικόνα 18 - Κατανομή των δειγμάτων σε κατηγορίες ως προς τον χαρακτηρισμό τους. 56
Εικόνα 19 - Κατανομές των μηκών των δύο υποσυνόλων του FakeNewsNet. 57
Εικόνα 20 - Word cloud του ψευδούς υποσυνόλου του FakeNewsNet. 58
Εικόνα 21 - Word cloud του αληθούς υποσυνόλου του FakeNewsNet. 58
Εικόνα 22 - Κατανομή των μηκών των διαφορετικών υποσυνόλων του συνόλου δεδομένων LIAR. 60
Εικόνα 23 - Κατανομή των δειγμάτων του συνόλου δεδομένων LIAR σε κατηγορίες ως προς την αξιοπιστία τους. 60
Εικόνα 24 - Κατανομή των δειγμάτων του κάθε υποσυνόλου σε κατηγορίες. 61
Εικόνα 25 - Κατανομή των μηκών των τίτλων και των άρθρων του υποσυνόλου εκπαίδευσης. 62
Εικόνα 26 - Κατανομή των μηκών των τίτλων και των άρθρων του υποσυνόλου επαλήθευσης. 62
Εικόνα 27 - Παρουσίαση της πειραματικής διαδικασίας. 73
Εικόνα 28 - Παρουσίαση δομής συνελικτικού στρώματος ακολουθούμενο από στρώμα μέγιστης συσσώρευσης και πλήρως συνδεδεμένου στρώματος. (Yoon, 2014)..... 78

Πίνακες

Πίνακας 1 - Συνοπτική παρουσίαση διαδεδομένων μεθόδων κωδικοποίησης κειμένου 44
Πίνακας 2 - Συνοπτική περιγραφή των συνόλων δεδομένων..... 63
Πίνακας 3 - Βέλτιστοι βαθμοί μάθησης DistilBERT ανά σύνολο δεδομένων 76
Πίνακας 4 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN ανά στρώμα 80
Πίνακας 5 - Αριθμός παραμέτρων της αρχιτεκτονικής Bidirectional LSTM ανά στρώμα 81
Πίνακας 6 - Αριθμός παραμέτρων της αρχιτεκτονικής FakeBERT ανά στρώμα..... 82
Πίνακας 7 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN-L2 Regularization ανά στρώμα83
Πίνακας 8 - Αριθμός παραμέτρων της αρχιτεκτονικής LSTM ανά στρώμα..... 83
Πίνακας 9 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN-Light ανά στρώμα 84
Πίνακας 10 - Αριθμός παραμέτρων της αρχιτεκτονικής Title-Text ανά στρώμα..... 84
Πίνακας 11 - Αποτελέσματα εφαρμογής της μεθοδολογίας στους τίτλους του συνόλου δεδομένων ISOT..... 85
Πίνακας 12 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα άρθρα του συνόλου δεδομένων ISOT..... 86
Πίνακας 13 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα συμπυκνωμένα άρθρα του συνόλου δεδομένων ISOT..... 86
Πίνακας 14 - Αποτελέσματα εφαρμογής της μεθοδολογίας στους τίτλους του συνόλου δεδομένων FakeNewsNet..... 87
Πίνακας 15 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα άρθρα του συνόλου δεδομένων FakeNewsChallenge **δύο** βαθμίδων αξιοπιστίας 88
Πίνακας 16 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα συμπυκνωμένα άρθρα του συνόλου δεδομένων FakeNewsChallenge **δύο** βαθμίδων αξιοπιστίας 88
Πίνακας 17 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα δείγματα του συνόλου δεδομένων LIAR **δύο** βαθμίδων αξιοπιστίας..... 89
Πίνακας 18 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα δείγματα του συνόλου δεδομένων PHEME **δύο** βαθμίδων αξιοπιστίας..... 90
Πίνακας 19 - Σύγκριση βέλτιστων αποτελεσμάτων ανά σύνολο δεδομένων ανεξαρτήτως του τύπου δείγματος..... 91
Πίνακας 20 – Αποτελέσματα συνόλου δεδομένων ISOT με προσεγγίσεις άλλων μελετητών 93
Πίνακας 21 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα άρθρα του συνόλου δεδομένων ISOT 93
Πίνακας 22 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα συμπυκνωμένα άρθρα του συνόλου δεδομένων ISOT 93
Πίνακας 23 – Αποτελέσματα συνόλου δεδομένων FakeNewsNet με προσεγγίσεις άλλων μελετητών..... 94
Πίνακας 24 - Αποτελέσματα συνόλου δεδομένων FakeNewsChallenge με προσεγγίσεις άλλων μελετητών 95
Πίνακας 25 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα άρθρα του συνόλου δεδομένων FakeNewsChallenge **τεσσάρων** βαθμίδων αξιοπιστίας .. 95
Πίνακας 26 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα συμπυκνωμένα άρθρα του συνόλου δεδομένων FakeNewsChallenge **τεσσάρων** βαθμίδων αξιοπιστίας 96
Πίνακας 27 – Αποτελέσματα συνόλου δεδομένων LIAR με προσεγγίσεις άλλων μελετητών 96
Πίνακας 28- Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στα δείγματα του συνόλου δεδομένων LIAR **έξι** βαθμίδων αξιοπιστίας..... 96
Πίνακας 29 - Αποτελέσματα συνόλου δεδομένων PHEME με προσεγγίσεις άλλων μελετητών 97

Πίνακας 30 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στα δείγματα του συνόλου δεδομένων PHEME **τριών** βαθμίδων αξιοπιστίας 98

1 Εισαγωγή

1.1 Ορισμός των fake news

Η επικύρωση εγκυρότητας περιεχομένου, αφορά την εξακρίβωση της αξιοπιστίας της πληροφορίας που φτάνει στους αποδέκτες και είναι άρρηκτα συνδεδεμένη στο μυαλό των περισσότερων ανθρώπων με τις ψευδείς ειδήσεις (fake news). Αυτό συμβαίνει, γιατί ο όρος fake news χρησιμοποιείται για να περιγράψει ένα ευρύτερο σύνολο εννοιών αναφορικά με περιπτώσεις διάδοσης περιεχομένου που σε έναν βαθμό δεν ευσταθεί (Simons & Manoilo, 2021).

Πολλές φορές, δεν είναι αντιληπτό πως πρόκειται για περιεχόμενο που μπορεί να ανήκει σε διάφορες κατηγορίες ως προς τον τόνο του συγγραφέα, το μέσο διάδοσής του (Mishra, Shukla, & Agarwal, 2022), τον σκοπό της δημιουργίας του, τα μέσα πειθούς που χρησιμοποιούνται (Shu, Wang, Silva, Tang, & Liu, 2017), καθώς και τις συνέπειες που επιφέρει. Σε συνδυασμό με το γεγονός, πως η μορφή των fake news αλλάζει και εξελίσσεται διαχρονικά, γίνεται ακόμα πιο δύσκολος ο σαφής προσδιορισμός της έννοιας (Choraś, και συν., 2019). Για αυτόν τον λόγο, αρκετοί μελετητές έχουν δημιουργήσει διαφορετικές κατατάξεις με σκοπό να ορίσουν όσο το δυνατό πιο ικανοποιητικά τον όρο αυτό.

Στη μελέτη των Simons και Manoilo (Simons & Manoilo, 2021), ο όρος fake news αναφέρεται ως σημασιολογικά ασαφής, αφού η έννοια συνδέεται έντονα με το περιεχόμενο και το επικοινωνιακό πλαίσιο στο οποίο εμφανίζεται. Αρχικά, τα fake news αποτελούσαν ένα μη επιβλαβές φαινόμενο το οποίο περιλάμβανε περιεχόμενο με υπερβολική ερμηνεία της πραγματικότητας και τόνο ενημερωτικό, που ταυτόχρονα είχε σαν σκοπό τη χρήση στη σάτιρα και σε ορισμένα talk shows. Στη συνέχεια, με την πάροδο του χρόνου τα fake news αποτέλεσαν μέσο χειραγώγησης και εξαπάτησης στο πεδίο της πληροφορίας και της ενημέρωσης, με τελικό στόχο την παραπληροφόρηση του κοινού και την επιβολή της άποψης του εκάστοτε δημιουργού. Πλέον, η έννοια αφορά κατά κύριο λόγο και πληροφορία που έρχεται σε αντίθεση με την κοσμοθεωρία ή τις γενικότερες απόψεις, τόσο μεμονωμένων ανθρώπων όσο και ολόκληρων κοινοτήτων. Όμως, στη σύγχρονη εποχή η κατάσταση έχει γίνει πολύ πιο δύσκολη από παλαιότερα χρόνια, γιατί η διασπορά των fake news γίνεται με πολύ μεγαλύτερη ταχύτητα από ότι στο παρελθόν και σε πολύ μεγαλύτερη κλίμακα, αφού εμπλέκει μεγαλύτερο αριθμό ανθρώπων. Αυτό το φαινόμενο, οφείλεται στην έντονη παρουσία των fake news στα περισσότερα κραταιά μέσα κοινωνικής δικτύωσης. Οπότε, δεν είναι εφικτό να οριστεί ξεκάθαρα ο όρος επειδή εξελίσσεται διαχρονικά η σημασία του και ίσως η διάκριση των επιμέρους κατηγοριών του να είναι πιο ουσιαστική και ωφέλιμη. Ακόμα, χρειάζεται να αναφερθεί, πως ο όρος fake news παρουσιάζει επικαλύψεις, με την έννοια misinformation, που αφορά την ακούσια διασπορά ψευδών πληροφοριών, καθώς και την έννοια disinformation, που αφορά την εκούσια διάδοση πληροφοριών που είναι πιθανότατα και ευρέως γνωστό πως είναι ψευδείς.

Ο παραπάνω προσδιορισμός, δεν επιχειρεί να γίνει ακριβής αλλά βασίζεται περισσότερο στην ανάδειξη της συνεχούς εξέλιξης της έννοιας των fake news με την πάροδο του χρόνου με συνέπεια την αδυναμία των ερευνητών να καταλήξουν σε έναν ικανοποιητικό ορισμό. Στην προσπάθειά τους, να ορίσουν την έννοια των fake news στη μελέτη Defining "Fake News" (Tandoc, Wei Lim, & Ling, 2018), οι συγγραφείς κατέληξαν, σε ένα σαφέστερο συμπέρασμα. Έγινε η διαπίστωση, πως τα fake news και οι υποκατηγορίες στις οποίες χωρίζονται, βασίζονται σε δύο βασικά χαρακτηριστικά και στον τρόπο που αυτά εκδηλώνονται. Το πρώτο χαρακτηριστικό, αφορά την αντικειμενικότητα ως προς τα πραγματικά δεδομένα, για παράδειγμα ένα σατιρικό κείμενο βασίζεται σε πραγματικά δεδομένα, τα οποία παραποιούνται ελάχιστα, για να προστεθεί συνήθως υπερβολή. Από την άλλη πλευρά, οι παρωδίες και τα πλήρως κατασκευασμένα γεγονότα βασίζονται σε μια πολύ αφηρημένη αντιμετώπιση των δεδομένων του πραγματικού κόσμου. Το δεύτερο χαρακτηριστικό, αφορά την άμεση πρόθεση του δημιουργού να παραπλανήσει το κοινό, που είναι αποδέκτης του συγκεκριμένου περιεχομένου. Ένα προφανές παράδειγμα, αποτελούν τα σατιρικά fake news, τα οποία έχουν σαν στόχο να διασκεδάσουν τον αναγνώστη και αυτό γίνεται εύκολα αντιληπτό από τον μέσο αποδέκτη και την πλειοψηφία του κοινού. Στον αντίποδα, ένα κατασκευασμένο άρθρο με σκοπό να χειραγωγήσει την κοινή γνώμη αποτελεί περιεχόμενο το οποίο είναι κατασκευασμένο με τέτοιον τρόπο, όπου είναι σχεδόν αδύνατο για τον μέσο άνθρωπο να διαχωρίσει την αλήθεια από το ψέμα.

Με βάση τους ορισμούς και τις πληροφορίες που προηγήθηκαν, είναι εμφανές πως ο όρος fake news δεν είναι σαφής, αφού περικλείει πολλές άλλες έννοιες οι οποίες έχουν να κάνουν με δυσλειτουργίες στη διασπορά πληροφορίας. Το φαινόμενο αυτό, περιγράφεται σε συνδυασμό με έναν εναλλακτικό ορισμό της έννοιας, στη μελέτη των Baptista, J.P. και Gradim A. με τίτλο A Working Definition of Fake News (Baptista & Gradim, 2022). Για αυτό, πολλοί ειδικοί επέλεξαν να σταματήσουν τη χρήση του όρου και να επικεντρωθούν στα επιμέρους προβλήματα που κρύβει όπως π.χ. παραπληροφόρηση. Ωστόσο, άλλοι μελετητές επέλεξαν τη χρήση του όρου ακριβώς επειδή χρησιμοποιείται πλέον από το ευρύ κοινό και παράλληλα συμπεριλαμβάνει πολλές διαφορετικές έννοιες, περιγράφοντας πλέον ένα κοινωνικό πρόβλημα. Το πρόβλημα, εντείνεται με την αντικειμενική παρουσίαση των γεγονότων να είναι λιγότερο σημαντική και σχετική με τη σκοπιά υπό την οποία μελετάται η κοινωνική και πολιτική σφαίρα. Ο ορισμός των fake news και ο διαχωρισμός των υποκατηγοριών στην παραπάνω εργασία, γίνεται όπως και σε παρεμφερή έργα, με βάση την πρόθεση του δημιουργού, το βαθμό των ανακριβειών που συναντώνται και επιπλέον το είδος του επικοινωνιακού πλαισίου που χρησιμοποιείται.

Τελικά, ένας πιο σαφής και απλοϊκός ορισμός, ο οποίος δεν βασίζεται απαραίτητα στις υποκατηγορίες που συνθέτουν την έννοια των fake news, αφορά την εκούσια διασπορά πληροφορίας που έχει σκοπό να κεντρίσει το ενδιαφέρον και να απευθυνθεί στο κοινό με τρόπο οικείο και φυσικό (Simons & Manoilo, 2021). Το περιεχόμενο της πληροφορίας, μπορεί να αφορά αληθή και ψευδή δεδομένα τα οποία μπορεί να είναι μεμονωμένα αποσπάσματα από ομιλίες, συζητήσεις και συνεντεύξεις. Έχουν διάφορους σκοπούς αλλά πιο συχνά έχουν στόχο να διασαλεύσουν την αλήθεια και να προκαλέσουν κάποιο πλήγμα στη φήμη ενός ανθρώπου, προϊόντος ή

γενικότερα να χειραγωγήσουν την άποψη των αποδεκτών πάνω σε ένα θέμα κοινού ενδιαφέροντος.

1.2 Παραδείγματα και πιθανές συνέπειες

Στη συνέχεια, προτού αναλυθούν τα σημαντικότερα χαρακτηριστικά των fake news με βάση τους παραπάνω ορισμούς, θα παρουσιαστούν κάποια παραδείγματα γνωστών περιστατικών που έχουν εμφανιστεί από το παρελθόν έως και σήμερα. Τα περιστατικά αυτά έχουν προκληθεί συνήθως λόγω εσφαλμένων πληροφοριών, οι οποίες σε συγκεκριμένες περιπτώσεις έχουν εξαπλωθεί με κακόβουλο στόχο. Επίσης, παρουσιάζει ιδιαίτερο ενδιαφέρον η ποικιλομορφία που υπάρχει ως προς τις συνέπειες των παραδειγμάτων που θα αναλυθούν παρακάτω.

Μια από τις πιο ενδιαφέρουσες ιστορικές περιόδους αποτελούν τα χρόνια της Ρωμαϊκής Αυτοκρατορίας. Πρόκειται για μια περίοδο, που απουσίαζαν τα μέσα κοινωνικής δικτύωσης, η πρόσβαση στην εκπαίδευση για τη πλειοψηφία του πληθυσμού καθώς και τα συμβατικά μέσα μαζικής ενημέρωσης. Οπότε, ο μόνος τρόπος να πειστεί η κοινή γνώμη για ένα ζήτημα και εντέλει να αποπροσανατολιστεί, θα ήταν αν η πληροφορία προερχόταν από ένα άτομο με εξουσία (Watson, 2018). Η κόντρα του Μάρκου Αντώνιου με τον θετό γιό του Ιούλιου Καίσαρα, Οκτάβιο, χαρακτηρίστηκε από τη χρήση μιας πρωτόγονης μορφής των fake news από τον δεύτερο για να υπερτερήσει στην πολιτική σκηνή. Συγκεκριμένα, φρόντισε να κυκλοφορήσουν νομίσματα με επιγραφές που δημιούργησαν την εντύπωση πως ο Μάρκος Αντώνιος ήταν αλκοολικός και φερέφωνο των απόψεων της Κλεοπάτρας. Ακόμα, έστρεψε τους πολιτικούς άρχοντες εναντίον της Κλεοπάτρας και του Μάρκου Αντώνιου, δημιουργώντας ένα πλαστό αντίγραφο της διαθήκης του, στην οποία αναφερόταν η επιθυμία του να ταφεί μαζί της. Σαν συνέπεια, ο Μάρκος Αντώνιος θεωρήθηκε προδότης, γεγονός που τον οδήγησε στην αυτοκτονία και παράλληλα δημιουργήθηκε η αφορμή για κήρυξη πολέμου εναντίον της Κλεοπάτρας και συνεπώς κατά της Αιγύπτου.

Στο κοντινότερο ιστορικό παρελθόν, έχουν σημειωθεί αρκετές φορές παραδείγματα που αφορούν τέτοια περιστατικά. Γνωστές φιγούρες, όπως ο Βενιαμίν Φραγκλίνος, αποτέλεσαν ακόμα και δημιουργοί δειγμάτων fake news (Watson, 2018). Το 1782, ο προαναφερόμενος πολιτικός διέδωσε το περιστατικό της βίαιης σφαγής κάποιων αποίκων, που βρίσκονταν σε αποστολή εκ μέρους του βασιλιά Γεωργίου Γ’, από αυτόχθονες. Αν και το συγκεκριμένο περιστατικό, ποτέ δεν έλαβε χώρα, ήταν αρκετό ώστε οι εφημερίδες της εποχής, να ξεκινήσουν να αναπαράγουν το γεγονός και τελικά να επιτευχθεί ο στόχος του Βενιαμίν Φραγκλίνου, που δεν ήταν άλλος παρά να στρέψει το πλήθος κατά των αυτόχθονων Αμερικάνων. Έτσι, οι άποικοι ενθάρρυναν τις σφαγές και τον πόλεμο εναντίων των πληθυσμών που ζούσαν στην περιοχή πολύ πριν καταφτάσουν εκείνοι.

Σε πιο σύγχρονες περιόδους και με την εδραίωση των πρώτων υποτυπωδών μέσων μαζικής ενημέρωσης, τα φαινόμενα αυτά εντάθηκαν (Watson, 2018). Οι εφημερίδες που έκαναν την εμφάνισή τους αρχικά, ήταν εξαιρετικά ακριβές και για

αυτόν τον λόγο απευθύνονταν σε κοινό με αντίστοιχη οικονομική άνεση. Ωστόσο, όταν οι τιμές των εφημερίδων έγιναν αρκετά φθηνές και περισσότεροι άνθρωποι απέκτησαν πρόσβαση, παρατηρήθηκε πως οι αναγνώστες, στρέφονταν περισσότερο σε περιεχόμενο το οποίο προκαλούσε συγκινησιακή φόρτιση και περιέργεια. Έτσι δημιουργήθηκε, ο κίτρινος τύπος και ξεκίνησε μια περίοδος κατά την οποία οι τίτλοι περιείχαν αναφορές σε γίγαντες, μυθικά τέρατα και εξωγήινους. Οι πωλήσεις εκτοξεύτηκαν και το φαινόμενο διαιωνίστηκε.

Τα παραπάνω περιστατικά αποτελούν δείγματα του φαινομένου, τα οποία δεν βασίζονται στη δυνατότητα εξάπλωσης της πληροφορίας ταυτόχρονα σε πολύ διευρυμένο κοινό, όπως συμβαίνει σήμερα λόγω του διαδικτύου και των μέσων κοινωνικής δικτύωσης (Goswami, Kaliyar, & Narang, 2021). Όμως, είναι αρκετά προφανές με βάση τα προηγούμενα παραδείγματα πως ακόμα και χωρίς τη συγκεκριμένη δυναμική, οι δημιουργοί πέτυχαν τον στόχο τους, ο οποίος παρουσίαζε μεγάλη ποικιλομορφία. Συγκεκριμένα, έγινε αναφορά σε οικονομικούς ή πολιτικούς σκοπούς και γενικότερα στην επιβολή της άποψης του δημιουργού, προς τους αποδέκτες του περιεχομένου. Οι πραγματικές διαστάσεις του προβλήματος, φαίνονται περισσότερο σήμερα παρά ποτέ άλλοτε και αυτό συμβαίνει γιατί το κύριο μέσο διάδοσης των fake news είναι το διαδίκτυο, το οποίο χρησιμοποιείται για την εργασία, την ενημέρωση, την ψυχαγωγία, την εξυπηρέτηση των καταναλωτικών αναγκών και τη διαμόρφωση γνώμης (Shu, Wang, Silva, Tang, & Liu, 2017).

Ένα παράδειγμα fake news, με πολύ σοβαρές συνέπειες στον τομέα της δημόσιας υγείας, εμφανίστηκε κατά την περίοδο της πανδημίας COVID-19. Το φαινόμενο αυτό, περιγράφεται στο έργο των Rocha, Y.M., de Moura, G.A., Desidério, G.A., που αναλύονται οι επιπτώσεις των fake news στο συγκεκριμένο χρονικό διάστημα (Rocha, 2023). Λόγω των μέσων κοινωνικής δικτύωσης υπήρξε ταχεία και ευρεία εξάπλωση των fake news σε μια περίοδο που παρατηρήθηκαν αυξημένα περιστατικά κατάθλιψης, διαταραχών άγχους και αυτοκτονιών. Η αρνητική επίδραση των fake news ωστόσο, πηγάζει κυρίως από τη διασπορά πληροφοριών που στήριξαν θεωρίες συνομωσίας. Κάποιες από τις πιο δημοφιλείς, αφορούσαν τη χρήση του ιού ως βιολογικό όπλο από την πλευρά της Κίνας καθώς και την αντιμετώπιση του ιού με φάρμακα που χρησιμοποιούνται σε παρόμοιες παθήσεις, χωρίς να υπάρχει κάποιο επιστημονικό έρεισμα για ίαση (Rocha, 2023). Το αποτέλεσμα ήταν να υπάρξουν χώρες όπου δημιουργήθηκαν ελλείψεις σημαντικών φαρμακευτικών σκευασμάτων, με αποτέλεσμα άνθρωποι με αυτοάνοσα νοσήματα και ρευματοπάθειες, να πρέπει να αλλάξουν φαρμακευτικές αγωγές. Ακόμα, πολλοί άνθρωποι, όπως για παράδειγμα στη Νιγηρία, πέθαναν λόγω υπερβολικής δόσης υδροξυχλωροκίνης (Rocha, 2023). Παράλληλα, υπήρξε μαζική υστερία και πανικός, για προμήθεια ειδών πρώτης ανάγκης.

Η πιο συνήθης θεματολογία fake news αφορά την πολιτική. Πολλές φορές τα fake news έχουν χρησιμοποιηθεί από ορισμένες χώρες για τη διαιώνιση αβάσιμων εντυπώσεων στο πλαίσιο εξωτερικών πολιτικών στηριγμένων στην προπαγάνδα (P. Ksieniewicz, 2020). Χαρακτηριστικό παράδειγμα, αποτελεί η Ρωσία η οποία έχει υιοθετήσει ως πάγια τακτική τη διασπορά ψευδών ειδήσεων, για να πετύχει τον

αποπροσανατολισμό του κοινού. Σαν απάντηση σε αυτήν την τακτική, οι χώρες της Ευρωπαϊκής Ένωσης δημιούργησαν την ιστοσελίδα EUvsDisnfo[1], η οποία έχει σαν ρόλο να καταγράφει τέτοια περιστατικά και τελικά επιδιώκει να σταματήσει τη διαιώνιση των συνεπειών που προκύπτουν από καμπάνιες παραπληροφόρησης. Επιπλέον, σχετικά πρόσφατα σε εκλογές στη Γαλλία και τις Ηνωμένες Πολιτείες Αμερικής έχουν υπάρξει υπόνοιες σχετικά με τη χειραγώγηση του εκλογικού σώματος (P. Ksieniewicz, 2020). Αυτός ο ισχυρισμός, εικάζεται πως ευσταθεί λόγω της διασποράς ψευδών ειδήσεων σε κοινότητες των μέσων κοινωνικής δικτύωσης που είχαν σκοπό να στρέψουν τους ψηφοφόρους προς συγκεκριμένους υποψήφιους για το εκάστοτε αξίωμα.

1.3 Μορφή των fake news και του περιεχομένου τους

Έως τώρα, έχει γίνει αναφορά στην έννοια των fake news, όμως παρατηρείται ασάφεια και ως προς τη μορφή τους. Είναι λογικό να υφίσταται το συγκεκριμένο φαινόμενο, αφού χωρίζονται σε διάφορες κατηγορίες και πολλές φορές δεν εντάσσονται αποκλειστικά σε μια από αυτές. Ο τύπος δεδομένων του περιεχομένου είναι μια συνιστώσα της μορφής των fake news και στη συνέχεια θα παρουσιαστεί με συντομία (Mishra, Shukla, & Agarwal, 2022):

- **Κείμενο:** Ο βασικότερος τύπος δεδομένων που εμφανίζεται είναι το κείμενο. Δεν αφορά αποκλειστικά τη χρήση της γλώσσας ως μέσο επικοινωνίας με τους αποδέκτες, αλλά την επιρροή της συνδυαστικά με τη χρήση της γραμματικής και του τόνου στην έκφραση του συγγραφέα. Όπως θα εξηγηθεί σε επόμενο τμήμα της εργασίας, το κείμενο και τα ιδιαίτερα χαρακτηριστικά του είναι πολύ βοηθητικά για την επικύρωση εγκυρότητας ενός δείγματος.

- **Πολυμέσα:** Ένας πολύ σημαντικός παράγοντας αναφορικά με τη διάδοση των fake news είναι να προκαλέσουν το ενδιαφέρον του αναγνώστη. Για να διασφαλιστεί αυτό, πολλές φορές χρησιμοποιούνται εικόνες, βίντεο, μουσική και διάφορα γραφικά με σκοπό, να κεντρίσουν την προσοχή του αποδέκτη όσο το δυνατό πιο αποδοτικά.

- **Ενσωματωμένο περιεχόμενο και υπερσυνδέσεις:** Για να προσδώσουν κύρος στα λεγόμενά τους, οι δημιουργοί των fake news, χρησιμοποιούν συνδέσεις με άλλα κείμενα, τα οποία επιβεβαιώνουν τα λεγόμενά τους. Παράλληλα, μπορεί να χρησιμοποιηθεί ενσωματωμένο περιεχόμενο (Twitter quote, Facebook post, YouTube video κ.α.) το οποίο προέρχεται απευθείας από κάποιο μέσο κοινωνικής δικτύωσης. Ωστόσο, πολλές φορές η εγκυρότητα των πηγών αυτών δεν είναι εξακριβωμένη και περισσότερο στην δεύτερη περίπτωση

[1] https://euvsdisinfo.eu

όπου το ενσωματωμένο περιεχόμενο μπορεί να προέρχεται από πληθώρα πηγών και δημιουργών.

- **Ήχος:** Ο συγκεκριμένος τύπος δεδομένων, αποτελεί μια ιδιαίτερη υποκατηγορία των πολυμέσων. Αξίζει η αναφορά του ήχου, ως ξεχωριστή κατηγορία, επειδή αποτελεί έναν ιδιαίτερο τρόπο μεταφοράς ειδήσεων. Αυτό συμβαίνει, διότι υπάρχουν μέσα διάδοσης που βασίζονται αποκλειστικά στον ήχο (π.χ. ραδιόφωνο).

Η δεύτερη συνιστώσα, η οποία επηρεάζει τη μορφή του περιεχομένου σημασιολογικά, αποτελεί το πλαίσιο εμφάνισης των fake news (Shu, Wang, Silva, Tang, & Liu, 2017). Σε παραδοσιακά μέσα μαζικής ενημέρωσης, επιστρατεύονται μέθοδοι που έχουν σαν υπόβαθρο βασικές αρχές της ψυχολογίας και των κοινωνικών αλληλεπιδράσεων. Ωστόσο, όταν γίνεται αναφορά σε fake news που διαδίδονται στα μέσα κοινωνικής δικτύωσης, η αποτελεσματικότητά τους βασίζεται σε στοιχεία όπως κακόβουλους λογαριασμούς χρηστών και το φαινόμενο του θαλάμου αντήχησης (Echo Chamber Effect). Στη συνέχεια, θα περιγραφούν εκτενέστερα, οι συγκεκριμένοι μηχανισμοί.

Υπάρχουν πολλοί παράγοντες που εξηγούν το ψυχολογικό υπόβαθρο με βάση το οποίο καθορίζεται το περιεχόμενο, στο οποίο είναι ευάλωτοι οι αποδέκτες, ωστόσο θα γίνει αναφορά στους δύο επικρατέστερους. Το φαινόμενο του αφελούς ρεαλισμού (Naïve Realism) περιγράφει την τάση των ανθρώπων, να θεωρούν τις προσωπικές τους απόψεις ως αυτονόητες και ρεαλιστικές. Ως επακόλουθο, υποτιμούν τις απόψεις άλλων ανθρώπων που διαφέρουν σημαντικά και τις αποδίδουν σε ελλιπή ενημέρωση, προκαταλήψεις και συναισθηματική φόρτιση. Ο δεύτερος μηχανισμός, γνωστός ως πλάνη της επιβεβαίωσης (Confirmation Bias) έχει να κάνει με το γεγονός πως η επιβεβαίωση κυρίαρχων ατομικών προκαταλήψεων και στερεοτύπων είναι πιθανότερο να είναι πιστευτή. Άρα, είναι προφανές πως περιεχόμενο που επιβεβαιώνει προκαταλήψεις ή στρέφεται εναντίον αντιλήψεων που δεν συνάδουν με τον αποδέκτη είναι πιθανώς πιο πιστευτό. Όμως, ακόμα και η προσπάθεια διόρθωσης των όποιων προκαταλήψεων μπορεί να αποβεί προβληματική, αφού έχει παρατηρηθεί πως οδηγεί σε σύγχυση και παρερμηνείες.

Το περιεχόμενο ωστόσο, δημιουργείται με γνώμονα και τις βασικές αρχές, που διέπουν την αλληλεπίδραση του ατόμου με τον κοινωνικό περίγυρο. Αυτό συμβαίνει, γιατί οι δημιουργοί γνωρίζουν πως οι άνθρωποι σε μια κοινότητα προκειμένου να ισχυροποιήσουν τη θέση τους θα ακολουθήσουν συγκεκριμένα μοτίβα συμπεριφοράς. Δηλαδή, συχνά οι αποδέκτες του εκάστοτε περιεχομένου υποστηρίζουν και διαδίδουν πληροφορίες στις οποίες δεν πιστεύουν, γιατί επιδιώκουν την επιβεβαίωση ως μέλη μιας ευρύτερης κοινότητας. Με αυτό τον τρόπο, διασπείρονται ψευδείς ειδήσεις, ακόμα και αν υπάρχουν υποψίες ως προς το ποιόν τους.

Σχετικά με τη μορφή των fake news στα μέσα κοινωνικής δικτύωσης, οι μηχανισμοί που εξασφαλίζουν την αποτελεσματικότητα διαφέρουν αρκετά από τη διάδοση στα παραδοσιακά μέσα ενημέρωσης. Αρχικά, θα αναλυθεί η επίδραση των κακόβουλων χρηστών. Οι χρήστες αυτοί χωρίζονται σε τρεις κατηγορίες social bots, cyborg users και trolls. Η κατηγορία των social bots, αφορά λογαριασμούς

ελεγχόμενους πλήρως από αλγορίθμους, που παράγουν αυτοματοποιημένο περιεχόμενο αλληλοεπιδρώντας μεταξύ τους αλλά και με πραγματικούς χρήστες. Η δεύτερη κατηγορία, αφορά λογαριασμούς που έχουν δημιουργηθεί από κάποιον άνθρωπο αλλά μεγάλο μέρος των διαδικασιών παραγωγής περιεχομένου και αλληλεπίδρασης γίνεται αυτοματοποιημένα από αλγορίθμους. Η τελευταία κατηγορία, στοχεύει επιμέρους κοινότητες ανθρώπων στο διαδίκτυο και προσπαθεί να προκαλέσει στους αποδέκτες συγκεκριμένες αντιδράσεις. Οι λογαριασμοί αυτού του είδους ανήκουν σε οργανικούς χρήστες οι οποίοι μπορεί να δρουν επί πληρωμή και συνήθως προκαλούν αρνητικά συναισθήματα, όπως φόβο και θυμό, για να οδηγήσουν το κοινό στη σύγχυση, στην έλλειψη εμπιστοσύνης και σε παράλογες συμπεριφορές. Στις προεδρικές εκλογές των ΗΠΑ το 2016, στοιχεία δείχνουν πως είχαν προσληφθεί τουλάχιστον 1000 τέτοιοι χρήστες, για διασπορά ψευδών ειδήσεων εις βάρος υποψηφίου (Shu, Wang, Silva, Tang, & Liu, 2017). Ενώ, συνολικά οι εκλογές συζητήθηκαν στο Twitter από τουλάχιστον 19 εκατομμύρια social bots χρήστες (Shu, Wang, Silva, Tang, & Liu, 2017). Οπότε, είναι προφανής η διαφορά που υπάρχει συγκριτικά με τα παραδοσιακά μέσα μαζικής ενημέρωσης. Το περιεχόμενο παράγεται μαζικά και πιθανότατα αυτοματοποιημένα ανάλογα με το είδος του κακόβουλου χρήστη.

Τέλος, ένας επιπλέον παράγοντας ο οποίος καθορίζει την μορφή του περιεχομένου, που εμφανίζεται στους χρήστες των μέσων κοινωνικής δικτύωσης, είναι το φαινόμενο του θαλάμου αντήχησης (Echo Chamber Effect) (Shu, Wang, Silva, Tang, & Liu, 2017). Το μοντέλο παραγωγής περιεχομένου αλλά και κατανάλωσης, είναι πολύ διαφορετικό συγκριτικά με τα συμβατικά μέσα ενημέρωσης, αφού πλέον η δημιουργία του δεν γίνεται απαραίτητα από κάποιον ειδικευμένο μεσάζοντα, όπως για παράδειγμα τους δημοσιογράφους. Οι χρήστες ως καταναλωτές, ακολουθούν επιλεκτικά συγκεκριμένους δημιουργούς περιεχομένου, με τους οποίους ταυτίζονται και έχουν παρόμοιες ιδεολογίες. Συνεπώς, το περιεχόμενο που παρουσιάζεται επιβεβαιώνει και ενισχύει τις αντιλήψεις των συγκεκριμένων αποδεκτών της πληροφορίας ακόμα και αν εμφανίζονται λανθασμένες πεποιθήσεις. Έτσι, σχηματίζονται κοινότητες ομοϊδεατών, στις οποίες αντηχούν παρόμοιες ιδεολογίες.

Η αρνητική απόρροια του παραπάνω φαινομένου, είναι αρκετά σημαντική ως προς τη μορφή και το είδος των fake news στα μέσα κοινωνικής δικτύωσης. Έχει παρατηρηθεί, πως πληροφορίες που θεωρούνται έγκυρες, από μικρότερες επιμέρους ομάδες αναπαράγονται και υιοθετούνται από το εκάστοτε μέλος τους ακόμα και όταν υπάρχουν ελλιπή στοιχεία, που υποστηρίζουν την ορθότητα της είδησης. Επίσης, ακριβώς το ίδιο συμβαίνει και με απόψεις, που απλά εμφανίζονται με υψηλή συχνότητα, ακόμα και αν αποτελούν ψευδείς πληροφορίες. Οπότε, το αποτέλεσμα είναι η δημιουργία πολλών τέτοιων επιμέρους ομογενών κοινοτήτων, οι οποίες όμως κατά τη διαδικασία διάχυσης της πληροφορίας σε μέσα κοινωνικής δικτύωσης οδηγούν τους υπόλοιπους χρήστες σε πόλωση, αφού υποστηρίζουν έντονα διαφορετικές απόψεις τις οποίες δεν έχουν πραγματικά εξακριβώσει.

1.4 Μέσα διάδοσης των fake news

Όπως, έχει ήδη ειπωθεί, τα fake news μπορεί να έχουν τη μορφή πολλών διαφορετικών τύπων δεδομένων. Ως λογική συνέπεια, γίνεται αντιληπτό πως ανάλογα με τον τύπο δεδομένων θα χρησιμοποιούνται και διαφορετικές πλατφόρμες για τη μετάδοσή τους (Mishra, Shukla, & Agarwal, 2022). Μελετώντας εκτενέστερα τα πιθανά μέσα διάδοσης, θα γίνει πιο φανερό σε ποια περίπτωση εμφανίζεται το μεγαλύτερο κοινό αποδεκτών και παράλληλα υποστηρίζονται οι περισσότεροι τύποι δεδομένων. Με αυτόν τον τρόπο, θα γίνει πιο κατανοητός ο λόγος που τα μέσα κοινωνικής δικτύωσης ως μέσο διάδοσης αποτελούν τομέα ενδιαφέροντος από πολλούς μελετητές. Τα πιο διαδεδομένα δίκτυα διαμοιρασμού πληροφορίας είναι τα ακόλουθα

1. **Μεμονωμένες ιστοσελίδες:** Η πρώτη και πιθανότατα πιο προφανής κατηγορία μέσου μετάδοσης είναι οι απλές ιστοσελίδες. Οι περισσότερες ιστοσελίδες εμπίπτουν σε μια από τις κατηγορίες των blogs, media και γνωστών ιστοσελίδων ενημέρωσης. Οι ιστοσελίδες ενημέρωσης είναι γνωστές συνήθως για το βαθμό αξιοπιστίας τους στο ευρύ κοινό και συχνά υπόκεινται σε κανόνες δημοσιογραφικής δεοντολογίας. Έτσι, δεν έχουν μεγάλη επίδραση στην διάδοση των fake news, όμως κάτι τέτοιο δεν μπορεί να ισχύσει για τις άλλες δύο κατηγορίες. Τα blogs βασίζονται στους χρήστες για την δημιουργία περιεχομένου, χωρίς απαραίτητα να υπάρχει κάποια επίβλεψη. Το γεγονός αυτό είναι βασικός παράγοντας επιδείνωσης των περιστατικών προώθησης παραπληροφόρησης. Οι ιστοσελίδες της κατηγορίας media (π.χ. Google) ασχολούνται με την δημιουργία ή και προώθηση ιστοσελίδων των πελατών τους. Οι ιστοσελίδες αυτές, έχουν υλικό και πληροφορίες που ποικίλουν ανάλογα τον πελάτη και δεν υπάρχουν αυστηρές δικλείδες ελέγχου για την εγκυρότητα του περιεχομένου.

2. **Μέσα κοινωνικής δικτύωσης:** Ο κύριος λόγος για την επικράτηση των fake news στα μέσα κοινωνικής δικτύωσης εντοπίζεται στο μεγάλο εύρος του κοινού και την αυξημένη ταχύτητα διάδοσης, που μπορεί να επιτευχθεί. Οι καθημερινές ειδήσεις και τα γεγονότα της επικαιρότητας, κοινοποιούνται από το 70% των αποδεκτών τους στα social media (Mishra, Shukla, & Agarwal, 2022). Δεν αποκλείεται όμως να συμπεριλαμβάνονται προσωπικές απόψεις των αποδεκτών και αλλοιώσεις. Για αυτόν τον λόγο, είναι πιθανό να δημιουργηθούν παρεξηγήσεις ως προς το νόημα του εκάστοτε περιεχομένου.

3. **Ηλεκτρονική αλληλογραφία:** Η ηλεκτρονική αλληλογραφία, θεωρείται ως ένα αξιόπιστο μέσο επικοινωνίας. Επιπλέον, αποτελεί και αξιόπιστο μέσο μεταφοράς ειδήσεων, όμως αυτό που συχνά δεν γίνεται αντιληπτό είναι πως η εξακρίβωση της εγκυρότητας του περιεχομένου παραμένει πρόβλημα, με αποτέλεσμα πολλοί αποδέκτες να πέφτουν θύματα παραπληροφόρησης. Δηλαδή, αν και θεωρείται

ένα ασφαλές μέσο για την επικοινωνία δεν σημαίνει πως απαραίτητα προσφέρει και εγγυήσεις για την εγκυρότητα του περιεχομένου που διακινείται.

4. **Podcast:** Το πλήθος των ανθρώπων και των ηλικιών, που επιλέγουν το συγκεκριμένο μέσο για την ενημέρωσή τους είναι πολύ περιορισμένο. Ωστόσο, αποτελεί ιδιαίτερη πρόκληση η εξακρίβωση της εγκυρότητας, αφού χρησιμοποιούνται αποκλειστικά δεδομένα ήχου, τα οποία είναι πολύ πιο περίπλοκο να μελετηθούν συγκριτικά με το απλό κείμενο και άλλους τύπους πολυμέσων.

5. **Ραδιόφωνο:** Το συμβατικό ή και διαδικτυακό ραδιόφωνο είναι πηγή ενημέρωσης, για αρκετά μεγάλο μέρος του πληθυσμού, λόγω του βαθμού ενσωμάτωσης του συγκεκριμένου μέσου διάδοσης στην καθημερινότητα. Όμως, αν και απασχολεί μεγαλύτερο μέρος αποδεκτών συγκριτικά με τα podcasts, έχει το ίδιο μειονέκτημα. Η πιστοποίηση εγκυρότητας αποκλειστικά σε δεδομένα ήχου είναι ιδιαίτερη πρόκληση.

Είναι φανερό πως υπάρχει μεγάλη ποικιλία τόσο ως προς τον τύπο των δεδομένων, όσο και ως προς τα μέσα μετάδοσης. Ωστόσο, η πλειοψηφία των περιστατικών βασίζεται κυρίως στο κείμενο, το οποίο μπορεί να χρησιμοποιηθεί με πολύ περίπλοκο τρόπο ανάλογα με την κατηγορία των fake news που εντάσσεται κάποιο δείγμα. Παράλληλα, τα δύο μέσα με την μεγαλύτερη επιρροή ανάλογα με τον όγκο της πληροφορίας που προσφέρουν και την δημοτικότητά τους, είναι οι μεμονωμένες ιστοσελίδες και τα μέσα κοινωνικής δικτύωσης.

1.5 Κατηγορίες των fake news

Παρακάτω θα μελετηθούν οι διαφορετικές κατηγορίες fake news που υπάρχουν. Η κατηγοριοποίηση είναι διαφορετική ανάλογα με τον κάθε ερευνητή, ωστόσο οι διαφορές είναι συνήθως μικρές όπως θα διαπιστωθεί στη συνέχεια. Πρώτα, θα μελετηθούν οι κατηγορίες με βάση τα μέσα που επιστρατεύονται για να πειστεί το κοινό (Ahmed, Hinkelmann, & Corradini, 2020).

- **Ψευδής Σύνδεση (False Connection):** Η συγκεκριμένη κατηγορία αφορά τις περιπτώσεις όπου οι επικεφαλίδες, οι λεζάντες και τα γραφικά στοιχεία στο κείμενο δεν συνάδουν με το κυρίως περιεχόμενο.

- **Ψευδές Πλαίσιο (False Context):** Η κατηγορία αυτή έχει να κάνει με περιεχόμενο που περιγράφει αληθή πληροφορία, όμως παρουσιάζεται σκόπιμα σε ένα λανθασμένο πλαίσιο. Έτσι μεταβάλλεται η σημασία των γεγονότων που αναφέρονται.

- **Αλλοιωμένο Περιεχόμενο (Manipulated Content):** Αυτό το είδος των fake news αφορά δείγματα, όπου γνήσιες πληροφορίες και εικόνες, αλλοιώνονται για να παραπλανήσουν τους αποδέκτες.

- **Σάτιρα (Satire):** Πρόκειται για περιεχόμενο, που δεν αποσκοπεί να βλάψει τους αποδέκτες ωστόσο είναι αρκετά πιθανό να τους παραπλανήσει. Η συνέπεια αυτή μπορεί να προκύψει όταν το κοινό δεν αντιλαμβάνεται την πρόθεση του δημιουργού για σάτιρα.

- **Αποπροσανατολιστικό Περιεχόμενο (Misleading Content):** Η κατηγορία αυτή, αφορά περιεχόμενο που αποσκοπεί στην παρουσίαση των πληροφοριών με τρόπο που θα χειραγωγήσει και θα οδηγήσει σε λανθασμένες εντυπώσεις το κοινό.

- **Περιεχόμενο Ψευδούς Αυθεντίας (Imposter Content):** Ο συγκεκριμένος τύπος fake news, έχει ως στόχο να ισχυριστεί πως ο συγγραφέας είναι αυθεντία σε έναν τομέα, για να προσδώσει επιπλέον αξιοπιστία στο περιεχόμενο. Με αυτό τον τρόπο οι αποδέκτες θα νιώθουν πιο ασφαλείς για την εγκυρότητα της πληροφορίας που λαμβάνουν.

- **Κατασκευασμένο Περιεχόμενο (Fabricated Content):** Πρόκειται για εντελώς πρωτότυπο περιεχόμενο, όμως είναι πιθανότατα εντελώς ανυπόστατο και αποσκοπεί στη πρόκληση προβλημάτων.

Άλλες κατηγοριοποιήσεις (Mishra, Shukla, & Agarwal, 2022), είναι λιγότερο αναλυτικές σε σχέση με την παραπάνω, όμως ταυτόχρονα είναι εξίσου ικανοποιητικές για τον χαρακτηρισμό των δειγμάτων. Για παράδειγμα, ο διαχωρισμός των fake news μπορεί να οριστεί στις κατηγορίες κατασκευασμένων νέων (Fabrication), φημών (Hoax) και σάτιρας (Satire). Τα (Fabricated) δείγματα κατασκευασμένων νέων συνήθως παραλείπουν σκόπιμα πληροφορία και προέρχονται από μόνο μια συγκεκριμένη πηγή. Τις περισσότερες φορές ο δημιουργός είναι ενημερωμένος πως πρόκειται για ψευδείς πληροφορίες. Η επιτυχία τους βασίζεται στη χρήση της τακτικής clickbait. Η τακτική αυτή, επιστρατεύει τίτλους που προκαλούν συγκινησιακή φόρτιση και περιέργεια στον αναγνώστη, έτσι ώστε να τον ωθήσουν να διαβάσει το περιεχόμενο του άρθρου. Τα δείγματα της κατηγορίας (Hoaxes) των φημών συγκριτικά με τα κατασκευασμένα νέα, κάνουν χρήση πιο πολύπλοκων μηχανισμών για να χειραγωγήσουν το κοινό. Για παράδειγμα, διαχέονται στο διαδίκτυο και τα μέσα κοινωνικής δικτύωσης από πολλές πηγές ταυτόχρονα και για αυτό το λόγο οι αποδέκτες είναι πιο πιθανό να πειστούν. Τέλος, η κατηγορία (Satire) της σάτιρας αναφέρεται σε περιεχόμενο με χιουμοριστική χροιά, ωστόσο είναι πολύ πιθανό να περιέχει πληροφορίες για θέμα άγνωστο στον αναγνώστη. Σαν απόρροια του φαινομένου μπορεί να υπάρξουν παρεξηγήσεις ως προς την ορθότητα του περιεχομένου.

Μια τελευταία κατηγοριοποίηση (Shu, Wang, Silva, Tang, & Liu, 2017) που παρουσιάζει ενδιαφέρον είναι ο διαχωρισμός των fake news σε θεωρίες συνομωσίας που είναι εύκολο να αποδειχτεί πως δεν ευσταθούν, παραπληροφόρηση η οποία είναι

εκούσια, σάτιρα που δεν παρέχει το πλήρες πλαίσιο της πληροφορίας καθώς και φήμες που δεν ισχύουν και συνήθως αφορούν την επικαιρότητα. Ο διαχωρισμός των fake news με αυτό τον τρόπο είναι αρκετά όμοιος με την κατηγοριοποίηση που αναλύθηκε προηγουμένως. Ωστόσο, συνδυάζοντας τις πιθανές κατηγοριοποιήσεις μπορεί ο κάθε μελετητής να καταλήξει σε ασφαλή συμπεράσματα ως προς τα πραγματικά δείγματα που αποτελούν fake news.

1.6 Μέθοδοι προσέγγισης του προβλήματος

Η επικύρωση περιεχομένου, αποτελεί ένα πρόβλημα το οποίο μελετάται από διαφορετικούς επιστημονικούς κλάδους. Οι προσεγγίσεις που ακολουθούνται είναι λογικό να διαφέρουν αρκετά και να εξετάζουν υπό διαφορετικό πρίσμα το εκάστοτε δείγμα (Shu, Wang, Silva, Tang, & Liu, 2017). Στη συνέχεια, θα αναλυθούν κάποιες από αυτές τις προσεγγίσεις. Στις πρώτες δύο παραγράφους, αναλύονται οι τακτικές που χρησιμοποιούνται αποκλειστικά για το περιεχόμενο και στις επόμενες τρεις εκείνες που αξιοποιούνται όταν διαχέεται στα μέσα κοινωνικής δικτύωσης.

Η πιο προφανής προσέγγιση, έχει να κάνει με την ταυτοποίηση συγκεκριμένων γλωσσολογικών στοιχείων (linguistic-based) (Shu, Wang, Silva, Tang, & Liu, 2017). Οι τίτλοι (clickbait) συχνά προδίδουν τις ψευδείς ειδήσεις σε συνδυασμό με κάποια καυστικά στοιχεία στον τόνο του δημιουργού, ο οποίος υποστηρίζει υπερβολικά έντονα τις απόψεις του. Ωστόσο, τα fake news περιέχουν κάποια ιδιαίτερα χαρακτηριστικά ως προς το λεξιλόγιο, που τελικά τα προδίδουν. Για παράδειγμα, παρατηρείται χρήση συγκεκριμένου μέσου αριθμού συνολικών λέξεων, μέσου αριθμού χαρακτήρων ανά λέξη και συχνότητα χρήσης ορισμένων όρων. Επίσης, η σύνταξη των κειμένων αυτών καθώς και η στίξη αποτελούν χρήσιμους δείκτες για τον εντοπισμό των fake news. Τέλος, η εμφάνιση ειδήσεων και γενικότερα πληροφοριών που αφορούν για παράδειγμα τον τομέα των ειδήσεων χαρακτηρίζεται από συγκεκριμένο μέσο μήκος παραγράφων, μέσο αριθμό παραγράφων και σαφώς καθορισμένη χρήση εξωτερικών συνδέσμων.

Μια δεύτερη προσέγγιση, μελετά κυρίως την όψη του περιεχομένου (visual-based) (Shu, Wang, Silva, Tang, & Liu, 2017). Τα fake news, βασίζονται σε εικόνες και λοιπά γραφικά για να προκαλέσουν την συναισθηματική αντίδραση του αποδέκτη. Η προσέγγιση, βασίζεται σε συγκεκριμένα γνωρίσματα όπως για παράδειγμα οι διαστάσεις και ο αριθμός των εικόνων που χρησιμοποιούνται. Ακόμα, μελετά και τα γενικότερα χαρακτηριστικά των εικόνων που είναι εξίσου σημαντικά. Όπως, το πόσο όμοιες είναι οι εικόνες μεταξύ τους, το πόσο συνάδουν με το υπόλοιπο περιεχόμενο, το κατά πόσο διαφέρουν μεταξύ τους και πόσο ξεκάθαρες είναι ως προς το νόημα που μεταφέρουν στον αποδέκτη. Οι εικόνες, τα βίντεο κ.τ.λ. εκτός από ένα χρήσιμο εργαλείο για την αποτελεσματικότητα των fake news μπορούν να βοηθήσουν και στην ταυτοποίηση περιεχομένου, που στοχεύει στην εξάπλωση της παραπληροφόρησης.

Επιπλέον, μια ακόμα προσέγγιση του προβλήματος αφορά τα χαρακτηριστικά των χρηστών στα μέσα κοινωνικής δικτύωσης (user-based) (Shu, Wang, Silva, Tang,

& Liu, 2017). Όπως έχει αναλυθεί παραπάνω, υπάρχουν διάφορες κατηγορίες κακόβουλων χρηστών οι οποίοι έχουν διαφορετικό προφίλ. Η ανάλυση των χαρακτηριστικών τους μπορεί να πραγματοποιηθεί σε ατομικό επίπεδο με σκοπό την εκτίμηση της αξιοπιστίας τους, όπου για κάθε χρήστη αξιολογείται η ηλικία κατά την εγγραφή, ο αριθμός των ακολούθων και ο αριθμός των δημοσιεύσεων. Επιπροσθέτως, η ανάλυση των χαρακτηριστικών μπορεί να γίνει και σε ομαδικό επίπεδο, όπου για συγκεκριμένες κοινότητες χρηστών εξάγεται ο μέσος όρος των στοιχείων που αναφέρθηκαν προηγουμένως. Έτσι, εξάγονται συμπεράσματα για ολόκληρες ομάδες δημιουργών περιεχομένου και προάγεται η διαδικασία εξακρίβωσης της εγκυρότητας.

Επίσης, μια ακόμα αντιμετώπιση του προβλήματος αφορά τη μελέτη των χαρακτηριστικών των ίδιων των δημοσιεύσεων (post-based) (Shu, Wang, Silva, Tang, & Liu, 2017). Είναι πολύ πιθανό, να εντοπιστούν fake news εξετάζοντας τα χαρακτηριστικά της δημοσίευσης κατά την αλληλεπίδραση με το κοινό και της αντίστοιχης ανταπόκρισής. Τα χαρακτηριστικά αυτά, μπορεί να αφορούν μεμονωμένες δημοσιεύσεις εξετάζοντας την στάση του κοινού ως προς τη δημοσίευση, την αξιοπιστία της και το θέμα που απασχολεί. Όμως, μπορεί να γίνει και μελέτη ολόκληρων ομάδων από δημοσιεύσεις η οποία προσφέρει ακόμα πιο ασφαλή αποτελέσματα. Το τελευταίο είδος στοιχείων, που εξετάζονται σε αυτή την προσέγγιση είναι οι αλλαγές στις δημοσιεύσεις με την πάροδο του χρόνου. Μια μέθοδος η οποία έχει εξεταστεί είναι η χρήση μη επιβλεπόμενης μάθησης για την ανάλυση των αλλαγών στις δημοσιεύσεις σε ένα συγκεκριμένο χρονικό διάστημα.

Τέλος, έχει πραγματοποιηθεί μελέτη και αναγνώριση των fake news με χρήση ανάλυσης δικτύων (network-based) (Shu, Wang, Silva, Tang, & Liu, 2017). Τα κοινωνικά δίκτυα και οι χρήστες τους μπορούν να παρασταθούν ως γράφοι. Το ίδιο ισχύει για τα νέα και τις μεταξύ τους συσχετίσεις. Η τακτική αυτή επιτρέπει την αναγνώριση του τρόπου διάχυσης της πληροφορίας και του τρόπου με τον οποίο συνδέονται μεταξύ τους οι χρήστες ανάλογα με το περιεχόμενο που δημοσιεύουν. Τα συμπεράσματα, προκύπτουν από τις μετρικές των γράφων που χρησιμοποιούνται σε πεδία όπως η ανάλυση κοινωνικών δικτύων. Τέτοιες μετρικές είναι για παράδειγμα ο βαθμός κόμβου (node degree) και ο συντελεστής ομαδοποίησης (clustering coefficient).

Με βάση τα χαρακτηριστικά της κάθε προσέγγισης, υπάρχουν και διαφορετικοί μηχανισμοί αναγνώρισης των fake news. Οι δύο βασικές κατηγορίες στις οποίες θα γίνει αναφορά είναι εκείνες που βασίζονται σε μηχανισμούς συστημάτων γνώσης (knowledge-based) και μηχανισμούς βασισμένους στην ανάλυση του συγγραφικού ύφους (style-based) (Shu, Wang, Silva, Tang, & Liu, 2017). Οι τρόποι ανίχνευσης αυτού του είδους εφαρμόζονται αποκλειστικά στο περιεχόμενο χωρίς να δίνεται βαρύτητα στα χαρακτηριστικά που προσδίδονται από τη διάθεση του στα μέσα κοινωνικής δικτύωσης.

Οι τρόποι ανίχνευσης της κατηγορίας μεθόδων που άπτονται των συστημάτων γνώσης (knowledge-based) είναι ίσως η πιο προφανής οδός ταυτοποίησης των fake news. Οι μέθοδοι που αξιοποιούνται βασίζονται στην εξακρίβωση των πληροφοριών που διαδίδονται πραγματοποιώντας έλεγχο της ορθότητας του πλαισίου και του περιεχομένου. Δηλαδή, δεν γίνεται απλά κάποια εκτίμηση με βάση οποιεσδήποτε

μετρικές, αλλά ελέγχεται η αξιοπιστία αποκλειστικά με κριτήριο προϋπάρχοντα δεδομένα και πολλαπλές απόψεις που συμπληρώνονται μεταξύ τους. Η πρώτη μέθοδος, βασίζεται στη μελέτη του περιεχομένου από κάποια αυθεντία στον εκάστοτε τομέα (expert-oriented fact-checking). Το πρόβλημα με τη μέθοδο αυτή εντοπίζεται στο γεγονός πως είναι ιδιαιτέρως χρονοβόρα και δεν είναι εύκολο να κλιμακωθεί. Η δεύτερη μέθοδος, εκμεταλλεύεται την βοήθεια των αποδεκτών (crowdsourcing-oriented fact-checking). Το κοινό ελέγχει την ακρίβεια και την αξιοπιστία του περιεχομένου και έπειτα κατατάσσει το δείγμα ως fake news. Η διαδικασία ολοκληρώνεται με την τελική αξιολόγηση, κατά την οποία χρησιμοποιείται το σύνολο των επιμέρους κριτικών και διεξάγεται ένας καταληκτικός έλεγχος από μια ομάδα ανεξάρτητων συντακτών. Η τρίτη και τελευταία μέθοδος, χρησιμοποιεί κυρίως υπολογιστική ισχύ, είναι αυτοματοποιημένη και κλιμακώσιμη (computational-oriented fact-checking). Υπάρχουν ωστόσο ζητήματα, όπως ποια χωρία του περιεχομένου, αξίζει να αναλυθούν και αν υφίσταται η ύπαρξη ορισμένης προτίμησης (bias) ως προς την αξιοπιστία των ισχυρισμών. Οι πόροι που χρησιμοποιεί η μέθοδος αυτή, βασίζονται σε δεδομένα από το διαδίκτυο αλλά και γράφους γνώσης που προκύπτουν από εργαλεία όπως π.χ. DBpedia.

Οι τρόποι επικύρωσης της εγκυρότητας του περιεχομένου με χρήση του συγγραφικού ύφους (style-based) βασίζονται στην ανάλυση του τόνου του δημιουργού, για να εξακριβωθούν προσπάθειες να χειραγωγηθεί ο αναγνώστης. Στη συνέχεια θα γίνει ανάλυση δύο βασικών μεθόδων. Η πρώτη μέθοδος, αφορά τον εντοπισμό της τάσης για παραπλάνηση στο ύφος του κειμένου. Για να εξακριβωθεί η αξιοπιστία χρησιμοποιούνται μοντέλα επεξεργασίας φυσικής γλώσσας, με σκοπό τον εντοπισμό φράσεων που υποδηλώνουν εξαπάτηση. Η δεύτερη μέθοδος, αφορά τη χρήση τακτικών για την ανίχνευση μοτίβων που δείχνουν έλλειψη αντικειμενικότητας, στο συγγραφικό ύφος του δημιουργού. Όπως έχει αναλυθεί εκτενέστερα και παραπάνω, αναφορικά με τα χαρακτηριστικά των fake news, συνήθως παρατηρείται clickbait τίτλος, κίτρινη δημοσιογραφία, υποκειμενικότητα και υπερβολές που έχουν στόχο να αποπροσανατολίσουν.

Ακριβώς επειδή πρόκειται για ένα διεπιστημονικό θέμα, υπάρχουν πάρα πολλές διαφορετικές προσεγγίσεις και ποικιλία, ως προς την απόδοση των μεθόδων σε κάθε περίπτωση. Κάποιες από τις μεθοδολογίες ταξινόμησης, έχουν μελετηθεί εκτενώς παρουσιάζοντας ποικίλα αποτελέσματα και ταυτόχρονα υπάρχουν κάποιες προσεγγίσεις που δεν έχουν ερευνηθεί πλήρως. Παρακάτω, παρουσιάζεται η συμβολική αναπαράσταση των εργαλείων έρευνας, ανά διαφορετικό προσανατολισμό σχετικά με τα χαρακτηριστικά που εξετάζονται ανά δείγμα. Στην παρούσα εργασία, θα εξεταστεί η προσέγγιση χρήσης μηχανικής μάθησης για τον έλεγχο εγκυρότητας περιεχομένου. Η συγκεκριμένη τεχνική θα αξιοποιεί το περιεχόμενο του εκάστοτε δείγματος αποκλειστικά ως προς το κυρίως κείμενο που παρέχει. Δηλαδή, δεν θα χρησιμοποιηθούν μεταδεδομένα τα οποία μπορεί να παρέχονται από κάποιο σύνολο δεδομένων.

Εικόνα 1 - Εργαλεία για την ανίχνευση fake news στα social media (Shu, Wang, Silva, Tang, & Liu, 2017)

1.7 Περιγραφή του προβλήματος και προτεινόμενη λύση

Έως το συγκεκριμένο σημείο, έχει γίνει αναφορά στον ορισμό των fake news, για να προσδιοριστεί με μεγαλύτερη σαφήνεια το ακριβές νόημα του όρου. Στη συνέχεια, αναλύθηκαν κάποια παραδείγματα με σκοπό να τονιστούν εκδηλώσεις του φαινομένου και οι αντίστοιχες συνέπειες. Πραγματοποιήθηκε αναφορά, στις περιόδους πριν την εμφάνιση των συμβατικών μέσων μαζικής ενημέρωσης, στο απόγειο της δημοτικότητάς τους καθώς και στην μετέπειτα εποχή όπου πλέον η ενημέρωση γίνεται μέσω του διαδικτύου, με έμφαση στα μέσα κοινωνικής δικτύωσης. Έπειτα, έγινε εμβάθυνση στη μορφή του περιεχομένου και των διαφορετικών τύπων δεδομένων που αξιοποιούνται στα fake news συνθέτοντας το τελικό τους προφίλ. Έχουν προσδιοριστεί όλα τα δημοφιλή μέσα διάδοσης και το σκεπτικό από το οποίο προκύπτει η διαπίστωση πως είναι συγκριτικά πιο σημαντικό να εξερευνηθεί η διάδοση στα μέσα κοινωνικής δικτύωσης και το διαδίκτυο, λόγω του μεγέθους του κοινού που επηρεάζεται. Ακόμα, έγινε επεξήγηση των κατηγοριών ταξινόμησης και του είδους του περιεχομένου που συνιστά fake news. Τέλος, έγινε αναφορά στις πιθανές προσεγγίσεις ταυτοποίησης των fake news και τα χαρακτηριστικά τους. Επιπρόσθετα, αναλύθηκαν οι ισχυρότεροι μηχανισμοί ανίχνευσης, βασισμένοι αποκλειστικά στο περιεχόμενο. Το εύλογο ερώτημα που προκύπτει, αφορά την ανάγκη ύπαρξης των προσεγγίσεων που αναφέρθηκαν και εν γένει των διαδικασιών επικύρωσης εγκυρότητας.

Είναι φανερή η ανάγκη της παρούσας μελέτης, αφού ο μέσος άνθρωπος δεν έχει τη γνωστική επάρκεια και γενικότερα την ικανότητα, να διακρίνει τα fake news διαισθητικά χωρίς τη χρήση κάποιας μεθόδου. Ένας μέσος αναγνώστης, μπορεί να ταξινομήσει με επιτυχία περίπου το 4% των κειμένων χρησιμοποιώντας σαν σημείο αναφοράς έναν τυχαίο ταξινομητή (Conroy, Rubin, & Yimin, 2015). Το γεγονός αυτό, σε συνδυασμό με την διαπίστωση πως τα fake news που αναλύουν αρνητικά συμβάντα είναι πολύ πιο αποτελεσματικά και διαδίδονται σε έξι φορές μεγαλύτερο κοινό, είναι ενδεικτικό του προβλήματος (Choraś, και συν., 2019). Οι συνέπειες μπορεί να οδηγήσουν σε συγκρούσεις, σύγχυση και λανθασμένες αποφάσεις στους τομείς των οικονομικών προβλέψεων, του χρηματιστηρίου, της δημόσιας και ατομικής υγείας, της πολιτικής και πολλών ακόμα πεδίων (Demestichas, Remoundou, & Adamopoulou, 2020).

Εκτός όμως από τους απλούς αναγνώστες, ούτε οι μεγάλοι οργανισμοί μαζικής ενημέρωσης μπορούν να ανταπεξέλθουν (Choraś, και συν., 2019). Οι εταιρείες βρίσκονται υπό πίεση λόγω της έντασης του φαινομένου και σαν αποτέλεσμα έχουν ορίσει δικλείδες για την καταπολέμηση του, με τεχνικές βασισμένες σε συστήματα γνώσης. Οι διαδικασίες αυτές, λόγω των συνεχώς αυξανόμενων πηγών πληροφορίας στα μέσα κοινωνικής δικτύωσης, δυσκολεύονται όλο και περισσότερο να προσφέρουν ικανοποιητικά αποτελέσματα. Πρόκειται για ένα διαρκώς μεταβαλλόμενο περιβάλλον, στο οποίο είναι δύσκολη και χρονοβόρα διαδικασία η εξακρίβωση των ειδήσεων. Ταυτόχρονα όμως, είναι αναγκαία η άμεση δημοσίευση των ειδήσεων, ο έλεγχος του υποβάθρου του εκάστοτε συμβάντος και ο έλεγχος της αξιοπιστίας των πηγών. Συνεπώς, ακόμα και οι οργανισμοί που μπορούν να διαθέσουν πόρους για το συγκεκριμένο ζήτημα παρατηρούν πως σταδιακά γίνεται όλο και πιο παρωχημένη η διαδικασία που ακολουθείται. Αξίζει επίσης να αναφερθεί, πως οι επιμέρους οργανισμοί, ακόμα και με τις δικλείδες που χρησιμοποιούνται, είναι πολύ πιθανό εκούσια ή ακούσια να παράγουν ενδεχομένως και κάποιες λανθασμένες αξιολογήσεις.

Μια ιδανική προτεινόμενη λύση, θα αποτελούσε μια αυτοματοποιημένη υπηρεσία η οποία θα είναι προσβάσιμη από όλους ως Software as a Service (SaaS) (Choraś, και συν., 2019). Ωστόσο, για να επιτευχθεί αυτός ο σκοπός χρειάζεται να αντιμετωπιστούν ορισμένα προβλήματα. Μια υπηρεσία σαν αυτή, απαιτεί τη χρήση μεθόδων μηχανικής μάθησης, έτσι ώστε να εφαρμοστούν οι τεχνικές ανάλυσης του συγγραφικού ύφους που αναφέρθηκαν παραπάνω. Αυτή η προσέγγιση, θα μπορέσει να εξετάσει με μεγάλη ακρίβεια τα σημασιολογικά χαρακτηριστικά του κειμένου και να εντοπίσει το φαινόμενο της λεκτικής διαρροής. Το συγκεκριμένο χαρακτηριστικό προδίδει τα fake news και αναφέρεται σε στοιχεία που έχουν να κάνουν με μήκος προτάσεων, λέξεων, παραγράφων, με τη χρήση της στίξης και την αρνητική χροιά συγγραφής. Άρα, δεν χρειάζεται απλά να χρησιμοποιηθούν μοντέλα που θα βασίζονται στη συχνότητα των όρων για την ταξινόμηση, αλλά θα πρέπει να γίνεται αντιληπτό και το πλαίσιο χρήσης των λέξεων. Οπότε, το ουσιαστικότερο πρόβλημα που θα εξεταστεί εντοπίζεται σε δύο βασικά σημεία. Αρχικά, εμφανίζεται έλλειψη αρκετών και ποικιλόμορφων συνόλων δεδομένων για την διαδικασία εκπαίδευσης, που να πληρούν συγκεκριμένες προδιαγραφές ως προς την ποιότητά τους. Επίσης, οι παραδοσιακές μέθοδοι δημιουργίας αναπαραστάσεων κειμένου για την εκπαίδευση των μοντέλων ταξινόμησης υστερούν, συγκριτικά με περισσότερο σύγχρονες τεχνικές.

Τα σύνολα δεδομένων χρειάζεται να περιέχουν μεγάλο αριθμό δειγμάτων, να είναι όσο το δυνατό πιο έμπιστα καθώς και να είναι ισορροπημένα. Στη συγκεκριμένη εργασία, θα χρησιμοποιηθούν διαφορετικά σύνολα δεδομένων που περιέχουν ποικιλία κατηγοριών ως προς τον χαρακτηρισμό των δειγμάτων. Θα εξεταστεί η ανταπόκριση του κάθε συνόλου δεδομένων, χρησιμοποιώντας διάφορα μοντέλα μηχανικής μάθησης, τα οποία βασίζονται σε εμφυτεύματα που αξιοποιούν το πλαίσιο χρήσης των λέξεων (contextual embeddings) του αρχικού κειμένου. Με τη χρήση της συγκεκριμένης προσέγγισης, θα είναι περισσότερο κατανοητό ποια είναι τα χαρακτηριστικά που συνθέτουν ένα ολοκληρωμένο σύνολο δεδομένων, το οποίο

μπορεί να αξιοποιηθεί κατάλληλα για τη δημιουργία υπηρεσιών επικύρωσης εγκυρότητας περιεχομένου.

Μια διαπίστωση η οποία αξίζει να αναφερθεί, αφορά την περίοδο κατά την οποία θα είναι χρήσιμες τέτοιου είδους υπηρεσίες. Τα fake news ως προς το περιεχόμενο και τον τρόπο συγγραφής τους εξελίσσονται με την πάροδο του χρόνου (Choraś, και συν., 2019). Το γεγονός αυτό, σημαίνει πως χρειάζεται η υιοθέτηση μηχανισμών που θα επιτρέπουν την επαναληπτική εκπαίδευση των μοντέλων σε τακτά χρονικά διαστήματα πάνω σε σύγχρονα δεδομένα (lifelong learning). Με αυτόν τον τρόπο, επιδιώκεται η συσσώρευση πληροφορίας για να υπάρξει όσο το δυνατό πιο αποτελεσματική διαδικασία εκπαίδευσης με την βέλτιστη δυνατή απόδοση.

Στα επόμενα τμήματα της εργασίας, θα πραγματοποιηθεί σύντομη παρουσίαση των μεθοδολογιών κωδικοποίησης κειμένου σε μορφή κατάλληλη για χρήση από μοντέλα μάθησης. Θα δοθεί έμφαση στο μοντέλο BERT και ιδιαίτερα στο DistilBERT. Έπειτα, θα επεξηγηθούν τα ξεχωριστά χαρακτηριστικά κάθε επιμέρους συνόλου δεδομένων που αξιοποιήθηκε καθώς και η αναλυτική προσέγγιση που χρησιμοποιήθηκε για την διάκριση των δειγμάτων ως προς την αξιοπιστία τους. Στη συνέχεια, θα γίνει παρουσίαση αλλά και σχολιασμός των αποτελεσμάτων και τελικά αναφορά σε μελλοντικές επεκτάσεις της συγκεκριμένης εργασίας.

2 Μέθοδοι κωδικοποίησης κειμένου

2.1 Προγενέστερες διαδικασίες κωδικοποίησης κειμένου

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, στην παρούσα εργασία θα χρησιμοποιηθούν μηχανισμοί κωδικοποίησης του κειμένου, που βασίζονται στο πλαίσιο στο οποίο χρησιμοποιείται ένας όρος και όχι αποκλειστικά στη συχνότητα εμφάνισής του. Ωστόσο, θα πρέπει να γίνει κατανοητό ποιες είναι οι υπόλοιπες προσεγγίσεις που μπορούν να χρησιμοποιηθούν, ως προς την κωδικοποίηση κειμένου για σκοπούς επεξεργασίας φυσικής γλώσσας (NLP: Natural Language Processing). Έτσι, θα γίνει προφανές, γιατί τα πλεονεκτήματα των μεθόδων που αξιοποιούν μηχανισμούς προσοχής, όπως το μοντέλο Bidirectional Encoder Representations from Transformers (BERT) υπερτερούν. Ακόμα, θα γίνει προσπάθεια επεξήγησης των μειονεκτημάτων του BERT και σύγκρισης με τα πλεονεκτήματα χρήσης της συμπυκνωμένης μορφής του, με ονομασία DistilBERT.

Στο συγκεκριμένο κεφάλαιο, θα γίνει επίσης αναφορά σε μεθόδους αναπαράστασης κειμένου όπως Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec και Glove. Οι συμβατικές μέθοδοι, αρχικά κωδικοποιούσαν το κείμενο σε διανύσματα όπου κάθε στοιχείο αντιστοιχούσε σε μια λέξη ανάλογα με τη σειρά που εμφανιζόταν (για παράδειγμα, η πρώτη λέξη της πρότασης κωδικοποιείται στο πρώτο στοιχείο του διανύσματος και η δεύτερη λέξη στο δεύτερο στοιχείο) (Mikolov, Chen, Corrado, & Dean, 2013). Η συγκεκριμένη πρακτική αφορούσε τη δημιουργία στατικών αναπαραστάσεων, με χρήση στατιστικών μεθόδων. Στη συνέχεια, η προσέγγιση μεταβλήθηκε και πλέον κάθε διάνυσμα αντιστοιχούσε σε μια λέξη και η κωδικοποίηση υποδείκνυε ομοιότητα μεταξύ όρων. Αυτό ήταν και το πρώτο βήμα, για να γίνει πολύ πιο εύκολη η υλοποίηση εργασιών επεξεργασίας φυσικής γλώσσας (NLP), όπως ταξινόμηση κειμένου, ανάλυση συναισθήματος, αναγνώριση ονομάτων ανθρώπων (named entity recognition) και μετάφραση κειμένων.

2.1.1 Επεξεργασία κειμένου πριν την κωδικοποίηση

Οι συμβατικές μέθοδοι συνήθως αξιοποιούν κάποιες συγκεκριμένες τεχνικές πρότερης επεξεργασίας του κειμένου, έτσι ώστε να δημιουργηθούν όσο το δυνατό καλύτερες αναπαραστάσεις. Στη συνέχεια, θα παρουσιαστούν κάποιες από τις δημοφιλέστερες τεχνικές αυτού του είδους καθώς και η επίδρασή τους στη μορφή του περιεχομένου προς κωδικοποίηση. Κάθε λέξη του κειμένου, χρειάζεται να αποτελεί ένα ξεχωριστό στοιχείο που χωρίζεται με κενό από την προηγούμενη και την επόμενη της και αντιμετωπίζεται ξεχωριστά σαν μονάδα (tokenization) κατά τη διαδικασία κωδικοποίησης. Ακόμα, μια διαδεδομένη πρακτική είναι όλα τα κεφαλαία γράμματα να μετατρέπονται σε πεζά (lowercasing) ανεξάρτητα από το αν βρίσκονται στην αρχή της λέξης. Η προσέγγιση αυτή έχει παρατηρηθεί, πως σε ορισμένες περιστάσεις δημιουργεί διφορούμενα συμπεράσματα και τελικά μπορεί το όνομα μιας οντότητας

(π.χ. Apple) να ταυτίζεται με μια απλή λέξη (π.χ. apple) (Camacho - Collados & Pilehvar, 2018). Μια άλλη μέθοδος, που εμφανίζεται συχνά είναι η αντικατάσταση των όρων με την αρχική λέξη από την οποία προέρχονται (π.χ. αντικατάσταση της λέξης ‘’are’’ από τη λέξη ‘’be’’) με σκοπό να γίνει αποδοτικότερη αποτύπωση της συχνότητας εμφάνισης μιας έννοιας (lemmatization) (Camacho - Collados & Pilehvar, 2018). Το σημαντικό μειονέκτημα της μεθόδου είναι πως με τις αλλαγές που δημιουργούνται, συχνά αγνοούνται οι λεπτομέρειες, που αφορούν τη σύνταξη του κειμένου. Μια τελευταία μέθοδος που είναι σημαντικό να αναφερθεί είναι η ένωση λέξεων που εμφανίζονται συχνά μαζί και αποτελούν ολοκληρωμένες έννοιες (multiword grouping). Το πλεονέκτημα είναι πως οι όροι αντιμετωπίζονται κατά την κωδικοποίηση σαν μια λέξη, γεγονός το οποίο μπορεί να φανεί χρήσιμο ως προς τη σαφέστερη απόδοση του νοήματος του κειμένου (Camacho - Collados & Pilehvar, 2018).

2.1.2 Term Frequency – Inverse Document Frequency (TF-IDF)

Η κωδικοποίηση διάφορων εγγράφων προς ταξινόμηση, γίνεται εύκολα μέσω της χρήσης του Vector Space Model (VSM). Κάθε ξεχωριστό έγγραφο d αντιπροσωπεύεται από ένα διάνυσμα V με βάρη για κάθε όρο που περιέχει και περιγράφεται ως $V_d = (v_{d1}, v_{d2}, \ldots, v_{dn})$ (Deng, 2004). Ένας βασικός μηχανισμός ο οποίος έχει χρησιμοποιηθεί σε μεγάλο βαθμό στο παρελθόν είναι τα διανύσματα TF-IDF, που θυμίζουν κατά κάποιον τρόπο την τεχνική VSM. Πρώτη φορά η ιδέα χρήσης TF-IDF τέθηκε πιο ξεκάθαρα στο έργο Probabilistic Models in Information Retrieval του μελετητή Norbert Fuhr, στη βάση της αναζήτησης μια προσέγγισης αποδοτικής ανάκτησης πληροφορίας (Fuhr, 1992). Τα στοιχεία του διανύσματος TF-IDF είναι στην πραγματικότητα το γινόμενο δύο μετρικών.

Το πρώτο τμήμα της ονομασίας (TF) είναι συντομογραφία των όρων term frequency. Η συγκεκριμένη μετρική αναφέρεται στην απόδοση βαρών στους όρους, αναφορικά με το κατά πόσο σημαντικοί είναι για το περιεχόμενο του εκάστοτε εγγράφου ανάλογα με τη συχνότητα εμφάνισής τους. Το δεύτερο τμήμα της ονομασίας (IDF) είναι συντομογραφία των όρων inverse document frequency και η μετρική αυτή αποτυπώνει την ικανότητα του κάθε όρου να διαχωρίζει όμοια έγγραφα. Η λογική πίσω από την χρήση της δεύτερης μετρικής βασίζεται στην υπόθεση πως όροι οι οποίοι εμφανίζονται συχνά σε πολλά κείμενα έχουν μικρότερη επιρροή ως προς τον διαχωρισμό παρόμοιων εγγράφων. Η συγκεκριμένη μέθοδος είναι ιδιαίτερα απλή και μπορεί να εφαρμοστεί αποδοτικά σε μεγάλα σύνολα κειμένων. Το μειονέκτημα είναι πως κατά τη χρήση της μετρικής IDF, δεν αποτυπώνεται το γεγονός πως σε διαφορετικές κατηγορίες ταξινόμησης οι ίδιες λέξεις διαδραματίζουν διαφορετικό ρόλο. Οπότε, στη συνέχεια έγιναν δοκιμές ώστε να βρεθούν εναλλακτικές ως προς τη χρήση της δεύτερης μετρικής. Το αποτέλεσμα ήταν η δημιουργία των TF-Category Relevance Factor (CRF), TF-OddsRatio και TF-CHI (όπου η δεύτερη μετρική έχει πάρει την ονομασία της από την κατανομή χ^2) (Deng, 2004). Οι δύο αποτελεσματικότερες προσεγγίσεις είναι οι TF-OddsRatio και TF-CHI (Deng, 2004).

Στη συνέχεια, θα γίνει αναφορά στις διαφορές που παρουσιάζονται σε κάθε περίπτωση σχετικά με τη δεύτερη μετρική. Στην περίπτωση TF-CRF (Category Relevance Factor), η δεύτερη μετρική αντιπροσωπεύει τη βαρύτητα ενός δείγματος για την ταξινόμηση ενός εγγράφου ανάμεσα σε διαφορετικές κατηγορίες (Deng, 2004). Η προσέγγιση TF-OddsRatio, χρησιμοποιείται κυρίως στο πεδίο ανάκτησης πληροφοριών όταν υπάρχει η ανάγκη να ταξινομηθούν τα έγγραφα ανάλογα με το κατά πόσο συναφή είναι με την θετική κατηγορία, λαμβάνοντας την εμφάνιση διαφορετικών λέξεων ως χαρακτηριστικό ομοιότητας (Deng, 2004). Τέλος, η πιο αποτελεσματική τακτική ανάμεσα σε όσες έχουν αναφερθεί είναι η TF-CHI που η δεύτερη μετρική αφορά την ποσοτικοποίηση της απουσίας ανεξαρτησίας μεταξύ μιας λέξης και μιας κατηγορίας. Η μετρική υπολογίζεται με τρόπο εμπνευσμένο από την κατανομή χ^2 (Deng, 2004).

Η μέθοδος TF-IDF είναι ο πιο αποδοτικός τρόπος για κωδικοποίηση μεγάλων συνόλων δεδομένων, με την μικρότερη δυνατή απώλεια ακρίβειας, συγκριτικά με τις προηγούμενες μεθόδους. Ωστόσο, όλες οι παραπάνω προσεγγίσεις προσφέρουν διαφορετικούς τρόπους να κωδικοποιηθεί ένα έγγραφο σε διανύσματα με βάρη (όπου αναπαρίστανται λέξεις) τα οποία μπορούν να αξιοποιηθούν από μοντέλα μάθησης. Έτσι, ολόκληρα σύνολα δεδομένων που περιέχουν κείμενο μπορούν να μελετηθούν εύκολα και εν συνεχεία να ταξινομηθούν τα επιμέρους δείγματα σε κατηγορίες. Τα διανύσματα όμως που δημιουργούνται σε κάθε περίπτωση, βασίζονται στη συχνότητα εμφάνισης λέξεων και τη συνάφεια των όρων με τις κατηγορίες ταξινόμησης.

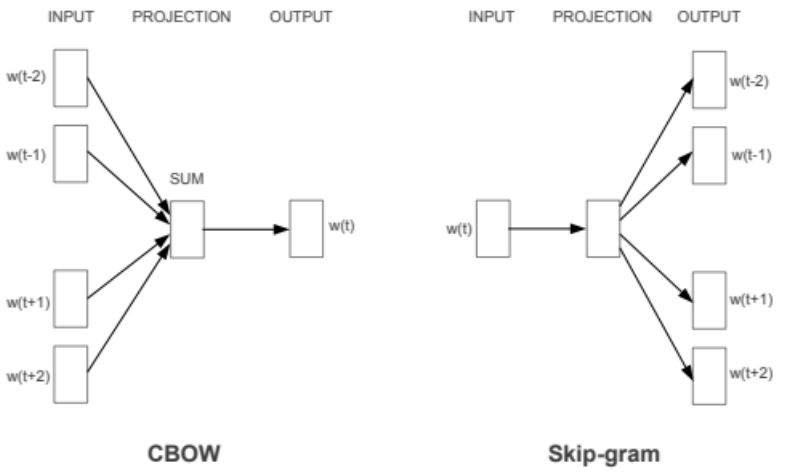
2.1.3 Word2Vec

Όπως αναφέρθηκε, τόσο στην αρχή της ενότητας όσο και στην περιγραφή της μεθόδου TF-IDF, η κωδικοποίηση του κειμένου αντιμετώπιζε τις λέξεις ως μονάδες χωρίς να χρησιμοποιείται η ομοιότητά τους ως προς τη σημασιολογική τους βαρύτητα. Η ανάγκη υιοθέτησης νέων πρακτικών, προέκυψε συνδυαστικά με τη συνειδητοποίηση πως ανεξάρτητα με τον πιθανότατα μεγάλο όγκο των διαθέσιμων δεδομένων δεν παρουσιάζονταν βελτιωμένα αποτελέσματα (Mikolov, Chen, Corrado, & Dean, 2013). Έτσι, υιοθετήθηκαν κατανεμημένες αναπαραστάσεις λέξεων, δηλαδή, πυκνά διανύσματα (dense vectors) όπου κάθε διάσταση κωδικοποιεί ένα διαφορετικό χαρακτηριστικό της λέξης με βάση τον τρόπο που χρησιμοποιείται.

Οι συγκεκριμένες αναπαραστάσεις ή αλλιώς εμφυτεύματα (embeddings) στοχεύουν στην κωδικοποίηση σημασιολογικής πληροφορίας. Προκύπτουν ως αποτέλεσμα μη επιβλεπόμενης μάθησης όπου γίνεται επεξεργασία μεγάλων όγκων δεδομένων για να βρεθούν επαναλαμβανόμενα μοτίβα στα κείμενα. Η υπόθεση με βάση την οποία χρησιμοποιείται η μέθοδος αυτή, είναι πως λέξεις με παρόμοιες έννοιες εμφανίζονται σε παρόμοιο πλαίσιο.

Δύο από τα πιο αποτελεσματικά και πρώιμα μοντέλα της μεθόδου ήταν το NNLM (Feedforward Neural Net Language Model) και το RNNLM (Recurrent Neural Net Language Model) και χρησιμοποιήθηκαν για την ανίχνευση εξαρτήσεων μεταξύ των όρων ενός κειμένου. Ωστόσο, έγιναν περαιτέρω προσπάθειες να μειωθεί

όσο το δυνατόν περισσότερο η υπολογιστική πολυπλοκότητα ανάλογα με τον αριθμό των παραμέτρων που χρειάζεται να εκπαιδευτούν. Παράλληλα υπήρξε η επιδίωξη να αυξηθεί η ακρίβεια. Έτσι προτάθηκαν τα μοντέλα Continuous Bag-of-Words Model (CBOW) και Continuous Skip-gram Model (Mikolov, Chen, Corrado, & Dean, 2013). Το μοντέλο CBOW χρησιμοποιεί ίσο αριθμό όρων, από λέξεις που έπονται όσο και από λέξεις που ήδη έχουν εμφανιστεί (2 πριν και μετά), με σκοπό να προσδιοριστεί ο όρος που βρίσκεται ενδιάμεσα. Οπότε, δέχεται πολλαπλές εισόδους και παράγει μια έξοδο. Το δεύτερο μοντέλο, αντίθετα με το προηγούμενο δέχεται ως είσοδο μια λέξη και την αξιοποιεί ώστε να προσδιοριστεί η κωδικοποίηση εκείνων που προηγούνται αλλά και αυτών που ακολουθούν στη συνέχεια. Η λογική πίσω από αυτή την προσέγγιση κωδικοποίησης, βασίζεται στο γεγονός πως μια λέξη μπορεί να αναπαρασταθεί ανάλογα με τις λέξεις που βρίσκονται κοντά της σε μια πρόταση, με το σκεπτικό πως οι όροι που βρίσκονται πιο μακριά δεν σχετίζονται σε τόσο μεγάλο βαθμό μαζί της.



Εικόνα 2- Παρουσίαση των αρχιτεκτονικών CBOW και Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013)

Αργότερα έγινε η παρατήρηση πως η εκπαίδευση των μοντέλων επεξεργασίας φυσικής γλώσσας μπορεί να γίνει επιτυχώς σε δύο βήματα. Αρχικά, τα μοντέλα που αναλύθηκαν παραπάνω μπορούν να δημιουργούν τα εμφυτεύματα και στη συνέχεια αρχιτεκτονικές όπως η NNLM εκπαιδεύονται με βάση τα εμφυτεύματα αυτά για να επιτευχθεί ο εκάστοτε τελικός στόχος. Οι αρχιτεκτονικές των νευρωνικών δικτύων CBOW και Skip-gram δεν είναι ιδιαίτερα μεγάλες ως προς τον αριθμό των στρωμάτων τους (shallow architectures) και για αυτό τον λόγο η χρήση τους είναι αρκετά αποδοτική καθώς μπορούν να διαχειριστούν σχετικά εύκολα μεγάλους όγκους δεδομένων.

Με σκοπό να γίνει πιο κατανοητό το αποτέλεσμα, θα παρουσιαστεί ένα παράδειγμα. Για να βρεθεί μια λέξη παρόμοια με τη λέξη small με τον ίδιο τρόπο με τον οποίο οι όροι bigger και big σχετίζονται μεταξύ τους υπολογίζεται το ακόλουθο διάνυσμα.

$$X = vector(\text{bigger}) - vector(\text{big}) + vector("small")$$

Στη συνέχεια, με χρήση της μετρικής cosine similarity εντοπίζεται το πιο σχετικό διάνυσμα ανάμεσα σε όσα έχουν ήδη δημιουργηθεί. Όταν υπάρχει ένα αρκετά μεγάλο σύνολο δεδομένων τότε μπορούν να εντοπιστούν πολύ λεπτές σημασιολογικές συσχετίσεις μεταξύ των λέξεων. Για παράδειγμα, η λέξη Paris συσχετίζεται με τη λέξη France, όπως η λέξη Berlin συσχετίζεται με τη λέξη Germany.

Ωστόσο, δεν υπάρχουν μόνο πλεονεκτήματα αλλά και σοβαρά μειονεκτήματα. Το βασικό μειονέκτημα της μεθόδου είναι πως δεν μπορεί να διαχειριστεί την πολυσημία κάποιων όρων, αφού όλες οι πληροφορίες της λέξης συγχέονται σε ένα διάνυσμα το οποίο περιέχει αριθμητικές τιμές για κάθε χαρακτηριστικό που έχει εξαχθεί (Mikolov, Suttskever, Chen, Corrado, & Dean Jeff, 2013). Ακόμα, ο κάθε όρος κωδικοποιείται ανεξάρτητα από το ευρύτερο πλαίσιο στο οποίο εμφανίζεται και δεν εξαρτάται η κωδικοποίηση από τις αλλαγές στον τρόπο χρήσης του στο συνολικό κείμενο. Επιπροσθέτως, δεν παρέχει τόσο καλή απόδοση σε περιπτώσεις λέξεων που δεν ανήκουν στο λεξιλόγιο των δεδομένων εκπαίδευσης και ακόμα λέξεις που εμφανίζονται σπάνια δεν κωδικοποιούνται με ακρίβεια συγκριτικά με τις υπόλοιπες. Η λογική της μεθόδου Word2Vec επεκτάθηκε και σε ολόκληρες φράσεις εκτός από μεμονωμένες λέξεις με την ονομασία Doc2Vec (Mikolov, Suttskever, Chen, Corrado, & Dean Jeff, 2013). Επίσης, συνήθως προτιμάται να γίνεται χρήση προεκπαιδευμένων εμφυτευμάτων για να μειώνεται όσο το δυνατό περισσότερο το υπολογιστικό κόστος εκπαίδευσης.

2.1.4 GloVe

Στην παρούσα ενότητα, θα γίνει σύντομη επεξήγηση της κωδικοποίησης κειμένου σε μορφή εμφυτευμάτων GloVe. Τα εμφυτεύματα που δημιουργούνται με τη συγκεκριμένη μέθοδο προκύπτουν χρησιμοποιώντας βασικές αρχές μηχανικής μάθησης. Όμως, η μέθοδος που ακολουθείται είναι διαφορετική από την προηγούμενη, αφού η κεντρική ιδέα δεν αφορά την κωδικοποίηση με βάση ένα μικρό σύνολο προγενέστερων και μεταγενέστερων λέξεων όπως στην περίπτωση χρήσης Word2Vec. Η λογική της προσέγγισης, βασίζεται στο γεγονός πως οι λέξεις κωδικοποιούνται ανάλογα με την πιθανότητα να εμφανίζονται συνδυαστικά με άλλους όρους στα επιμέρους δείγματα του συνόλου δεδομένων (Pennington, Socher, & Manning, 2014).

Η διαδικασία για να δημιουργηθούν τα εμφυτεύματα είναι αρκετά διαφορετική (Pennington, Socher, & Manning, 2014). Αρχικά, δημιουργείται ένας πίνακας με όλα τα πιθανά ζεύγη λέξεων στο σύνολο δεδομένων και υπολογίζονται οι πιθανότητες να εμφανίζονται οι όροι συνδυαστικά στα δείγματα. Στη συνέχεια, ακολουθεί παραγοντοποίηση του πίνακα και προκύπτουν δύο πίνακες (word vector matrix και context vector matrix) που αποτελούν μια πρώιμη μορφή των τελικών εμφυτευμάτων. Στο στάδιο αυτό, γίνεται συνδυασμός των πινάκων που έχουν προκύψει σε μια ενιαία δομή. Έπειτα, με τη μέθοδο της παλινδρόμησης ελάχιστων τετραγώνων εφαρμοζόμενη ξεχωριστά για όλα τα πιθανά ζεύγη λέξεων του τελικού πίνακα, δημιουργούνται τα εμφυτεύματα. Το κριτήριο τερματισμού της διαδικασίας παλινδρόμησης, αποτελεί η ελαχιστοποίηση της διαφοράς του εσωτερικού γινομένου

των διανυσμάτων (γραμμών πίνακα), με τον λογάριθμο της πιθανότητας να συνυπάρχουν οι δύο λέξεις στο κείμενο.

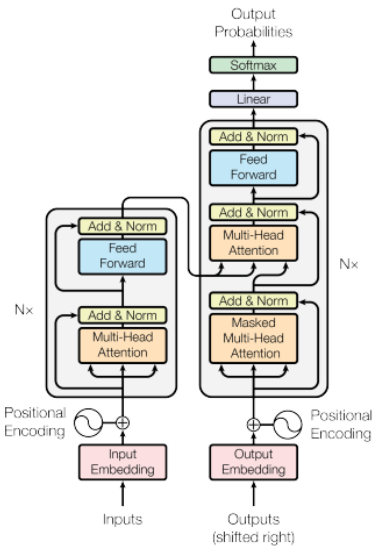
Αν και πρόκειται για μια διαδικασία φαινομενικά απλή και κλιμακώσιμη, υπάρχουν κάποια μειονεκτήματα. Το βασικότερο μειονέκτημα είναι πως τα σύνολα δεδομένων μπορεί να περιέχουν εκατοντάδες ή χιλιάδες ξεχωριστές λέξεις ακόμα και αν προηγηθεί επεξεργασία στο κείμενο (π.χ. lemmatization). Το γεγονός αυτό, σημαίνει πως για να υπολογιστούν οι πιθανότητες για όλα τα ζεύγη και στη συνέχεια να ολοκληρωθεί η υπόλοιπη διαδικασία θα χρειαστούν σημαντικοί υπολογιστικοί πόροι, αφού οι συνολικοί συνδυασμοί είναι πάρα πολλοί (Pennington, Socher, & Manning, 2014). Σαν αποτέλεσμα, είναι πολλές φορές θεμιτό να χρησιμοποιείται η συγκεκριμένη διαδικασία, ενώ έχει προηγηθεί κάποια εκπαίδευση σε μεγάλα σύνολα δεδομένων. Επιπλέον, το γεγονός πως η προσέγγιση μελετά τις λέξεις, με βάση ολόκληρο το σύνολο δεδομένων, χωρίς να δίνει βαρύτητα στις επιμέρους περιπτώσεις χρήσης τους, είναι πιθανό να οδηγήσει σε περιορισμένη ικανότητα να αποτυπωθεί η πολυσημία των λέξεων και οι συσχετισμοί τους με το συγγραφικό ύφος του υπόλοιπου κειμένου.

2.2 Μέθοδοι κωδικοποίησης βασισμένες στο πλαίσιο χρήσης των λέξεων

Στην προηγούμενη ενότητα, έγινε ανάλυση των πιο δημοφιλών μεθόδων κωδικοποίησης όμως παρουσιάστηκαν και σοβαρά μειονεκτήματα αυτών. Το σημαντικότερο, είναι η αδυναμία σύλληψης πιο περίπλοκων χαρακτηριστικών χρήσης των λέξεων συντακτικά και σημασιολογικά. Παράλληλα, δεν υπάρχει σε καμία από τις προηγούμενες μεθόδους η δυνατότητα να αποτυπωθεί στα εμφυτεύματα η πολυσημία μιας λέξης ανάλογα με το ύφος του συγγραφέα. Οπότε, χρειάζονται μέθοδοι οι οποίες ανταποκρίνονται και στις δύο απαιτήσεις. Οι προτεινόμενες προσεγγίσεις πρέπει να είναι αποδοτικές και να μπορούν να συνδυαστούν όσο το δυνατό πιο εύκολα με τα μοντέλα που χρησιμοποιούνται ήδη για σκοπούς επεξεργασίας φυσικής γλώσσας (Peters, et al., 2018).

2.2.1 Transformers

Στη συνέχεια της εργασίας, παρατηρείται συχνή αναφορά στην έννοια των μετασχηματιστών (transformers). Είναι αρκετά σημαντικό να εξηγηθεί η βασική δομή, ο σκοπός τους και τα ιδιαίτερα πλεονεκτήματα που προσφέρουν. Οι μετασχηματιστές έχουν σχετικά απλή οργάνωση και βασίζονται σχεδόν αποκλειστικά σε μηχανισμούς προσοχής εξαλείφοντας την ανάγκη χρήσης παλινδρόμησης και συνέλιξης. Οι μηχανισμοί προσοχής εντοπίζουν εξαρτήσεις λέξεων μεταξύ των όρων ενός κειμένου αποδοτικά και με ακρίβεια. Χρησιμοποιούνται για την επεξεργασία ακολουθιών λέξεων (προτάσεων) και την επίτευξη σκοπών επεξεργασίας φυσικής γλώσσας. Οι μετασχηματιστές περιέχουν δομές που επιτρέπουν τόσο την κωδικοποίηση όσο και την αποκωδικοποίηση του κειμένου (Vaswani, και συν., 2017).



Εικόνα 3 - Βασική αρχιτεκτονική κωδικοποίησης - αποκωδικοποίησης των transformers (Vaswani, et al., 2017)

Ο βασικός μηχανισμός προσοχής ονομάζεται self-attention και πρόκειται για έναν όρο για τον οποίο δεν υπάρχει ακριβής μετάφραση. Ο μηχανισμός αυτός, επιτρέπει σε κάθε λέξη να συσχετιστεί με τις υπόλοιπες σε ένα συγκεκριμένο εύρος και να εξεταστεί σε τι βαθμό προσδιορίζουν τη σημασία της. Στη συνέχεια, οι εξαρτήσεις μεταξύ όλων των συνδυασμών λέξεων ποσοτικοποιούνται με τη μορφή βαρών που δείχνουν πόσο μεγάλη σημασία πρέπει να δοθεί σε κάθε λέξη κατά την επεξεργασία του κειμένου. Με βάση το μηχανισμό αυτό, δημιουργήθηκε η έννοια της δομής προσοχής πολλαπλών κεφαλών (Multi-Head Attention), που κάθε κεφαλή αποτελείται από ένα σύνολο self-attention δομών. Το αποτέλεσμα είναι οι διαφορετικές κεφαλές να εξετάζουν την είσοδο και να προσφέρουν διάφορες προσεγγίσεις ως προς τη σημασία των λέξεων (Vaswani, και συν., 2017). Έτσι, η τελική εκτίμηση που προκύπτει από τη συνένωση των επιμέρους βαρών είναι πιο ακριβής, αφού είναι αποτέλεσμα συνδυασμού των διαφορετικών εκτιμήσεων.

Η αρχιτεκτονική των μετασχηματιστών, αποτελείται από επιμέρους στρώματα τα οποία περιέχουν μια δομή μηχανισμού προσοχής πολλαπλών κεφαλών που ακολουθείται από ένα δίκτυο εμπρόσθιας τροφοδότησης το οποίο με τη σειρά του προωθεί το αποτέλεσμα στο επόμενο στρώμα (Vaswani, και συν., 2017). Αυτές οι δύο δομικές μονάδες ανά στρώμα προσφέρουν σημαντικά πλεονεκτήματα. Η μονάδα του μηχανισμού προσοχής πολλαπλών κεφαλών προσφέρει μια γενικευμένη εκτίμηση της σημασίας των επιμέρους λέξεων σχετικά με την κωδικοποίηση, ενώ τα δίκτυα εμπρόσθιας τροφοδότησης προσφέρουν τη βελτιστοποίηση των αποτελεσμάτων και την εξαγωγή τοπικών χαρακτηριστικών για τις επιμέρους λέξεις.

2.2.2 Embeddings from Language Models (ELMo)

Η μέθοδος δημιουργίας εμφυτευμάτων ELMo (Embeddings from Language Models) αποτυπώνει ένα κείμενο κωδικοποιώντας τις λέξεις ως προς τον τρόπο που χρησιμοποιούνται καθώς και ως προς το νόημά τους. Ωστόσο, δεν αξιοποιεί

μηχανισμούς προσοχής για να προσδιορίσει τη σημασία των λέξεων, όπως συμβαίνει σε περιπτώσεις που θα παρουσιαστούν στη συνέχεια. Η συγκεκριμένη προσέγγιση, βασίζεται στη δημιουργία εμφυτευμάτων μέσω μοντέλων βαθιάς μάθησης. Ειδικότερα, η τελική μορφή της μεθόδου προέκυψε από την παρατήρηση πως γλωσσικά μοντέλα τα οποία προβλέπουν την επόμενη λέξη και βασίζονται σε προηγούμενους και επόμενους όρους ταυτόχρονα (biLM-bidirectional Language Models), οδηγούν σε αναπαραστάσεις λέξεων που αποτυπώνουν πολλαπλές ερμηνείες της σημασίας τους. Για παράδειγμα, παρατηρήθηκε πως τα ανώτερα στρώματα μακράς βραχύχρονης μνήμης (Long Short Term Memory) δύο κατευθύνσεων σε μοντέλα biLM συλλαμβάνουν ικανοποιητικά τη σημασιολογία των λέξεων (Peters, και συν., 2018). Ταυτόχρονα, τα κατώτερα στρώματα είναι πιο ακριβή στην ανάθεση ετικετών στο κείμενο που χρησιμεύουν στην κατηγοριοποίηση διαφορετικών ιδιοτήτων των λέξεων (POS tags), όπως για παράδειγμα χαρακτηρισμό προσωπικών αντωνυμιών και επιρρημάτων (Peters, και συν., 2018).

Στη συνέχεια, θα πραγματοποιηθεί σύντομη ανάλυση της διαδικασίας έως και την τελική δημιουργία των εμφυτευμάτων χρησιμοποιώντας μοντέλα μάθησης τα οποία βασίζονται σε στρώματα μακράς βραχύχρονης μνήμης δύο κατευθύνσεων (bidirectional LSTM - biLSTM) (Peters, και συν., 2018). Στο πρώτο στάδιο, πραγματοποιείται μη επιβλεπόμενη εκπαίδευση του μοντέλου γλώσσας σε μεγάλα σύνολα δεδομένων με σκοπό την πρόβλεψη της επόμενης λέξης. Στην είσοδο του μοντέλου παρέχεται το κείμενο και ακολουθείται επεξεργασία ως προς δύο κατευθύνσεις από τις δομές του. Για διαφορετικούς σκοπούς επεξεργασίας φυσικής γλώσσας είναι αναγκαίο να γίνει προσδιορισμός της σημασίας κάθε στρώματος biLSTM ως προς την τελική συμμετοχή του στα εμφυτεύματα. Οπότε, πραγματοποιείται μια δεύτερη διαδικασία επιβλεπόμενης μάθησης στο εκάστοτε σύνολο δεδομένων για να υπάρξει προσαρμογή των βαρών σε κάθε διαφορετική εργασία επεξεργασίας φυσικής γλώσσας (NLP: Natural Language Processing). Ο σκοπός είναι να δημιουργηθεί μια όσο το δυνατό καλύτερη αναπαράσταση των δειγμάτων (Peters, και συν., 2018). Κάθε κωδικοποίηση λέξης, αποδίδεται ως τα αποτελέσματα που προκύπτουν στις επιμέρους δομές μακράς βραχύχρονης μνήμης δύο κατευθύνσεων ανά στρώμα. Τελικά, κάθε δείγμα κειμένου αναπαρίσταται από τα εμφυτεύματα των επιμέρους λέξεων που το συνθέτουν. Η διαφορά συγκριτικά με τα προηγούμενα μέσα κωδικοποίησης κειμένου είναι πως τα εμφυτεύματα των λέξεων δεν είναι στατικά αλλά μεταβάλλονται ανάλογα με την σημασία του όρου στο κάθε δείγμα.

Το βασικό πλεονέκτημα της μεθόδου είναι πως αποδίδει μια λεπτομερή αναπαράσταση της εκάστοτε λέξης η οποία ανταποκρίνεται στο νόημά της, όπως προκύπτει από το ύφος του συγγραφέα. Επομένως, προσφέρει μια διαφορετική χρησιμότητα συγκριτικά με προγενέστερες μεθόδους που αναφέρθηκαν παραπάνω και θεμελιώνουν τη λειτουργία τους σε άλλες θεωρητικές βάσεις. Ωστόσο, υπάρχουν ορισμένα μειονεκτήματα με χαρακτηριστικό παράδειγμα τη χαμηλή απόδοση σε λέξεις που δεν εμφανίζονται στο στάδιο της αρχικής εκπαίδευσης και δεν είναι γνωστό πως μπορούν να αναπαρασταθούν. Επιπλέον, οι διαδικασίες μάθησης που αναφέρθηκαν έχουν αυξημένες απαιτήσεις σε υπολογιστικούς πόρους. Γενικότερα, οι διαδικασίες κωδικοποίησης με βάση το πλαίσιο εμφάνισης μιας λέξης κρύβουν

αδυναμίες όπως θα αναλυθεί παρακάτω και θα γίνει αντιληπτό από ορισμένα αποτελέσματα της πειραματικής διαδικασίας.

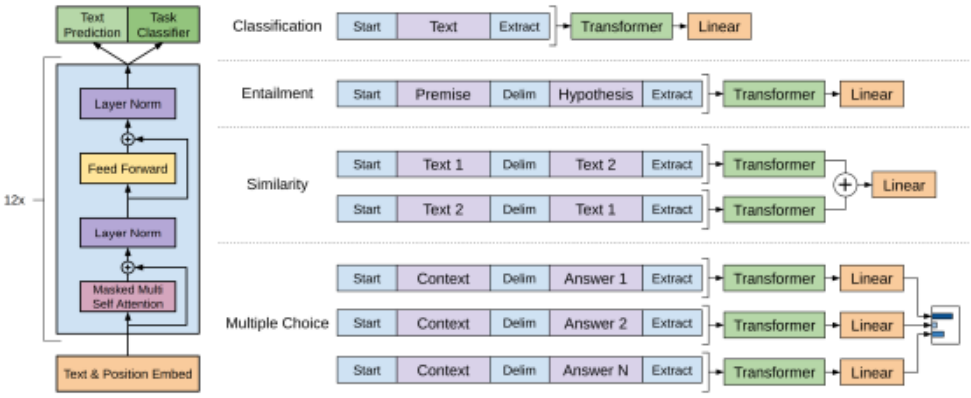
2.2.3 Generative Pretrained Transformer (GPT)

Στη συνέχεια της εργασίας, θα γίνει προσπάθεια σύντομης ανάλυσης κάποιων δημοφιλών διαδικασιών κωδικοποίησης κειμένου, με χρήση μηχανισμών προσοχής. Αρχικά, θα γίνει αναφορά στην κατηγορία μοντέλων παραγωγικού προεκπαιδευμένου μετασχηματιστή (GPT: Generative Pretrained Transformer) ως μέσο για τη δημιουργία εμφυτευμάτων. Στις περισσότερες από τις προσεγγίσεις που έχουν παρουσιαστεί έως τώρα η κωδικοποίηση προκύπτει ως αποτέλεσμα μη επιβλεπόμενης μάθησης. Σε προηγούμενες περιπτώσεις, ακολουθούσε βελτιστοποίηση των εμφυτευμάτων με προσαρμογή των βαρών ανάλογα με το σκοπό της εκάστοτε εργασίας επεξεργασίας φυσικής γλώσσας που μελετάται (π.χ. classification, similarity). Η συνολική προσέγγιση που θα αναλυθεί παρακάτω δεν διαφέρει ιδιαίτερα συγκριτικά με την τεχνική ELMo. Μια βασική διαφορά εντοπίζεται στην αξιοποίηση δομών μετασχηματιστών (transformers).

Στο πρώτο στάδιο, πραγματοποιείται η διαδικασία tokenization η οποία έχει αναφερθεί και σε προηγούμενη ενότητα. Κατά τη διαδικασία αυτή το κείμενο χωρίζεται στις επιμέρους λέξεις του και δημιουργείται μια ακολουθία από λεκτικές μονάδες. Κάθε μονάδα αντιπροσωπεύεται από ένα αναγνωριστικό (token id) και αντιστοιχεί σε ένα εμφύτευμα. Ακόμα, στην κατηγορία μοντέλων παραγωγικού προεκπαιδευμένου μετασχηματιστή η διαδικασία tokenization παράγει και κωδικοποίηση που αφορά τη θέση των λέξεων της ακολουθίας, με αποτέλεσμα να μπορούν να αποτυπωθούν οι σχέσεις διαδοχής των όρων (Radford, Narasimhan, Salimans, & Sutskever, 2018). Στη συνέχεια, οι αναπαραστάσεις προωθούνται ως είσοδος σε ένα μοντέλο αποτελούμενο μεταξύ άλλων από διαδοχικά στρώματα όμοια με εκείνα που αναλύθηκαν στην ενότητα που αφορά τους μετασχηματιστές (transformer blocks). Έπειτα, πραγματοποιείται η πρώτη διαδικασία εκπαίδευσης του μοντέλου, σε ένα μεγάλο σύνολο δεδομένων, με μη επιβλεπόμενο τρόπο που αποσκοπεί στην πρόβλεψη της επόμενης λέξης σε μια ακολουθία λέξεων (Radford, Narasimhan, Salimans, & Sutskever, 2018). Μετά, πραγματοποιείται μια δεύτερη διαδικασία εκπαίδευσης ανάλογα με τον σκοπό της εκάστοτε εργασίας (π.χ. named entity recognition, text classification, machine translation), που είναι επιβλεπόμενη. Στο συγκεκριμένο στάδιο, γίνεται βελτιστοποίηση των εμφυτευμάτων και κωδικοποιείται ταυτόχρονα αναπαράσταση τόσο της σημασίας των λέξεων όσο και του πλαισίου μέσα στο οποίο χρησιμοποιούνται. Μόλις ολοκληρωθούν οι επιμέρους λειτουργίες που αναλύθηκαν παραπάνω μπορούν να εξαχθούν οι αναπαραστάσεις και να χρησιμοποιηθούν σε διαδικασίες μάθησης αντίστοιχες της βελτιστοποίησης που έχει πραγματοποιηθεί.

Παρά τα πλεονεκτήματα που παρέχει η συγκεκριμένη μέθοδος κωδικοποίησης κειμένου έναντι άλλων τεχνικών που αναφέρθηκαν παραπάνω, παρουσιάζει και ορισμένα μειονεκτήματα. Ένα σημαντικό μειονέκτημα εντοπίζεται στη διαδικασία συσχετισμού των λέξεων και εντέλει τον εντοπισμό του πλαισίου χρήσης. Δηλαδή,

δεν γίνεται να εξαχθεί κάποιο συμπέρασμα, όταν ο αριθμός των λέξεων ανά δείγμα υπερβαίνει ένα όριο που μπορεί να διαχειριστεί το μοντέλο μάθησης. Παράλληλα, ένα ακόμα μειονέκτημα είναι πως κατά τη διάρκεια της μη επιβλεπόμενης εκπαίδευσης μπορεί λόγω του συνόλου δεδομένων να υπάρξουν ανακρίβειες ως προς την κωδικοποίηση. Δηλαδή, αν υπάρχουν λέξεις του συνόλου δεδομένων που δεν εμφανίζονται καθόλου ή παρουσιάζονται σε διαφορετικό πλαίσιο συγκριτικά με τον συνήθη τρόπο χρήσης τους θα υπάρξουν προβληματικές αναπαραστάσεις. Τέλος, μπορεί να θεωρηθεί αρνητικό το γεγονός, πως για την διαδικασία που αναλύθηκε προηγουμένως, υπάρχει μεγάλη ανάγκη υπολογιστικών πόρων και πιο συγκεκριμένα χρήση μονάδας επεξεργασίας γραφικών (GPU), συνδυαστικά με μεγάλη απαίτηση μνήμης.



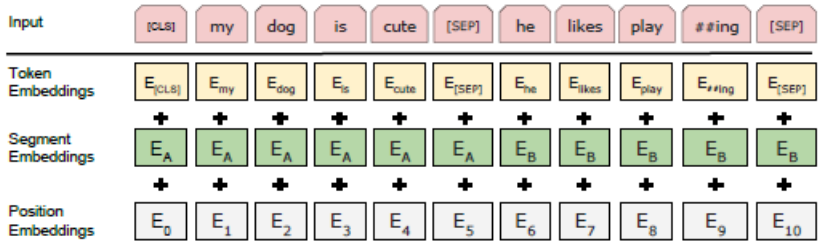
Εικόνα 4 - Περιγραφή της δημιουργίας των εμφυτευμάτων και βελτιστοποίηση ανά διαφορετικό σκοπό. (Radford, Narasimhan, Salimans, & Sutskever, 2018)

2.2.4 Bidirectional Encoder Representations from Transformers (BERT)

Στην παρούσα ενότητα, θα παρουσιαστεί η προσέγγιση από την οποία προέκυψε η μέθοδος κωδικοποίησης που χρησιμοποιείται στο πειραματικό μέρος της εργασίας. Το συγκεκριμένο μοντέλο, κωδικοποιεί τις αναπαραστάσεις με χρήση μετασχηματιστών μελετώντας και τις δύο κατευθύνσεις, δεξιά και αριστερά από την εκάστοτε λέξη. Για συντομία, θα γίνεται αναφορά στο μοντέλο αυτό με τη συντομογραφία BERT (Bidirectional Encoder Representations from Transformers). Η πρώτη και πιο φανερή διαφορά του συγκριτικά με ένα υποτυπώδες μοντέλο GPT είναι πως δεν περιορίζεται στην κατανόηση του πλαισίου χρήσης μια λέξης μελετώντας το κείμενο μόνο από τα αριστερά προς τα δεξιά (unidirectional model). Το γεγονός αυτό, σημαίνει ότι η κάθε λέξη μπορεί να συσχετιστεί ταυτόχρονα από τους μηχανισμούς προσοχής όχι μόνο με όρους που έχουν προηγηθεί αλλά και με όρους που έπονται. Τα αποτελέσματα των συσχετίσεων αυτών συνδυάζονται με σκοπό να προκύψουν όσο το δυνατό πληρέστερες κωδικοποιήσεις.

Η διαδικασία tokenization του κειμένου είναι αρκετά διαφορετική συγκριτικά με τη χρήση ενός μοντέλου GPT. Η επεξεργασία των δειγμάτων προς μελέτη πραγματοποιείται με τη μέθοδο WordPiece (Sennrich, Haddow, & Birch, 2015), η οποία προσφέρει προστασία ως προς τη διαχείριση λέξεων που δεν ανήκουν στο αρχικό λεξιλόγιο του συνόλου δεδομένων. Αυτό συμβαίνει, επειδή οι λέξεις που περιλαμβάνονται στα δείγματα διασπώνται σε μικρότερες λεκτικές μονάδες για να

συνυπολογισθούν όλες οι πιθανές μορφολογικές ιδιαιτερότητες. Παράλληλα, για να λειτουργήσει αποδοτικά η υπόλοιπη διαδικασία προστίθενται ετικέτες που σηματοδοτούν την έναρξη του εκάστοτε δείγματος ([CLS]) και το διαχωρισμό προτάσεων ([SEP]). Όσο αφορά τα επιμέρους δείγματα που αναφέρθηκαν παραπάνω, κατά τη διαδικασία tokenization δημιουργούνται τρεις αντίστοιχες αναπαραστάσεις για κάθε λεκτική μονάδα που περιέχουν (Devlin, Chang, Lee, & Toutanova, 2019). Το πρώτο είδος αναπαραστάσεων ονομάζεται token embeddings όπου όλες οι λεκτικές μονάδες και οι ετικέτες αντικαθίστανται από κάποιο αναγνωριστικό με τυχαίο τρόπο ανάλογα με την αρχιτεκτονική του μοντέλου. Το δεύτερο είδος αναπαράστασης αφορά το χωρισμό του κειμένου σε τμήματα (segment embeddings) με την απόδοση κάποιων αναγνωριστικών. Οι λέξεις που έχουν ίδιο αναγνωριστικό στο συγκεκριμένο είδος αναπαράστασης ανήκουν στο ίδιο τμήμα αλλιώς βρίσκονται σε διαφορετικά τμήματα του κειμένου. Το τελευταίο είδος αναπαράστασης αφορά τη χωρική κωδικοποίηση (position embeddings) των επιμέρους λεκτικών μονάδων και περιέχει πληροφορίες αναφορικά με την απόλυτη και σχετική θέση τους στην ακολουθία λέξεων του αρχικού δείγματος. Στο τέλος, συνδυάζονται τα τρία εμφυτεύματα ανά λεκτική μονάδα και δίνονται ως είσοδο στο μοντέλο (Devlin, Chang, Lee, & Toutanova, 2019). Στο σημείο αυτό, χρειάζεται να αναφερθεί πως η δομή του μοντέλου αξιοποιεί όπως και στην περίπτωση αρχιτεκτονικής παραγωγικού προεκπαιδευμένου μετασχηματιστή (GPT) στρώματα που περιλαμβάνουν συνδυασμό ενός μηχανισμού προσοχής πολλαπλών κεφαλών και ενός δικτύου εμπρόσθιας τροφοδότησης. Έπειτα, σε επόμενο στάδιο πραγματοποιείται η προεκπαίδευση του μοντέλου με σκοπό την επίτευξη δύο βασικών στόχων.

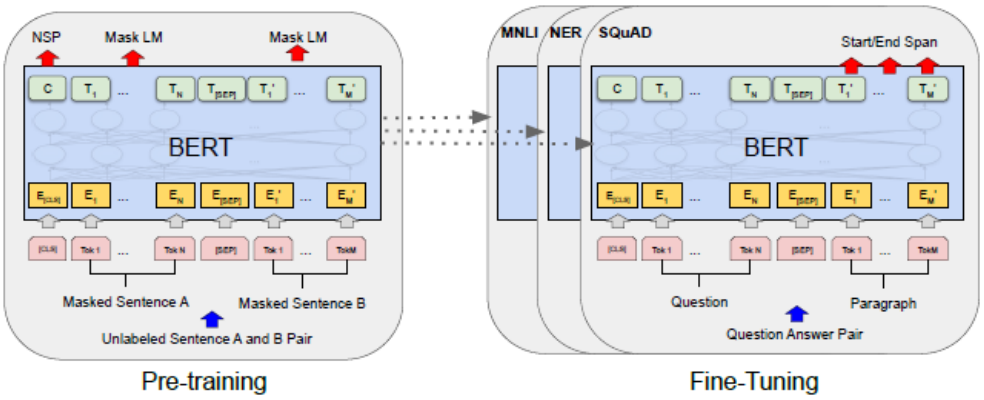


Εικόνα 5-Συνδυασμός των επιμέρους εμφυτευμάτων πριν την διαδικασία της αρχικής μη επιβλεπόμενης εκπαίδευσης (Devlin, Chang, Lee, & Toutanova, 2019)

Όπως αναλύθηκε παραπάνω, γίνεται μελέτη του πλαισίου χρήσης των λέξεων σε δύο κατευθύνσεις. Όμως, δεν είναι εφικτό να συμβεί η διαδικασία αυτή ταυτόχρονα από ένα μοντέλο, γιατί ο συνδυασμός των δύο ερμηνειών μια λεκτικής μονάδας θα οδηγήσει σε περιπτώσεις όπου μια λέξη συσχετίζεται με τον εαυτό της (Devlin, Chang, Lee, & Toutanova, 2019). Έτσι, προκύπτει ο πρώτος στόχος της αρχικής εκπαίδευσης που στοχεύει στην επίλυση αυτού του προβλήματος. Εφαρμόζεται μια μάσκα με τυχαίο τρόπο σε ορισμένες λέξεις των δειγμάτων της εισόδου και στη συνέχεια πραγματοποιείται προσπάθεια προσδιορισμού της εκάστοτε λεκτικής μονάδας με χρήση των δύο ξεχωριστών ερμηνειών. Ο συγκεκριμένος στόχος της εκπαίδευσης αναφέρεται ως Masked Language Model και συναντάται στη βιβλιογραφία με τη συντομογραφία MLM. Ο δεύτερος στόχος της αρχικής εκπαίδευσης, στοχεύει στην πρόβλεψη της επόμενης πρότασης (NSP: Next Sentence Prediction). Είναι μια λογική επιλογή ο συγκεκριμένος στόχος αφού σε πολλές

κατηγορίες εργασιών επεξεργασίας φυσικής γλώσσας παρουσιάζεται η ανάγκη να αναγνωριστεί η συσχέτιση μεταξύ προτάσεων οι οποίες δεν συνδέονται με φανερό τρόπο. Χαρακτηριστικά παραδείγματα τέτοιου είδους εργασιών αποτελούν η απάντηση ερωτήσεων (QA: Question Answering) και ο συμπερασμός φυσικής γλώσσας (NLI: Natural Language Inference).

Στην συνέχεια, γίνεται δεύτερη εκπαίδευση του μοντέλου η οποία αποτελεί μια επιβλεπόμενη διαδικασία μάθησης. Η συγκεκριμένη φάση εκπαίδευσης, βασίζεται σε δεδομένα τα οποία έχουν χαρακτηριστεί κατάλληλα με ετικέτες. Τελικά, μόλις ολοκληρωθεί η διαδικασία μπορεί να γίνει εξαγωγή των εμφυτευμάτων από διαφορετικά επίπεδα του μοντέλου ανάλογα με την επιθυμητή χρήση τους. Τα εμφυτεύματα μπορούν να χρησιμοποιηθούν στη συνέχεια από μοντέλα μάθησης για διάφορους σκοπούς με την προϋπόθεση πως συμβαδίζουν με το στάδιο βελτιστοποίησης που έχει προηγηθεί. Για παράδειγμα, αν έχει γίνει βελτιστοποίηση των εμφυτευμάτων σε εργασίες συμπερασμού φυσικής γλώσσας δεν είναι ιδανικό να χρησιμοποιηθούν οι αναπαραστάσεις στην πρόβλεψη απαντήσεων σε ερωτήσεις. Ταυτόχρονα, υπάρχουν και μειονεκτήματα στη χρήση του συγκεκριμένου μοντέλου κωδικοποίησης κειμένου όπως για παράδειγμα οι απαιτήσεις για υπολογιστικούς πόρους, ο υπερβολικά μεγάλος αριθμός παραμέτρων και οι μεγάλοι χρόνοι εκπαίδευσης για προσαρμογή των βαρών.

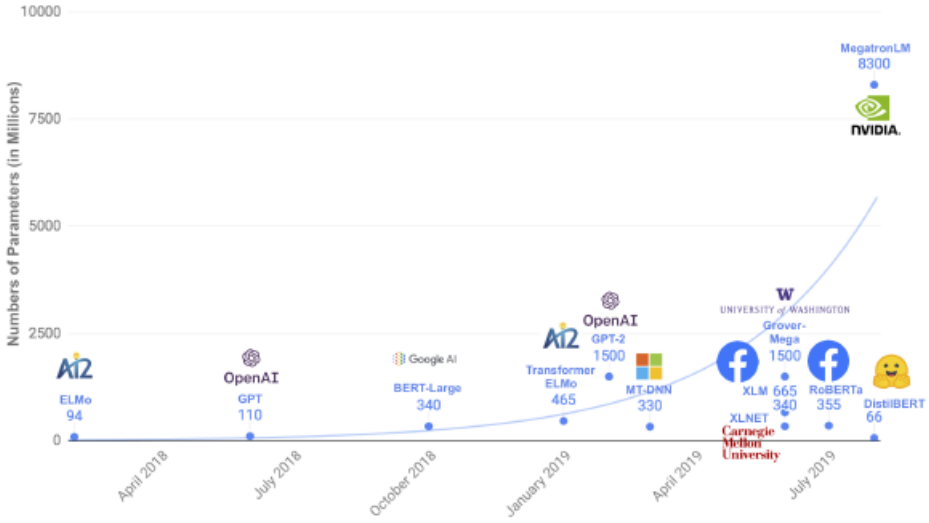


Εικόνα 6 - Περιγραφή των δύο διαδικασιών εκπαίδευσης του BERT (Devlin, Chang, Lee, & Toutanova, 2019)

2.3 DistilBERT

Έπειτα, θα πραγματοποιηθεί παρουσίαση της μεθόδου κωδικοποίησης που χρησιμοποιήθηκε στην παρούσα εργασία. Τα χρόνια που ακολούθησαν μετά τη δημιουργία μοντέλων μάθησης όπως το BERT και το GPT άρχισαν να εμφανίζονται αρχιτεκτονικές με όλο και περισσότερες παραμέτρους. Μάλιστα, υπήρξε η διαπίστωση πως όσο αυξάνονται τα εκατομμύρια παραμέτρων που ήδη χρησιμοποιούνται σε προεκπαιδευμένα μοντέλα κωδικοποίησης τα αποτελέσματα θα είναι όλο και καλύτερα (Sanh, Debut, Chaumond, & Wolf, 2020). Η αρνητική επίδραση αυτού του φαινομένου αφορά το περιβαλλοντικό κόστος χρήσης όλο και πιο απαιτητικών προσεγγίσεων σε υπολογιστικούς πόρους και συνεπώς σε ενεργειακούς πόρους (Sanh, Debut, Chaumond, & Wolf, 2020). Ακόμα, η χρήση τέτοιου είδους αρχιτεκτονικών οδηγεί σε μεγάλες περιόδους αναμονής σε υπηρεσίες

πραγματικού χρόνου και χαμηλή απόδοση σε συσκευές με περιορισμένες δυνατότητες.

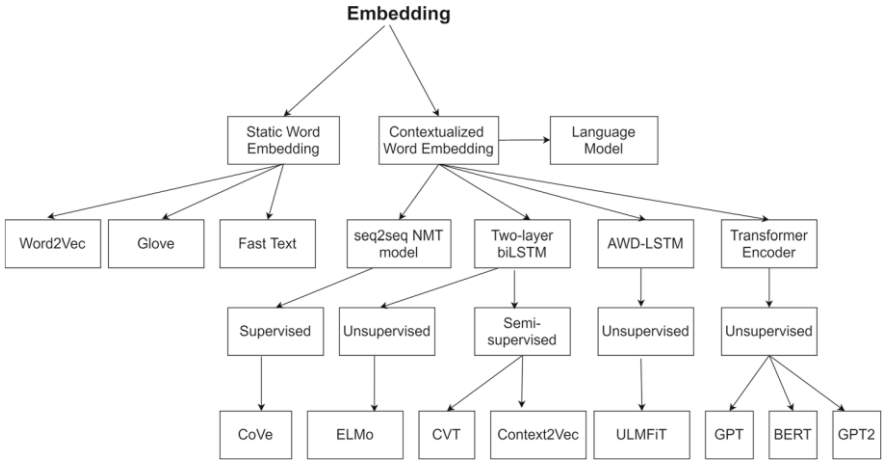


Εικόνα 7 - Παρουσίαση του αυξανόμενου αριθμού παραμέτρων με την πάροδο του χρόνου (Sanh, Debut, Chaumond, & Wolf, 2020).

Η προτεινόμενη λύση στο πρόβλημα αυτό ήταν η δημιουργία μοντέλων τα οποία θα είναι λιγότερο απαιτητικά πετυχαίνοντας όμως παρόμοιες επιδόσεις. Έτσι, παρουσιάστηκαν προεκπαιδευμένα μοντέλα βασισμένα στην τεχνική της απόσταξης γνώσης (knowledge distillation). Πρόκειται για μια μέθοδο συμπίεσης που στοχεύει στην εκπαίδευση ενός μικρότερου μοντέλου (student) με σκοπό να αναπαράγει τη συμπεριφορά ενός μεγαλύτερου μοντέλου (teacher) ή μιας ολόκληρης ομάδας αρχιτεκτονικών. Το αποτέλεσμα της μεθόδου στην περίπτωση των εργασιών επεξεργασίας φυσικής γλώσσας χρησιμοποιώντας το BERT, οδήγησε στη δημιουργία του DistilBERT. Το μοντέλο που προέκυψε είναι συγκριτικά ελαφρύτερο και ταχύτερο στο συμπερασμό των αναπαραστάσεων κειμένου. Παράλληλα, χρειάζεται λιγότερους υπολογιστικούς πόρους και παρουσιάζει μειωμένη διάρκεια ως προς τον χρόνο εκπαίδευσης.

Όπως συμβαίνει στο BERT, το DistilBERT έχει δύο σκοπούς κατά την πρώτη διαδικασία εκπαίδευσης. Ο πρώτος σκοπός αφορά την επεξεργασία των λεκτικών μονάδων στις οποίες έχει εφαρμοστεί κάποια μάσκα και αξιοποιείται για την κωδικοποίηση δύο κατευθύνσεων όπως στο μοντέλο BERT. Ο δεύτερος σκοπός αφορά την εκπαίδευση του μοντέλου να μιμείται τη συμπεριφορά του BERT με τη διαδικασία απόσταξης γνώσης (knowledge distillation). Δηλαδή, το μοντέλο DistilBERT εκπαιδεύεται στην παραγωγή αποτελεσμάτων όμοιων του BERT και για αυτό το λόγο δεν χρειάζεται να έχει τόσο μεγάλο αριθμό παραμέτρων όσο εκείνο. Η αρχιτεκτονική του DistilBERT δεν διαφέρει σε μεγάλο βαθμό από εκείνη που παρουσιάστηκε στην προηγούμενη ενότητα. Η διαφοροποίηση ως προς την δομή του μοντέλου είναι πως πλέον τα επιμέρους στρώματα έχουν μειωθεί κατά το ήμισυ όμως διατηρούν την ίδια λειτουργικότητα και την ίδια δομή. Μια σημαντική ομοιότητα μεταξύ των δύο προσεγγίσεων είναι το γεγονός, πως μπορούν να δεχθούν το πολύ έως 512 λεκτικές μονάδες, μετά την προσθήκη των ειδικών ετικετών, διατηρώντας την αποτελεσματικότητά τους.

Το DistilBERT είναι 40% μικρότερο, 60% ταχύτερο από το BERT και διατηρεί το 97% της ικανότητάς του να κατανοεί το κείμενο που επεξεργάζεται (Sanh, Debut, Chaumond, & Wolf, 2020). Η συγκεκριμένη περίπτωση αρχιτεκτονικής είναι ιδανική για εργασίες σε περιβάλλοντα με περιορισμένη διαθεσιμότητα πόρων και για εφαρμογές υπολογιστικής παρυφών (edge computing). Στη συνέχεια, δίνεται περιγραφή των διαφορετικών κατηγοριών κωδικοποίησης κειμένου.



Εικόνα 8 - Παρουσίαση μεθόδων κωδικοποίησης κειμένου (Goswami, Kaliyar, & Narang, 2021)

Το τελικό συμπέρασμα είναι πως υπάρχουν πολλές διαφορετικές μέθοδοι κωδικοποίησης με διαφορετικά μειονεκτήματα και πλεονεκτήματα. Ανάλογα με τη φύση των δεδομένων και το σκοπό της εκάστοτε διαδικασίας μάθησης χρειάζεται να γίνει επιλογή της καταλληλότερης μεθόδου.

Πίνακας 1 - Συνοπτική παρουσίαση διαδεδομένων μεθόδων κωδικοποίησης κειμένου

	Μεθοδολογία	**Πλεονεκτήματα & Μειονεκτήματα**
TF-IDF	Το κάθε έγγραφο (δείγμα) του συνόλου δεδομένων κωδικοποιείται σε ένα διάνυσμα. Κάθε λέξη αντιστοιχεί σε ένα στοιχείο του διανύσματος ανάλογα με τη σειρά που εμφανίζεται. Οι λέξεις κωδικοποιούνται με βάση το γινόμενο δύο μετρικών, τη μετρική TF (term-frequency) και τη μετρική IDF (inverse document frequency). Η πρώτη μετρική αποδίδει τη βαρύτητα μιας λέξης ανάλογα με τη συχνότητά εμφάνισής της στο κείμενο. Η δεύτερη μετρική αποδίδει την ικανότητα μια λέξης να διαχωρίσει τα κείμενα σε κατηγορίες. Σε όσα περισσότερα κείμενα συναντάται η λέξη τόσο μικρότερη είναι η ικανότητά της να διαχωρίσει τα δείγματα.	+ Πρόκειται για μια αποτελεσματική μέθοδο στον τομέα της ανάκτησης πληροφορίας. Η ιδιότητα αυτή προκύπτει επειδή η μέθοδος βασίζεται κυρίως στη συχνότητα εμφάνισης των λέξεων. + Είναι μια υπολογιστικά οικονομική μέθοδος και για αυτό το λόγο μπορεί να κλιμακωθεί πολύ πιο εύκολα συγκριτικά με άλλες τεχνικές. - Δεν αξιοποιείται το πλαίσιο χρήσης της κάθε λέξης.
Word2Vec	Η μέθοδος κωδικοποιεί τις λέξεις του συνόλου δεδομένων σε διανύσματα αξιοποιώντας τα μοντέλα μάθησης CBOW και Skip-gram. Η βασική αρχή της μεθόδου είναι πως οι λέξεις αποκτούν νόημα από τους κοντινούς όρος που προηγούνται και έπονται. Τα μοντέλα μάθησης που αναφέρθηκαν αξιοποιούν τις κοντινές λέξεις των όρων για να δημιουργήσουν τις αντίστοιχες αναπαραστάσεις.	+ Μελετάται κάθε λέξη βάσει των κοντινών όρων. + Πρόκειται για μια υπολογιστικά οικονομική μέθοδο όταν χρησιμοποιούνται προεκπαιδευμένα εμφυτεύματα. + Οι αναπαραστάσεις για κάθε όρο είναι πιο περιγραφικές καθώς αυξάνονται οι διαστάσεις των εμφυτευμάτων. - Δημιουργούνται στατικές αναπαραστάσεις για κάθε λέξη. - Αμελείται η επιρροή όρων, που βρίσκονται πολύ μακριά από μια λέξη. - Πρόκειται για μια μέθοδο που δεν μπορεί να διαχειριστεί επιτυχώς λέξεις, τις οποίες δεν έχει συναντήσει στα δεδομένα εκπαίδευσης. - Η δημιουργία των εμφυτευμάτων μπορεί να είναι υπολογιστικά ακριβή.

GloVe	Η μέθοδος κωδικοποιεί τις λέξεις του συνόλου δεδομένων σε διανύσματα. Η κωδικοποίηση πραγματοποιείται με βάση την πιθανότητα να συνυπάρχουν δύο οποιεσδήποτε λέξεις του συνόλου δεδομένων μεταξύ τους. Η ιδέα από την οποία πηγάζει αυτή η προσέγγιση είναι πως η σημασιολογική ερμηνεία του εκάστοτε όρου προσδιορίζεται από το σύνολο των λέξεων που τον περιτριγυρίζουν.	+ Οι λέξεις δεν αναπαρίστανται με βάση μόνο τους κοντινούς όρους. + Πρόκειται για μια αποδοτική μέθοδο όταν χρησιμοποιούνται προεκπαιδευμένα εμφυτεύματα. - Δεν δίνεται έμφαση στις περιπτώσεις που μια λέξη καθορίζεται πλήρως από τους όρους που εμφανίζονται πριν και μετά από αυτή. - Οι αναπαραστάσεις που προκύπτουν είναι στατικές. - Η διαδικασία δημιουργίας των εμφυτευμάτων είναι υπολογιστικά ακριβή.
ELMo	Η παρούσα μέθοδος κωδικοποιεί τις λέξεις των δειγμάτων σε εμφυτεύματα. Η συγκεκριμένη τεχνική, αναπαριστά το κείμενο αξιοποιώντας μοντέλα μάθησης που επιστρατεύουν στρώματα μακράς βραχύχρονης μνήμης δύο κατευθύνσεων (bidirectional LSTM). Αρχικά, πραγματοποιείται μη επιβλεπόμενη εκπαίδευση στα δείγματα του συνόλου δεδομένων με σκοπό την πρόβλεψη επόμενης λέξης. Έπειτα, ακολουθεί δεύτερη διαδικασία επιβλεπόμενης μάθησης που καθορίζει το ποσοστό συμμετοχής των εμφυτευμάτων, που προκύπτουν από τα επιμέρους στρώματα, στον εκάστοτε σκοπό επεξεργασίας φυσικής γλώσσας	+ Πρόκειται για μια μέθοδο η οποία συλλαμβάνει το νόημα των όρων ανάλογα με τον τρόπο που χρησιμοποιούνται. + Τα εμφυτεύματα είναι δυναμικά. Πρακτικά μια λέξη ενός δείγματος κειμένου μπορεί να κωδικοποιείται σε δύο προτάσεις, με διαφορετικό τρόπο, ανάλογα με τη χρήση της. + Η παρούσα μέθοδος μελετά τόσο τις λέξεις που προηγούνται όσο και εκείνες που έπονται για να κωδικοποιήσει μια λεκτική μονάδα. + Μπορούν να χρησιμοποιηθούν προεκπαιδευμένα εμφυτεύματα για να αποφευχθεί το υψηλό υπολογιστικό κόστος. - Πρόκειται για μια μέθοδο που απαιτεί αρκετούς υπολογιστικούς πόρους. - Υπάρχει μεγάλη εξάρτηση της απόδοσης του μοντέλου στα δεδομένα που χρησιμοποιούνται κατά τη διαδικασία εκπαίδευσης. - Όπως και σε όλες τις προηγούμενες τεχνικές ένα σύνολο δεδομένων με έντονη ανισορροπία δεν μπορεί να αντιμετωπιστεί αν δεν υπάρξει κάποια προηγούμενη επεξεργασία.

GPT	Πρόκειται για μια μέθοδο κωδικοποίησης των λέξεων σε εμφυτεύματα. Η συγκεκριμένη τεχνική αξιοποιεί μοντέλα μάθησης τα οποία βασίζουν τη λειτουργία τους σε μηχανισμούς προσοχής που παρέχονται από δομές μετασχηματιστών (transformers). Αρχικά, πραγματοποιείται διαδικασία μη επιβλεπόμενης εκπαίδευσης στο σύνολο δεδομένων και παράγονται τα εμφυτεύματα. Στη συνέχεια, γίνεται βελτιστοποίηση των εμφυτευμάτων ανάλογα με την εκάστοτε διαδικασία επεξεργασίας φυσικής γλώσσας.	+ Η συγκεκριμένη τεχνική αξιοποιεί μηχανισμούς προσοχής για να συλλάβει το πλαίσιο χρήσης κάθε λέξης. Μελετά τις λέξεις ανάλογα με τη σχέση που έχουν με όρους που έχουν προηγηθεί, ενώ ταυτόχρονα μελετώνται και οι κοντινοί τους όροι. + Δημιουργούνται εμφυτεύματα τα οποία δεν είναι στατικά για κάθε διαφορετική λέξη. + Μπορεί να είναι μια ιδιαίτερα αποδοτική τεχνική όταν χρησιμοποιούνται προεκπαιδευμένα εμφυτεύματα. - Είναι μια μέθοδος υπολογιστικά ακριβή. - Η ποιότητα των εμφυτευμάτων εξαρτάται σε μεγάλο βαθμό από το σύνολο δεδομένων εκπαίδευσης όπως συμβαίνει και στην περίπτωση της τεχνικής ELMo.
BERT	Είναι μια μέθοδος κωδικοποίησης των λέξεων σε εμφυτεύματα. Η συγκεκριμένη τεχνική αξιοποιεί μοντέλα μάθησης τα οποία βασίζουν τη λειτουργία τους σε μηχανισμούς προσοχής που παρέχονται από δομές μετασχηματιστών (transformers). Αρχικά, πραγματοποιείται διαδικασία μη επιβλεπόμενης εκπαίδευσης στο σύνολο δεδομένων και παράγονται τα εμφυτεύματα. Οι δύο βασικοί στόχοι της πρώτης διαδικασίας μάθησης είναι οι MLM και NSP που έχουν αναλυθεί λεπτομερώς σε προηγούμενη ενότητα. Ο πρώτος σκοπός αφορά τη δυνατότητα του μοντέλου να προσδιορίζει το νόημα των λέξεων ως προς δύο κατευθύνσεις με χρήση μετασχηματιστών. Ο δεύτερος σκοπός αφορά τη δυνατότητα πρόβλεψης της επόμενης πρότασης. Στη συνέχεια με διαδικασία επιβλεπόμενης μάθησης γίνεται βελτιστοποίηση των εμφυτευμάτων ανάλογα με την εκάστοτε διαδικασία επεξεργασίας φυσικής γλώσσας.	+ Η συγκεκριμένη τεχνική αξιοποιεί μηχανισμούς προσοχής για να συλλάβει το πλαίσιο χρήσης κάθε λέξης. Μελετά τις λέξεις ανάλογα με τη σχέση που έχουν με όρους που προηγούνται ή έπονται, ενώ ταυτόχρονα μελετώνται και οι κοντινοί της όροι. + Δημιουργούνται εμφυτεύματα τα οποία δεν είναι στατικά για κάθε διαφορετική λέξη. + Μπορεί να είναι μια ιδιαίτερα αποδοτική τεχνική όταν χρησιμοποιούνται προεκπαιδευμένα εμφυτεύματα. - Είναι μια μέθοδος υπολογιστικά ακριβή. - Η ποιότητα των εμφυτευμάτων εξαρτάται σε μεγάλο βαθμό από το σύνολο δεδομένων εκπαίδευσης όπως συμβαίνει και στην περίπτωση της τεχνικής ELMo.

DistilBERT	Η συγκεκριμένη τεχνική είναι παρόμοια με την μέθοδο κωδικοποίησης που αξιοποιεί το μοντέλο BERT. Το DistilBERT αποτελεί μια συμπιεσμένη έκδοση του BERT που διαφέρει ως προς τον τελικό αριθμό παραμέτρων και των στόχων στο στάδιο της μη επιβλεπόμενης εκπαίδευσης. Το DistilBERT κατά τη διάρκεια της πρώτης διαδικασίας μάθησης έχει σαν στόχο να μπορεί να αναπαράγει τη συμπεριφορά του μοντέλου BERT. Παράλληλα, εκπαιδεύεται για το σκοπό MLM που αφορά την ανίχνευση του πλαισίου χρήσης μιας λέξης ως προς τις δύο κατευθύνσεις του κειμένου.	+ Προσφέρει τα σημαντικότερα πλεονεκτήματα του μοντέλου BERT. + Έχει μειωμένο αριθμό παραμέτρων. Για αυτό το λόγο είναι μια αποδοτικότερη μέθοδος συγκριτικά με άλλες τεχνικές που αξιοποιούν μηχανισμούς προσοχής
		- Όπως και σε άλλες περιπτώσεις υπάρχει εξάρτηση από τα σύνολα δεδομένων που αξιοποιούνται. - Δεν διατηρείται πλήρως η απόδοση του μοντέλου BERT αφού το DistilBERT είναι εκπαιδευμένο απλώς να μιμείται τη συμπεριφορά του.

3 Ανάλυση συνόλων δεδομένων

3.1 Σημασία εύρεσης κατάλληλων συνόλων δεδομένων

Έχοντας αναλύσει τη βασική μέθοδο κωδικοποίησης που θα χρησιμοποιηθεί στη συγκεκριμένη εργασία, καθώς και τη σύνδεσή της με προηγούμενες μεθοδολογίες, θα ακολουθήσει παρουσίαση των συνόλων δεδομένων. Στο παρόν κεφάλαιο, θα εξηγηθεί η ανάγκη ανάλυσης διαφορετικών συνόλων δεδομένων και θα αναφερθούν οι δυσκολίες πραγματοποίησης της συγκεκριμένης διαδικασίας. Ακόμα, θα γίνει προσπάθεια να εξακριβωθούν τα θεωρητικά συμπεράσματα που έχουν ήδη παρουσιαστεί αναφορικά με τα χαρακτηριστικά των fake news, όπως για παράδειγμα ο αρνητικός τόνος του συγγραφέα. Επιπροσθέτως, θα παρουσιαστούν μέσω των επιμέρους συνόλων δεδομένων επιπλέον κατηγοριοποιήσεις ως προς τα είδη των fake news και την εγκυρότητα των δειγμάτων.

Ο βασικότερος παράγοντας, που καθιστά δύσκολη την εύρεση και την επιλογή κάποιου συγκεκριμένου συνόλου δεδομένων, είναι το γεγονός πως χρειάζεται τα δεδομένα να είναι χαρακτηρισμένα ως αληθή ή ως ψευδή. Δηλαδή, είναι αναγκαίο να υπάρχει ένας μηχανισμός, που να έχει ταξινομήσει το κάθε δείγμα με αξιοπιστία έτσι ώστε το αποτέλεσμα των μεθόδων μάθησης να είναι όσο το δυνατό πιο ακριβές. Για αυτό το λόγο χρησιμοποιούνται τεχνικές μηχανισμών γνώσης (knowledge-based) για να αξιολογηθούν τα δείγματα. Συγκεκριμένα, χρειάζονται ειδικοί σε διαφορετικούς τομείς, οι οποίοι μπορούν να αξιολογήσουν τα δείγματα, που σε συνδυασμό με έμπιστες πηγές, τον προσδιορισμό του επικοινωνιακού πλαισίου και επιπλέον στοιχεία, μπορούν να οδηγήσουν σε ένα τελικό συμπέρασμα (Shu, Wang, Silva, Tang, & Liu, 2017). Ωστόσο, ακόμα και στην περίπτωση αυτή, δεν υπάρχει εγγύηση για τους τελικούς χαρακτηρισμούς, αφού είναι υπαρκτό το ενδεχόμενο λανθασμένων εκτιμήσεων. Ακόμα, έχουν κάνει την εμφάνισή τους σύνολα δεδομένων για ανίχνευση fake news, τα οποία βασίζονται σε απόδοση ετικετών με χρήση μεθόδων μάθησης (Shu, Wang, Silva, Tang, & Liu, 2017). Συνήθως, γίνεται εξαγωγή χαρακτηριστικών από το εκάστοτε μοντέλο μέσω μη επιβλεπόμενης εκπαίδευσης και στη συνέχεια αποδίδονται οι κατάλληλες ετικέτες στα δείγματα.

Ένας ακόμα παράγοντας που δημιουργεί επιπλοκές είναι η απουσία κάποιου καθιερωμένου συνόλου δεδομένων ως μέτρο σύγκρισης για την αποτελεσματικότητα των πιθανών μεθόδων ανίχνευσης. Δηλαδή, δεν υπάρχει ένα σύνολο δεδομένων με αρκετά δείγματα το οποίο να συνδυάζει τα γλωσσικά χαρακτηριστικά όλων των περιβαλλόντων στα οποία συναντώνται fake news (Shu, Wang, Silva, Tang, & Liu, 2017). Έτσι, σε ορισμένα σύνολα δεδομένων περιέχονται άρθρα από πρακτορεία ειδήσεων, σε άλλα υπάρχουν δεδομένα από ομιλίες και συνεντεύξεις πολιτικών προσώπων και σε άλλες περιπτώσεις υπάρχουν δεδομένα από δημοσιεύσεις σε μέσα κοινωνικής δικτύωσης, όπως για παράδειγμα το Twitter. Επιπλέον, σε κάθε ξεχωριστό σύνολο δεδομένων παρουσιάζεται διαφορετικός τρόπος κατηγοριοποίησης των δειγμάτων ο οποίος δεν περιορίζεται αποκλειστικά σε διαχωρισμό ψευδών και αληθών δειγμάτων. Για αυτόν τον λόγο, προκειμένου να διαπιστωθεί η

αποτελεσματικότητα μιας προσέγγισης είναι αναγκαίο να δοκιμαστεί η απόδοσή της σε ένα πλήθος διαφορετικών συνόλων.

Προτού παρουσιαστεί αναλυτικά η οργάνωση και το περιεχόμενο του κάθε συνόλου δεδομένων, θα γίνει μια σύντομη αναφορά σε έναν φορέα που είναι υπεύθυνος για χαρακτηρισμό των δειγμάτων. Με αυτόν τον τρόπο, θα γίνει πιο κατανοητή η διαδικασία απόδοσης ετικετών και τα χαρακτηριστικά τέτοιου είδους οργανισμών. Η ανεξάρτητη ιστοσελίδα PolitiFact, που ανήκει σε μη κερδοσκοπική εταιρεία, έχει δημοσιογραφικό χαρακτήρα και ειδικεύεται στη διασταύρωση ειδήσεων, όπως αναφέρει και στον ιστότοπό[2] της. Η ιστοσελίδα λαμβάνει περιεχόμενο άξιο ελέγχου από το κοινό, από άλλες δημοσιογραφικές ιστοσελίδες και από μέσα κοινωνικής δικτύωσης όπως το Facebook και το TikTok. Οι δημοσιογράφοι δεν εκφράζουν την προσωπική τους άποψη, επίσης η τελική αξιολόγηση της πληροφορίας βασίζεται αποκλειστικά σε έγκυρες πηγές και πραγματοποιείται μόνο για ειδήσεις που είναι εφικτό να διασταυρωθούν. Τα περισσότερα σύνολα δεδομένων που χρησιμοποιούνται για την ανίχνευση των fake news, έχουν δημιουργηθεί με βάση ψευδή νέα που έχουν εντοπιστεί από ιστοσελίδες όπως η PolitiFact. Πρόκειται για οργανισμούς, που ακολουθούν διαφανή πρωτόκολλα ως προς τις αξιολογήσεις τους και αυτό επιβεβαιώνεται από το γεγονός πως κάποιοι από αυτούς έχουν λάβει ακόμα και βραβείο Pulitzer[3].

3.2 Παρουσίαση και διαδικασία ανάλυσης συνόλων δεδομένων

Στη συνέχεια του κεφαλαίου, θα πραγματοποιηθεί ανάλυση των συνόλων δεδομένων ως προς το περιεχόμενο των δειγμάτων τους, των θεματικών ενοτήτων που απασχολούν, το μέσο μήκος του κειμένου ανά δείγμα και άλλων χρήσιμων πληροφοριών που μπορούν να προσφέρουν.

3.2.1 ISOT

Το παρόν σύνολο δεδομένων ανίχνευσης fake news περιέχει 44.898 δείγματα και αποτελείται από δύο κύρια υποσύνολα ένα με αληθή και ένα με ψευδή δείγματα. Η δημιουργία του συνόλου δεδομένων πραγματοποιήθηκε ως τμήμα του κύκλου εργασιών του τμήματος μηχανικής (ISOT: Information Security and Object Technology Research Lab) του University of Victoria και αφορά δείγματα της περιόδου 2016-2017 (Ahmed, Traore, & Saad, 2018). Το αληθές περιεχόμενο (21.417 δείγματα) αφορά άρθρα που έχουν συλλεχθεί από την ειδησεογραφική ιστοσελίδα Reuters.com (Ahmed, Traore, & Saad, 2018). Τα δείγματα που είναι χαρακτηρισμένα ως ψευδή (συνολικά 23.481 δείγματα) έχουν συλλεχθεί από πηγές οι οποίες έχουν χαρακτηριστεί ως ύποπτες για διασπορά ψευδών ειδήσεων από ιστοσελίδες που

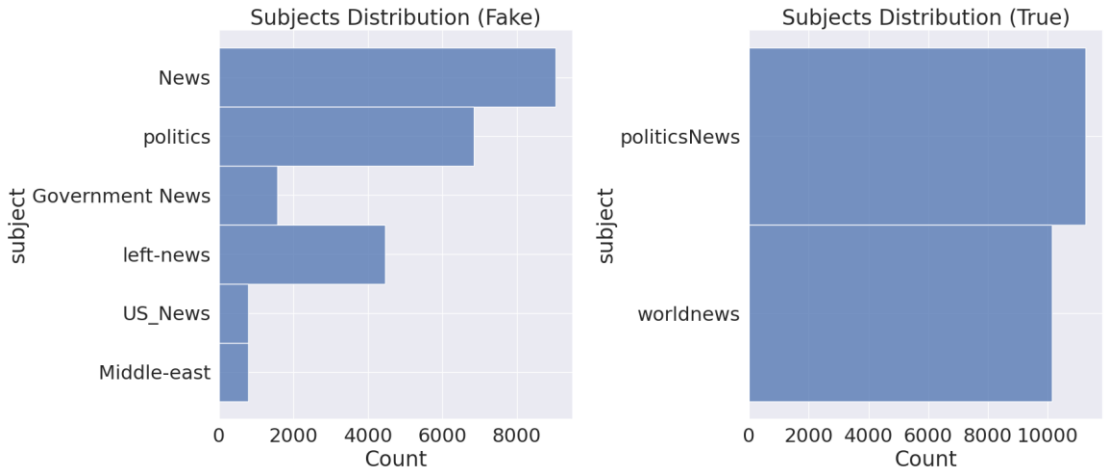
[2] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/
[3] https://www.pulitzer.org/winners/staff-69

εξετάζουν την αξιοπιστία των γεγονότων, όπως PolitiFact.com και Wikipedia.com (Ahmed, Traore, & Saad, 2018).

Το κάθε δείγμα του συνόλου δεδομένων αποτελείται από τέσσερα χαρακτηριστικά τα οποία είναι ο τίτλος του άρθρου, το ίδιο το άρθρο, η θεματική στην οποία εντάσσεται και τέλος η αρχική ημερομηνία δημοσίευσης. Οι ονομασίες των χαρακτηριστικών αυτών παρατίθενται με τη σειρά που αναφέρθηκαν, όπως εμφανίζονται στο σύνολο δεδομένων.

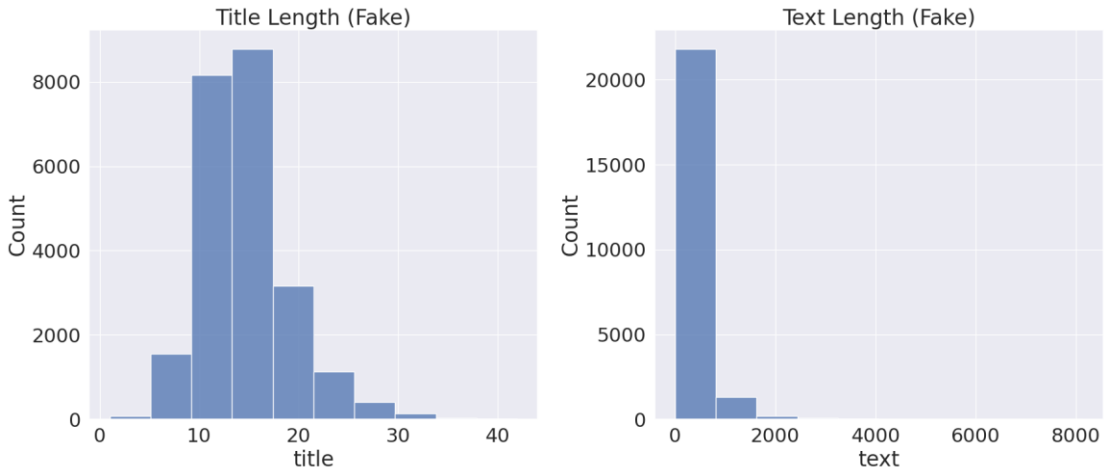
- **title**
- **text**
- **subject**
- **date**

Το σύνολο δεδομένων με βάση τις πληροφορίες που έχουν αναφερθεί και παραπάνω, μπορεί να θεωρηθεί ισορροπημένο και προσφέρει αρκετά ικανοποιητικό όγκο πληροφορίας. Οι θεματικές κατηγορίες που εντοπίζονται και η κατανομή των άρθρων σε αυτές δίνονται παρακάτω.



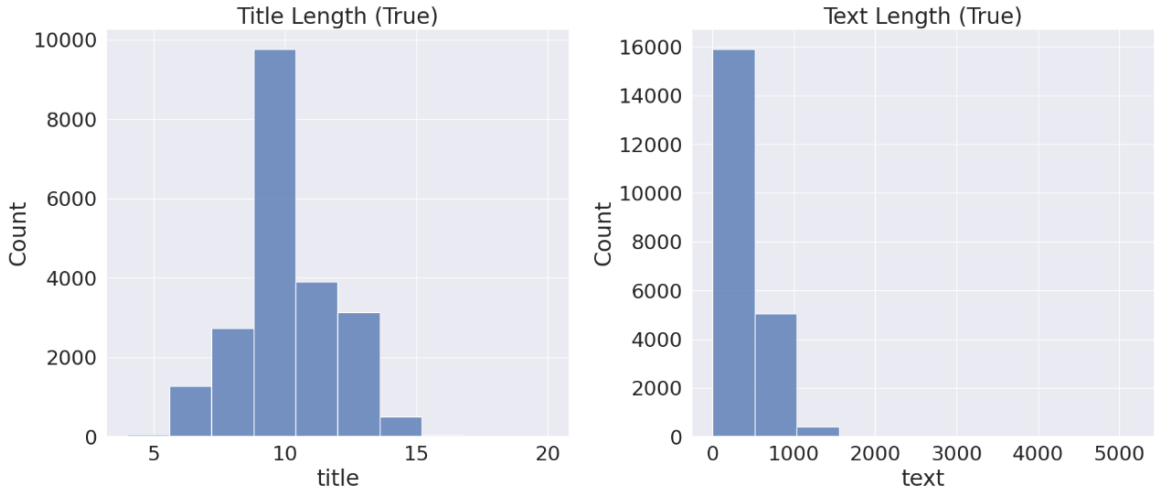
Εικόνα 9 - Κατανομή των άρθρων σε κάθε σύνολο (Fake/True) ως προς τη θεματολογία τους.

Στα προηγούμενα διαγράμματα, τα ψευδή άρθρα καλύπτουν ένα μεγαλύτερο εύρος θεμάτων, όμως στην πλειοψηφία τους ασχολούνται με ειδήσεις γενικότερου ενδιαφέροντος και την πολιτική. Παρατηρείται, πως τα αληθή άρθρα επικεντρώνονται μόνο σε δύο βασικές κατηγορίες όπως η πολιτική και τα διεθνή νέα. Στη συνέχεια, θα εξεταστεί η κατανομή των μηκών για τα ψευδή δείγματα τόσο ως προς τα άρθρα όσο και ως προς τον τίτλο.



Εικόνα 10 - Παρουσίαση κατανομών των μηκών των ψευδών τίτλων και άρθρων.

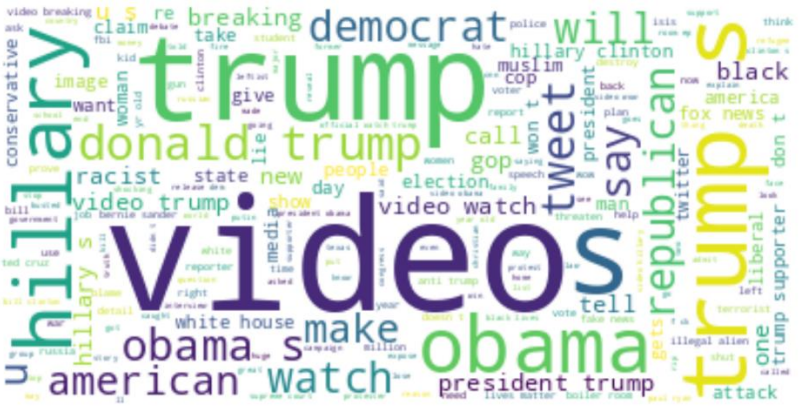
Παρατηρείται πως οι τίτλοι έχουν σχετικά μικρό μήκος. Το μέγιστο μήκος των τίτλων είναι 42 λέξεις και το μέσο μήκος τους είναι 15 λέξεις. Όλοι οι τίτλοι μπορούν να αποτελέσουν είσοδο στο μοντέλο που χρησιμοποιείται κατά την κωδικοποίηση του κειμένου, αφού δεν υπερβαίνουν το όριο των 512 λεκτικών μονάδων που αναφέρθηκε πρωτύτερα. Ωστόσο, δεν ισχύει το ίδιο για το περιεχόμενο των άρθρων. Το μεγαλύτερο άρθρο έχει μήκος 8.136 λεκτικών μονάδων και κατά μέσο όρο τα άρθρα έχουν μήκος 424 λέξεις. Από τα συνολικά 23.481 ψευδή δείγματα μόλις τα 16.740 έχουν μήκος μικρότερο από 480 λέξεις, στις οποίες όμως θα πρέπει να προστεθούν και οι ειδικές ετικέτες. Οπότε με σχετική σιγουριά μόνο για τα συγκεκριμένα δείγματα μπορεί να διασφαλιστεί πως δεν θα ξεπεραστούν οι 512 λεκτικές μονάδες. Στη συνέχεια θα γίνει η ίδια ανάλυση για τα μήκη των αληθών τίτλων και άρθρων.



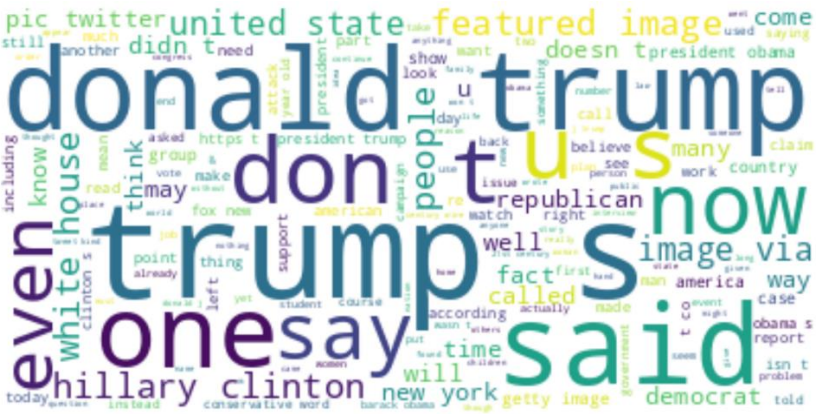
Εικόνα 11 - Παρουσίαση κατανομών των μηκών των αληθών τίτλων και άρθρων.

Παρατηρώντας προσεκτικά τις κατανομές είναι φανερό πως οι τίτλοι δεν αποτελούν πρόβλημα για τη διαδικασία κωδικοποίησης. Το μέσο μήκος των τίτλων είναι 10 λέξεις και ο μεγαλύτερος τίτλος αποτελείται από 20 λέξεις. Παρατηρείται πως τα άρθρα έχουν κατά μέσο όρο μήκος 387 λέξεις και το άρθρο με το μεγαλύτερο μήκος περιέχει 5.173 λέξεις. Από τα 21.417 αληθή δείγματα μόλις τα 15.205 περιέχουν λιγότερες από 480 λέξεις. Οπότε και στις δύο περιπτώσεις το μήκος του περιεχομένου των άρθρων θα αποτελέσει πρόβλημα για τη διαδικασία της

κωδικοποίησης. Το συγκεκριμένο σύνολο δεδομένων δεν παρέχεται σε διακριτά τμήματα (training set, testing set, validation set) από τους δημιουργούς του, οπότε θα συνδυαστούν τα δύο επιμέρους σύνολα που αναφέρθηκαν έως τώρα και στη συνέχεια με αναλογία 20/80 θα δημιουργηθεί το σύνολο ελέγχου (testing data) και εκπαίδευσης (training data) αντίστοιχα. Το τελικό σύνολο δεδομένων που προέκυψε από το συνδυασμό των επιμέρους τμημάτων έχει δείγματα με χαρακτηριστικά τον τίτλο, το περιεχόμενο του άρθρου και την ετικέτα της κατηγορίας που ανήκουν. Σε μια προσπάθεια να εξακριβωθεί η υπόθεση πως τα fake news αφορούν συνήθως την πολιτική και εκμεταλλεύονται αρνητικές έννοιες για την αποτελεσματικότητά τους θα παρουσιαστούν τα word clouds των ψευδών τίτλων και άρθρων.



Εικόνα 12 - Word cloud των τίτλων των ψευδών δειγμάτων του ISOT.



Εικόνα 13- Word cloud των άρθρων των ψευδών δειγμάτων του ISOT.

Οι λέξεις που συναντώνται συχνά στους τίτλους και στο περιεχόμενο των άρθρων, ξεκάθαρα αφορούν την πολιτική αφού πρόκειται για ονόματα όπως “Trump”, ”Clinton”, ”Obama”. Ακόμα, εμφανίζονται και έννοιες οι οποίες είναι φανερό πως μεταδίδουν μια αρνητική χροιά στον αποδέκτη όπως “racist”, ”lie”, ”black” και “lives matter”. Οπότε πρόκειται για περιεχόμενο που μπορεί να θεωρηθεί πως πληροί τις προϋποθέσεις του πρώτου κεφαλαίου για να κατηγοριοποιηθεί ως ψευδές.

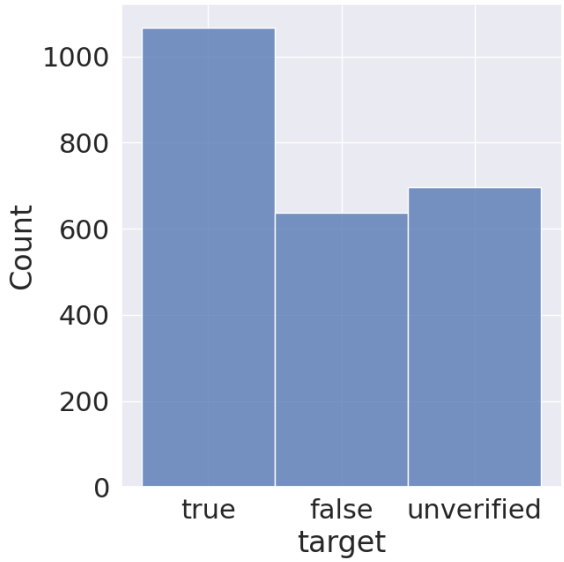
3.2.2 PHEME

Το συγκεκριμένο σύνολο δεδομένων προέκυψε ως αποτέλεσμα της μελέτης με τίτλο All-in-one: Multi-task Learning for Rumour Verification (Kochkina, Liakata, & Arkaitz, 2018), η πρότερη μορφή του εμπλουτίστηκε με σκοπό την διάκριση δειγμάτων που αποτελούν φήμες. Το περιεχόμενο των δημοσιεύσεων αφορά 9 διαφορετικά περιστατικά της επικαιρότητας (Kochkina, Liakata, & Arkaitz, 2018). Τα συνολικά δείγματα που εμφανίζονται είναι 2.402. Σε καθένα από αυτά περιέχονται διάφορα χαρακτηριστικά όπως το κείμενο της δημοσίευσης, λεπτομέρειες της αλληλεπίδρασης του κοινού μαζί της, λεπτομέρειες για τον ίδιο το δημιουργό της δημοσίευσης καθώς και την ετικέτα της κατηγορίας που ανήκει. Παρακάτω, παρουσιάζονται τα χαρακτηριστικά του κάθε δείγματος με τη μορφή και τη σειρά που εμφανίζονται στο σύνολο δεδομένων.

- **text:** Το κείμενο της δημοσίευσης.
- **date:** Ημερομηνία δημιουργίας της δημοσίευσης.
- **fav_count:** Μετρική που υποδεικνύει πόσο αρεστή είναι η δημοσίευση από τους υπόλοιπους χρήστες.
- **retweet_count:** Μετρική που υποδεικνύει τον αριθμό των αναδημοσιεύσεων του περιεχομένου από άλλους χρήστες.
- **username:** Όνομα χρήστη που δημιούργησε το δείγμα.
- **account_date:** Ημερομηνία δημιουργίας του λογαριασμού.
- **followers:** Αριθμός χρηστών που «ακολουθούν» τον λογαριασμό του δημιουργού.
- **followings:** Αριθμός χρηστών που «ακολουθεί» ο λογαριασμός του δημιουργού.
- **tweet_count:** Μετρική που δείχνει τον αριθμό των φορών που έχει εμφανιστεί μια δημοσίευση τόσο στο δημιουργό όσο και σε απλούς χρήστες.
- **protected:** Ο συγκεκριμένος χαρακτηρισμός δηλώνει ποιοι χρήστες μπορούν να δουν την αντίστοιχη δημοσίευση. Μόνο οι ακόλουθοι ενός δημιουργού μπορούν να διαβάσουν μια προστατευμένη δημοσίευση.
- **verified:** Ο συγκεκριμένος χαρακτηρισμός αφορά τον εκάστοτε δημιουργό περιεχομένου. Οι δημιουργοί που χαρακτηρίζονται ως verified δηλώνουν πως είναι αυθεντικοί. Για παράδειγμα ο λογαριασμός ενός δημοσιογραφικού πρακτορείου είναι πιστοποιημένος σε αντίθεση με άλλους χρήστες που προσπαθούν να τον μιμηθούν υιοθετώντας την ίδια ονομασία.
- **no_hashtags:** Αριθμός των hashtags[4] στη δημοσίευση. Πρόκειται για ετικέτες που εμφανίζονται εμβόλιμα στο κείμενο και συμβάλλουν στην κατηγοριοποίησή του ως προς τη θεματολογία που απασχολεί.
- **urls:** Σύνδεσμοι που περιέχονται στο δείγμα υπό μελέτη.
- **event:** Το συγκεκριμένο χαρακτηριστικό υποδεικνύει το περιστατικό που αφορά το εκάστοτε δείγμα.
- **target:** Χαρακτηρισμός του δείγματος σε μια από τις βαθμίδες αξιοπιστίας του συνόλου δεδομένων.

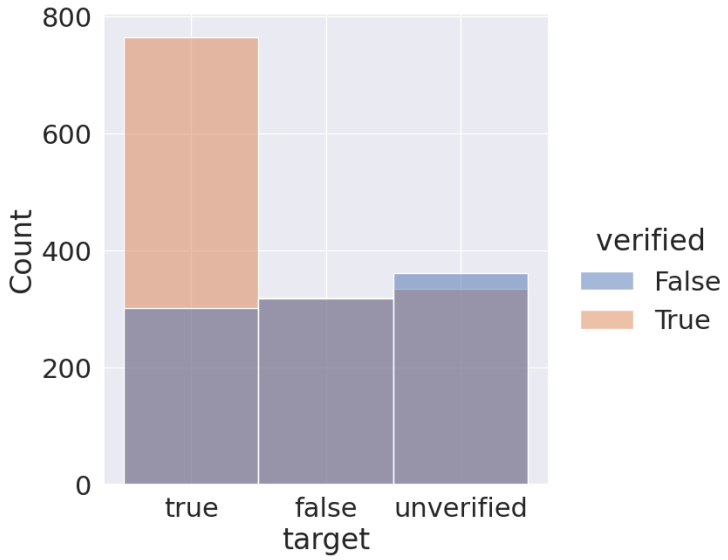
[4] https://help.twitter.com/en/using-twitter/how-to-use-hashtags

Υπάρχουν δείγματα τα οποία δεν έχουν κάποια ετικέτα και συνεπώς αφαιρούνται. Οι πιθανοί χαρακτηρισμοί των δειγμάτων είναι ψευδές (false), αληθές (true) και μη εξακριβωμένο (unverified).



Εικόνα 14 - Κατανομή των δειγμάτων ως προς τον χαρακτηρισμό τους.

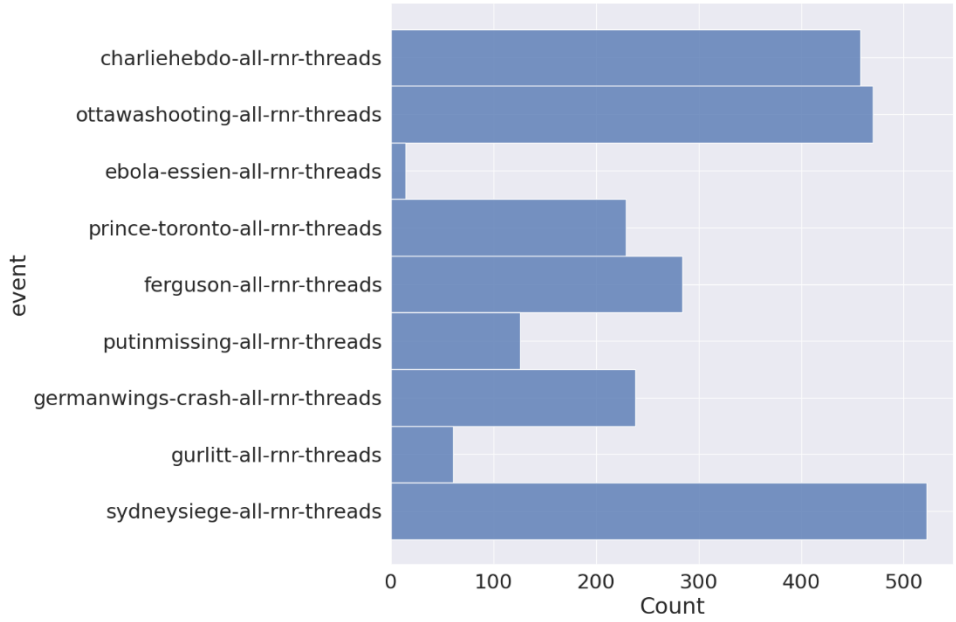
Παρατηρείται πως η κατανομή των δειγμάτων σε κατηγορίες δεν παρουσιάζει ιδιαίτερα έντονες ανισότητες. Η εφαρμογή της μεθοδολογίας που θα προταθεί στη συνέχεια θα εφαρμοστεί τόσο για την περίπτωση που συμπεριλαμβάνεται η κατηγορία μη επιβεβαιωμένων δειγμάτων όσο και για την περίπτωση που απουσιάζουν. Μια ακόμα απεικόνιση που παρουσιάζει ενδιαφέρον είναι η κατανομή των δεδομένων σε κατηγορίες, συνδυαστικά με το κύρος του λογαριασμού που έκανε τη δημοσίευση.



Εικόνα 15 - Κατανομή των δειγμάτων σε κατηγορίες συνδυαστικά με το κύρος του δημιουργού.

Με βάση την παραπάνω απεικόνιση μπορούν να εξαχθούν δύο πολύ σημαντικά συμπεράσματα. Στο Twitter υπάρχουν εξακριβωμένοι χρήστες που τους έχει αποδοθεί ο χαρακτηρισμός verified, που υποδηλώνει την αυθεντικότητα του

λογαριασμού. Η πλειοψηφία των δειγμάτων που είναι αληθή προέρχονται από τέτοιου είδους χρήστες. Η δεύτερη παρατήρηση αφορά τα ψευδή και μη εξακριβωμένα δείγματα. Παρατηρείται πως η πλειοψηφία δειγμάτων αυτών των κατηγοριών προέρχεται από χρήστες ο οποίοι δεν είναι χαρακτηρισμένοι ως verified.



Εικόνα 16 - Κατανομή των δειγμάτων ως προς το πλαίσιο στο οποίο αναφέρονται.

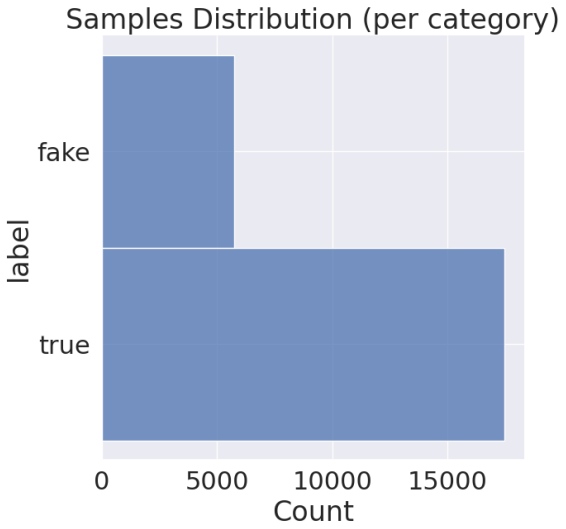
Είναι προφανές πως δεν υπάρχει ο ίδιος βαθμός κάλυψης για καθένα από τα 9 γεγονότα. Σε ορισμένες περιπτώσεις εμφανίζονται πολύ περισσότερες δημοσιεύσεις συγκριτικά με κάποιες άλλες. Το μέσο μήκος των δημοσιεύσεων είναι 17 λέξεις ενώ η πιο εκτενής δημοσίευση περιέχει 29 λέξεις. Έπειτα από επεξεργασία το σύνολο δεδομένων αποτελείται από δύο στήλες. Η μια στήλη περιέχει το κείμενο των δημοσιεύσεων και η δεύτερη στήλη το χαρακτηρισμό τους σε μια από τις τρεις κατηγορίες που έχουν αναλυθεί παραπάνω. Τέλος, παρουσιάζεται το word cloud των δημοσιεύσεων.

Εικόνα 17 - Word cloud των δημοσιεύσεων του συνόλου δεδομένων PHEME.

Συγκριτικά με το προηγούμενο σύνολο δεδομένων είναι πολύ πιο έντονη η χρήση αρνητικών όρων που σε πολλές περιπτώσεις παρατηρείται πως εκφράζουν βία (όπως για παράδειγμα ‘war memorial”). Υπάρχουν όροι που αναφέρονται σε πυροβολισμούς και τραυματισμούς από πυροβολισμό όπως “soldier shot”, “Ottawa shooting”, “shooting”, “shots fired”. Το μοτίβο αυτό αν και δεν υπονοεί τόσο έντονα κάποια σύνδεση με την πολιτική, όπως συμβαίνει με άλλα σύνολα δεδομένων, είναι φανερό πως προωθεί έντονα αρνητικές έννοιες.

3.2.3 FakeNewsNet

Το παρόν σύνολο δεδομένων δημοσιεύθηκε το 2018 σε μια προσπάθεια των δημιουργών του να υπάρξει ένα σύνολο δεδομένων το οποίο να συνδυάζει το περιεχόμενο των ψευδών νέων, το κοινωνικό πλαίσιο στο οποίο εμφανίζεται η πληροφορία και τέλος τις δυναμικές πληροφορίες ανά δείγμα (Shu K. , Mahudeswaran, Wang, Lee, & Liu, 2020). Οι πληροφορίες είναι επικυρωμένες από το PolitiFact.com και το GossipCop.com. Η θεματολογία των δειγμάτων αφορά την πολιτική αλλά και δημοφιλή πρόσωπα της επικαιρότητας. Το βασικό τμήμα του συνόλου δεδομένων αποτελείται από δύο υποσύνολα, ένα με αληθή (17.441 δείγματα) και ένα με ψευδή (5.755 δείγματα) δεδομένα. Όπως φαίνεται από τα προηγούμενα στοιχεία θα πρέπει να προβλεφθεί η ανισορροπία του συνόλου δεδομένων για να προκύψουν στη συνέχεια όσο το δυνατό καλύτερα αποτελέσματα.



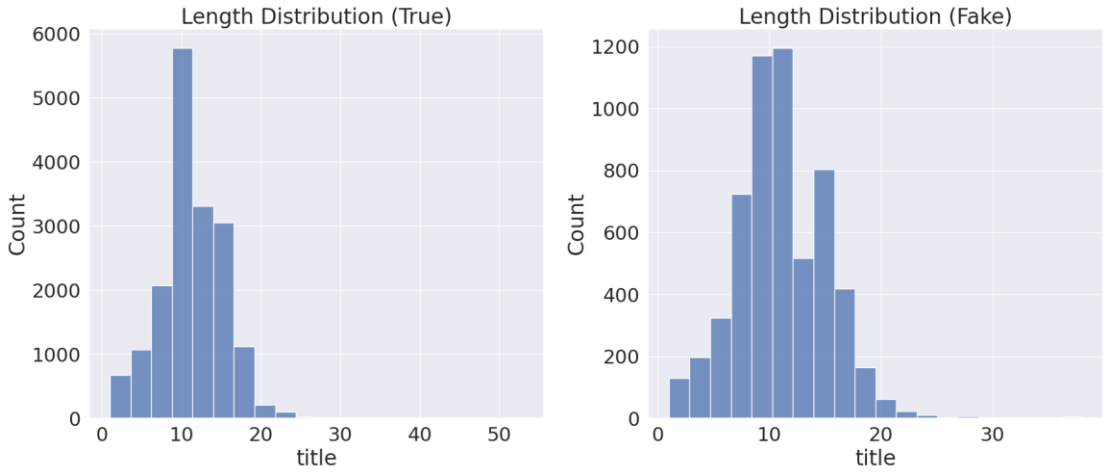
Εικόνα 18 - Κατανομή των δειγμάτων σε κατηγορίες ως προς τον χαρακτηρισμό τους.

Το κάθε δείγμα αποτελείται από ένα αναγνωριστικό, τον σύνδεσμο όπου περιέχεται το πλήρες άρθρο, τον τίτλο του άρθρου και τα αναγνωριστικά των tweets που αναφέρονται σε αυτό. Παρακάτω, παρουσιάζονται τα χαρακτηριστικά των δειγμάτων με τη σειρά και τον τρόπο που εμφανίζονται στο σύνολο δεδομένων.

- **id:** Αναγνωριστικό του εκάστοτε δείγματος.
- **news_url:** Σύνδεσμος προς το πλήρες άρθρο.
- **title:** Τίτλος του άρθρου.

- **tweet_ids:** Αναγνωριστικά των λογαριασμών του Twitter που αλληλοεπίδρασαν με τη δημοσίευση.

Αν το επιθυμεί, κάποιος μελετητής δύναται να αναλύσει επί πληρωμή μέσω της διεπαφής προγραμματισμού εφαρμογών (API) που προσφέρει το Twitter, μεταδεδομένα των λογαριασμών που αλληλοεπίδρασαν με το εκάστοτε δείγμα (Shu K. , Mahudeswaran, Wang, Lee, & Liu, 2020). Ωστόσο, στο πλαίσιο της συγκεκριμένης εργασίας χρησιμοποιούνται τεχνικές που βασίζονται στην ανάλυση του συγγραφικού ύφους (style-based) και όχι σε άλλες πτυχές των δειγμάτων όπως έχουν αναφερθεί στο πρώτο κεφάλαιο. Για αυτό το λόγο, το συγκεκριμένο σύνολο δεδομένων θα αναλυθεί ως προς τους τίτλους που δίνονται, για να εξεταστεί η ικανότητα της μεθόδου που θα παρουσιαστεί στη συνέχεια σε μικρού μήκους κείμενα. Κάθε δείγμα τελικά θα έχει σαν χαρακτηριστικά το κείμενο του τίτλου και την ετικέτα του δείγματος. Τα δύο υποσύνολα θα συγχωνευτούν προτού εφαρμοστεί η μεθοδολογία ταξινόμησης. Στη συνέχεια παρουσιάζονται οι κατανομές των μηκών για καθένα από τα δύο υποσύνολα.



Εικόνα 19 - Κατανομές των μηκών των δύο υποσυνόλων του FakeNewsNet.

Με βάση τις παραπάνω κατανομές γίνεται κατανοητό πως δεν θα υπάρξει πρόβλημα κατά τη διαδικασία της κωδικοποίησης. Αναφορικά με το υποσύνολο που περιέχει αληθή δείγματα το μέσο μήκος των τίτλων είναι 12 λέξεις και ο μεγαλύτερος τίτλος αποτελείται από 53 λέξεις. Παράλληλα τα ψευδή δείγματα έχουν μέσο μήκος τίτλου 12 λέξεις και ο μεγαλύτερος τίτλος έχει μήκος 38 λέξεις. Στη συνέχεια παρουσιάζεται το word cloud του ψευδούς υποσυνόλου για να παρατηρηθούν οι δημοφιλέστεροι του όροι.

Εικόνα 20 - Word cloud του ψευδούς υποσυνόλου του FakeNewsNet.

Με βάση τους όρους που παρουσιάζονται παρατηρείται πως γίνεται συχνή αναφορά σε δημοφιλή πρόσωπα και ταυτόχρονα παρουσιάζονται έννοιες με αρνητική χροιά. Κάποιες από αυτές τις έννοιες αναφέρονται σε χωρισμούς γνωστών προσώπων “split”, ”divorce”, ”lie”. Το λεξιλόγιο αυτό ικανοποιεί ορισμένες από τις προϋποθέσεις για να είναι ένα δείγμα ψευδές. Δηλαδή, χρησιμοποιούνται τα ονόματα δημοφιλών προσώπων για να εξάψουν την περιέργεια του αναγνώστη και ταυτόχρονα τονίζονται τα αρνητικά γεγονότα. Παρακάτω για λόγους σύγκρισης παρατίθεται και το word cloud του υποσυνόλου αληθών δειγμάτων.

Εικόνα 21 - Word cloud του αληθούς υποσυνόλου του FakeNewsNet.

Λόγω της σύγκρισης γίνεται αρκετά φανερό πως αν και υπάρχει έντονη αναφορά σε δημοφιλή πρόσωπα και στις δύο περιπτώσεις, στο υποσύνολο των αληθών δειγμάτων απουσιάζει η εμφάνιση εννοιών με αρνητική βαρύτητα (π.χ. διαζύγια).

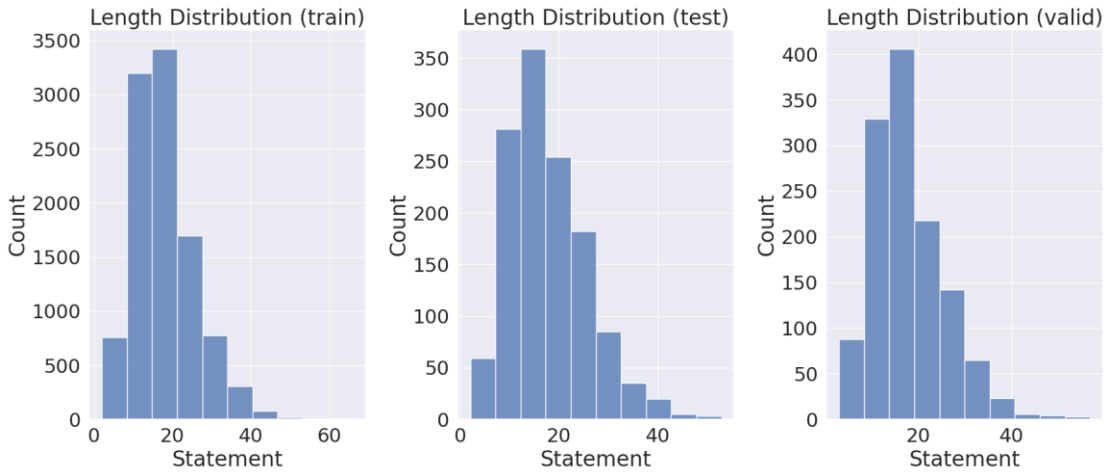
3.2.4 LIAR

Το συγκεκριμένο σύνολο δεδομένων δημοσιεύθηκε το 2017 και περιέχει συνολικά 12.800 δείγματα από ομιλίες, διάφορες δηλώσεις ή αποσπάσματα συνεντεύξεων (William Yang, 2017). Τα δείγματα είναι χωρισμένα σε 6 διαφορετικές κατηγορίες. Συλλέχθηκαν από την ιστοσελίδα PolitiFact.com σε διαφορετικές περιστάσεις και έχουν χαρακτηριστεί από τους συντάκτες της χειροκίνητα (William Yang, 2017). Το σύνολο δεδομένων χωρίζεται σε τρία τμήματα τα οποία έχουν διαφορετικό σκοπό. Το πρώτο υποσύνολο περιέχει δεδομένα τα οποία θα χρησιμοποιηθούν κατά τη διαδικασία εκπαίδευσης των μοντέλων μάθησης (training partition), το δεύτερο υποσύνολο αφορά την επαλήθευση των αποτελεσμάτων μετά από κάθε εποχή εκπαίδευσης του εκάστοτε μοντέλου (validation partition) και τέλος το τρίτο υποσύνολο περιέχει δεδομένα που αξιοποιούνται για την εξακρίβωση της αποτελεσματικότητας της μεθοδολογίας ταξινόμησης (testing partition).

Το κάθε δείγμα περιέχει τα εξής χαρακτηριστικά (William Yang, 2017):

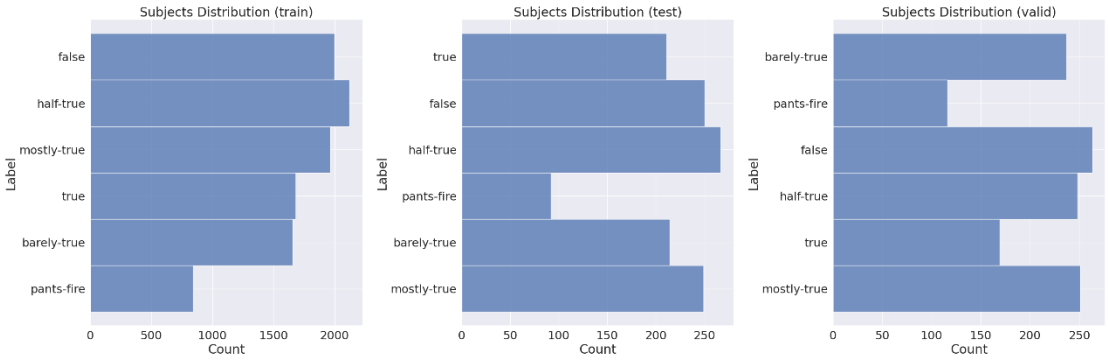
- **ID:** Το αναγνωριστικό του δείγματος.
- **Label:** Η τελική κατηγορία στην οποία ταξινομήθηκε το δείγμα ανάλογα με τις αξιολογήσεις.
- **Statement:** Το κείμενο της δήλωσης που θα χρησιμοποιηθεί.
- **Subject:** Η θεματική με την οποία σχετίζεται η δήλωση.
- **Speaker:** Ο ομιλητής ή το μέσο διάδοσης του δείγματος.
- **Speaker's Job:** Το είδος της απασχόλησης του ομιλητή.
- **State:** Η πολιτεία των Η.Π.Α. στην οποία πραγματοποιήθηκε η δήλωση.
- **Party:** Το τμήμα του κοινοβουλευτικού τόξου στο οποίο εντάσσεται ο ομιλητής.
- **Barely True Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **False Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **Half True Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **Mostly True Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **Pants Fire Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **True Count:** Ψήφοι που εντάσσουν το κείμενο στη συγκεκριμένη βαθμίδα αξιοπιστίας.
- **Context:** Το πλαίσιο με βάση το οποίο δημιουργήθηκε το εκάστοτε δείγμα.

Παρακάτω παρουσιάζονται οι κατανομές των μηκών των δειγμάτων για κάθε ξεχωριστό υποσύνολο. Είναι προφανές πως δεν θα υπάρξει κάποιο σημαντικό πρόβλημα κατά τη διαδικασία της κωδικοποίησης αφού τα μήκη δεν είναι ιδιαιτέρως μεγάλα.



Εικόνα 22 - Κατανομή των μηκών των διαφορετικών υποσυνόλων του συνόλου δεδομένων LIAR.

Το υποσύνολο που χρησιμοποιείται κατά τη διαδικασία εκπαίδευσης έχει δείγματα με μέσο όρο μήκους 18 λέξεων και το μεγαλύτερο έχει μήκος 66 λέξεις. Το υποσύνολο που χρησιμεύει στην επαλήθευση των αποτελεσμάτων ανά εποχή εκπαίδευσης έχει δείγματα με μέσο όρο μήκους 18 λέξεις και το πιο εκτενές έχει μήκος 57 λέξεων. Το μέσο μήκος των δειγμάτων του τελευταίου υποσυνόλου είναι 18 λέξεις και παράλληλα το μέγιστο μήκος δείγματος είναι ίσο με 53 λέξεις. Στη συνέχεια παρουσιάζονται οι κατανομές των δειγμάτων στις 6 βαθμίδες αξιοπιστίας του συνόλου δεδομένων.



Εικόνα 23 - Κατανομή των δειγμάτων του συνόλου δεδομένων LIAR σε κατηγορίες ως προς την αξιοπιστία τους.

Δεν πρόκειται για ένα ιδιαίτερα ισορροπημένο σύνολο δεδομένων γεγονός το οποίο θα πρέπει να προβλεφθεί κατά την πειραματική διαδικασία. Επίσης, η μεθοδολογία ταξινόμησης δειγμάτων θα πρέπει να εφαρμοστεί δύο φορές τόσο για την περίπτωση όπου υπάρχουν 6 βαθμίδες αξιοπιστίας όσο και για την περίπτωση όπου γίνεται διαχωρισμός μόνο σε αληθή και ψευδή δείγματα. Για να πραγματοποιηθεί ο συγκεκριμένος διαχωρισμός τα δείγματα που ανήκουν στις δύο υψηλότερες βαθμίδες αξιοπιστίας θα χαρακτηριστούν ως αληθή και όλα τα υπόλοιπα ως ψευδή.

3.2.5 FakeNewsChallenge

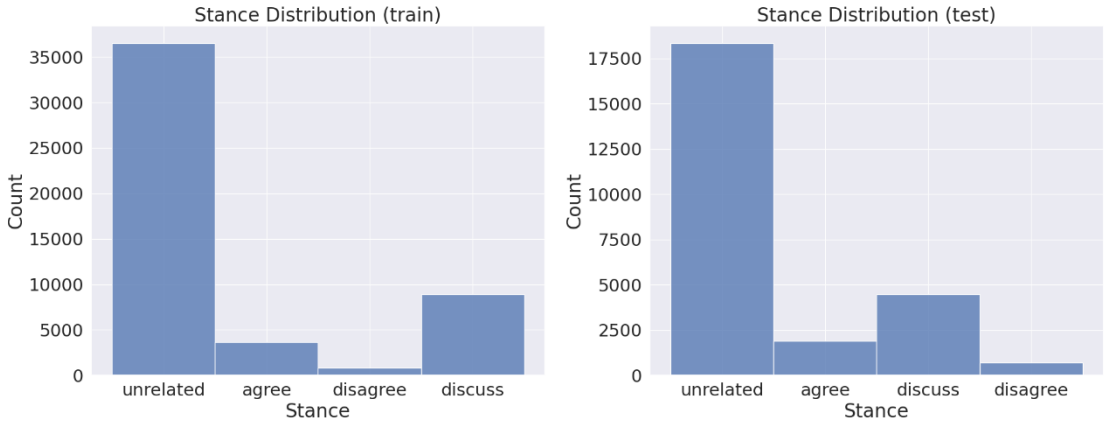
Το συγκεκριμένο σύνολο δεδομένων προέκυψε από την πρωτοβουλία Fake News Challenge[5] το 2017, που σαν σκοπό είχε να ενθαρρύνει τη δημιουργία συστημάτων ταξινόμησης περιεχομένου βασισμένων στη χρήση μηχανικής μάθησης. Κάθε δείγμα χαρακτηρίζεται από ένα αναγνωριστικό, τον τίτλο, την κατηγορία στην οποία ανήκει καθώς και το ίδιο το άρθρο. Παρακάτω, θα παρουσιαστούν τα χαρακτηριστικά ανά δείγμα με τη σειρά και τη μορφή που εμφανίζονται στο σύνολο δεδομένων.

- **Headline:** Η επικεφαλίδα του άρθρου στο υπό μελέτη δείγμα.
- **Stance:** Ο χαρακτηρισμός του δείγματος ως προς την αξιοπιστία του.
- **BodyID:** Το αναγνωριστικό του άρθρου στο σύνολο δεδομένων.
- **articleBody:** Το περιεχόμενο του άρθρου.

Το σύνολο δεδομένων αποτελείται από δύο υποσύνολα δειγμάτων ένα που θα χρησιμοποιηθεί για την εκπαίδευση (49.972 δείγματα) και ένα που θα χρησιμοποιηθεί για τον έλεγχο απόδοσης των μοντέλων (25.413 δείγματα). Οι κατηγορίες στις οποίες χωρίζονται τα δείγματα είναι οι ακόλουθες:

- **Disagree:** Το κείμενο που εμφανίζεται στο άρθρο διαφωνεί με τον τίτλο οπότε πρόκειται για fake news.
- **Agree:** Το κείμενο που εμφανίζεται στο άρθρο συμφωνεί με τον τίτλο.
- **Discuss:** Το κείμενο που εμφανίζεται στο άρθρο αποτελεί μια συζήτηση του τίτλου.
- **Unrelated:** Το κείμενο που εμφανίζεται στο άρθρο δεν εμφανίζει καμία συνάφεια με τον τίτλο.

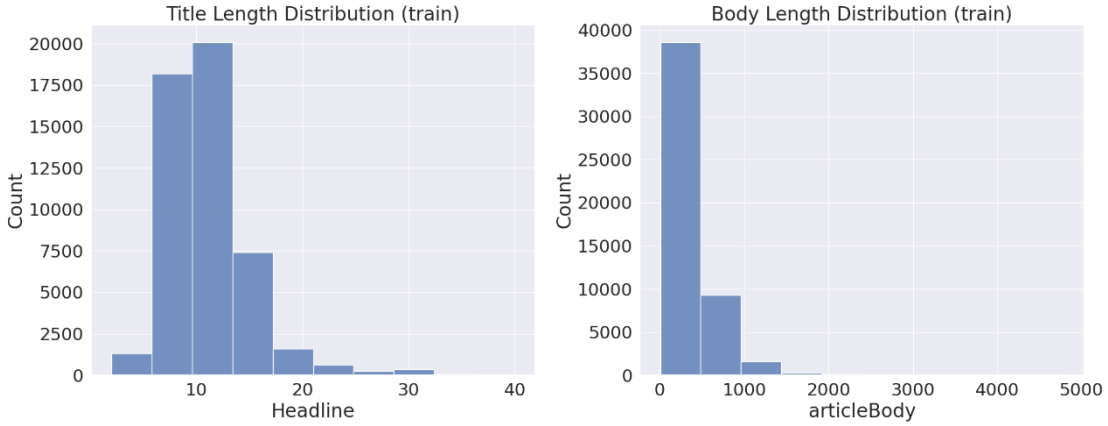
Με βάση τις παραπάνω ταξινομήσεις, ως αληθή νέα μπορούν να χαρακτηριστούν μόνο εκείνα που εντάσσονται στην κατηγορίες agree και discuss, ενώ οι υπόλοιπες θεωρείται πως περιέχουν παραπλανητικό περιεχόμενο. Στη συνέχεια, θα παρουσιαστεί η κατανομή των δειγμάτων στα δύο υποσύνολα ως προς τις κατηγορίες στις οποίες εντάσσονται.



Εικόνα 24 - Κατανομή των δειγμάτων του κάθε υποσυνόλου σε κατηγορίες.

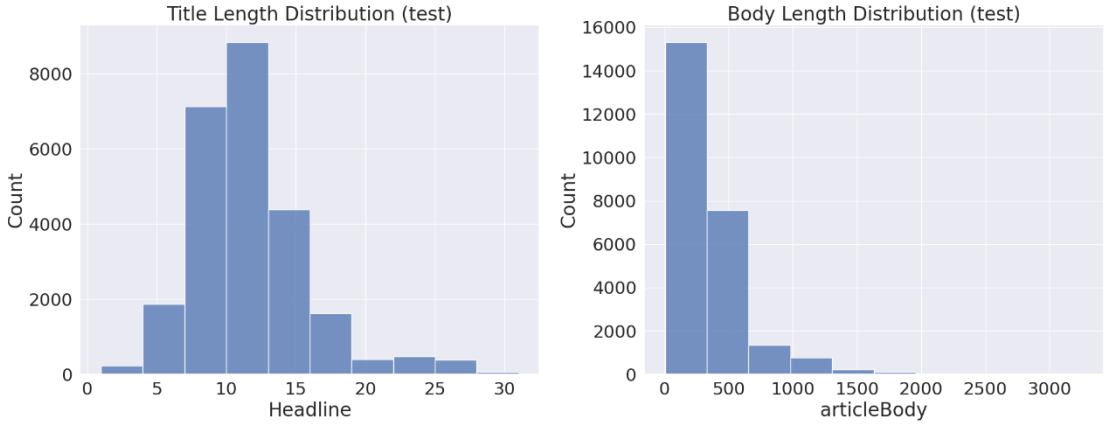
[5] http://www.fakenewschallenge.org

Είναι προφανής η μη ομοιόμορφη κατανομή των δειγμάτων. Το γεγονός αυτό σημαίνει πως πρόκειται για ένα σύνολο δεδομένων με πάρα πολύ έντονη ανισορροπία που είναι πιθανό να προκαλέσει μείωση στην τελική απόδοση. Στη συνέχεια, θα πρέπει να εξεταστεί και το πλήθος των λέξεων των δειγμάτων τόσο στους τίτλους όσο και στο ίδιο το περιεχόμενο των άρθρων για κάθε υποσύνολο. Αρχικά, θα παρουσιαστούν τα μήκη λέξεων των δειγμάτων του υποσυνόλου που χρησιμοποιείται κατά την εκπαίδευση. Το μέσο μήκος των τίτλων είναι 12 λέξεις και ο μεγαλύτερος τίτλος έχει μήκος 40 λέξεις. Το μεγαλύτερο άρθρο του υποσυνόλου έχει μήκος 4.788 λέξεις και κατά μέσο όρο το μήκος των άρθρων είναι 370 λέξεις.



Εικόνα 25 - Κατανομή των μηκών των τίτλων και των άρθρων του υποσυνόλου εκπαίδευσης.

Έπειτα παρουσιάζεται η αντίστοιχη απεικόνιση και για τα δείγματα του συνόλου επαλήθευσης. Ο τίτλος με το μέγιστο μήκος περιέχει 31 λέξεις και κατά μέσο όρο οι τίτλοι έχουν μήκος 12 λέξεις. Τα άρθρα έχουν κατά μέσο όρο μήκος 348 λέξεις και το μεγαλύτερο άρθρο περιέχει συνολικά 3.257 λέξεις.



Εικόνα 26 - Κατανομή των μηκών των τίτλων και των άρθρων του υποσυνόλου επαλήθευσης.

Είναι φανερό πως τα άρθρα έχουν ιδιαίτερα μεγάλο μήκος το οποίο δυσκολεύει την διαδικασία κωδικοποίησης. Το μοντέλο μάθησης DistilBERT, που θα χρησιμοποιηθεί για την κωδικοποίηση, μπορεί να δεχθεί μέχρι 512 λεκτικές μονάδες (tokens). Συνεπώς, χρειάζεται να υπάρξει πρόβλεψη στη μεθοδολογία του πειράματος για αυτή την ιδιαιτερότητα.

Πίνακας 2 - Συνοπτική περιγραφή των συνόλων δεδομένων

	Είδος περιεχομένου	**Συνολικά δείγματα**	**Βαθμίδες αξιοπιστίας**
ISOT	Τίτλοι - Άρθρα	44.898	2
PHEME	Δημοσιεύσεις Twitter	2.402	3
FakeNewsNet	Τίτλοι	23.196	2
LIAR	Σύντομες δηλώσεις	12.800	6
FakeNewsChallenge	Τίτλοι - Άρθρα	75.385	4

4 Περιγραφή μεθοδολογίας

4.1 Παρουσίαση χρήσιμων βιβλιοθηκών

Στο παρόν κεφάλαιο, θα γίνει αναφορά στην αξιοποίηση παραδοσιακών τεχνικών καθώς και περισσότερο σύγχρονων μεθόδων μάθησης για την ταξινόμηση περιεχομένου ως προς την εγκυρότητά του. Ακόμα, θα παρουσιαστεί η πειραματική προσέγγιση της εργασίας και ο τρόπος εφαρμογής της ανά διαφορετικό σύνολο δεδομένων. Τέλος, θα παρουσιαστούν με συντομία και θα συγκριθούν μεταξύ τους οι αρχιτεκτονικές που χρησιμοποιήθηκαν. Ειδικότερα, θα αναλυθεί η δομή τους και γενικότερα ο τρόπος με τον οποίο οργανώνονται τα επιμέρους στρώματα μαζί με τις αντίστοιχες παραμέτρους τους.

Η ανάλυση των συνόλων δεδομένων και η εκτέλεση της πειραματικής διαδικασίας πραγματοποιήθηκαν στο προγραμματιστικό περιβάλλον που παρέχεται από την πλατφόρμα Google Colab. Στις επόμενες ενότητες εκτός από τη σύντομη περιγραφή της πλατφόρμας θα γίνει αναφορά και σε κάποιες από τις βιβλιοθήκες που αξιοποιήθηκαν στην πειραματική διαδικασία. Από το σύνολο των εργαλείων που χρησιμοποιήθηκαν θα παρουσιαστούν οι πιο κρίσιμες βιβλιοθήκες, για σκοπούς ανάλυσης δεδομένων και δημιουργίας νευρωνικών δικτύων βασισμένων σε εμφυτεύματα που έχουν προκύψει από μοντέλα μετασχηματιστών (transformers).

Συγκεκριμένα, ήταν κρίσιμο να αξιοποιηθούν βιβλιοθήκες οι οποίες προσφέρουν τις επιμέρους δομές που συνθέτουν τις αρχιτεκτονικές και παρέχουν μεγάλο περιθώριο παραμετροποίησης προσφέροντας σημαντικές δυνατότητες προς τη βελτιστοποίηση του τελικού αποτελέσματος. Ακόμα, ήταν απαραίτητη η χρήση κάποιας βιβλιοθήκης η οποία να προσφέρει τις υλοποιήσεις των μηχανισμών που αναφέρθηκαν στο δεύτερο κεφάλαιο για την κωδικοποίηση των δειγμάτων, όπως είναι ο μηχανισμός της διαδικασίας tokenization καθώς και το μοντέλο DistilBERT.

4.1.1 Google Colab

Πρόκειται για ένα προϊόν από την εταιρεία Google. Η πλατφόρμα αυτή είναι προσβάσιμη από περιηγητές ιστού και επιτρέπει στους χρήστες να γράψουν και να εκτελέσουν κώδικα χρησιμοποιώντας τη γλώσσα Python. Χρησιμοποιείται κυρίως στους τομείς της μηχανικής μάθησης, της ανάλυσης δεδομένων και της εκπαίδευσης όπως αναφέρεται από την ίδια την εταιρεία[6] που δημιούργησε την πλατφόρμα. Στην πραγματικότητα, πρόκειται για μια υπηρεσία η οποία προσφέρει τη φιλοξενία αρχείων κατηγορίας Jupyter notebook (έχουν κατάληξη .ipynb) και δεν απαιτεί καμία επιπλέον διαδικασία παραμετροποίησης για την εκτέλεσή τους. Ακόμα, προσφέρεται η πρόσβαση σε υπολογιστικούς πόρους όπως μονάδες επεξεργασίας γραφικών (GPUs), απλές μονάδες επεξεργασίας (CPUs) και μονάδες επεξεργασίας τανυστών (TPUs).

[6] https://research.google.com/colaboratory/faq.html

Η πλατφόρμα έχει βασίσει τη δημιουργία της στο έργο ανοικτού κώδικα με την ονομασία Project Jupyter και στοχεύει στις ανάγκες που είχε σκοπό να ικανοποιήσει. Η πρώτη ανάγκη που εξυπηρετεί το Google Colab είναι πως προσφέρει τη δυνατότητα παρουσίασης μιας ομογενούς υπολογιστικής λύσης σε ένα πρόβλημα, με μορφή που επιτρέπει την παρουσίαση συμπερασμάτων σε ένα μεγάλο εύρος διαφορετικών περιστάσεων και κοινών (Perez & Granger, 2015). Το δεύτερο σημαντικό πλεονέκτημα που προσφέρει είναι το γεγονός πως η εκάστοτε μελέτη είναι εύκολο να αναπαραχθεί από επόμενους ερευνητές αν ακολουθηθεί επακριβώς η ίδια πειραματική διαδικασία (Perez & Granger, 2015). Γεγονός που διευκολύνει τους μελετητές σε τομείς όπως η ανάλυση δεδομένων. Τέλος, επιτρέπει την συνεργασία μεταξύ πολλαπλών μελετητών αφού κάθε έργο μπορεί να εξαχθεί σε διαφορετικούς τύπους αρχείων και να διαμοιραστεί σε όλους τους ενδιαφερόμενους (Perez & Granger, 2015).

4.1.2 NumPy

Η συγκεκριμένη βιβλιοθήκη ανοικτού κώδικα της γλώσσας Python χρησιμοποιείται σχεδόν σε κάθε επιστημονικό πεδίο. Αυτό συμβαίνει γιατί θεωρείται από τις πιο διαδεδομένες μεθόδους διαχείρισης αριθμητικών δεδομένων στη συγκεκριμένη γλώσσα προγραμματισμού[7]. Η βιβλιοθήκη NumPy μπορεί να χρησιμοποιηθεί από άπειρους προγραμματιστές έως και πιο έμπειρους χρήστες στο πλαίσιο της επιστημονικής έρευνας ή της βιομηχανικής έρευνας και ανάπτυξης. Η διεπαφή προγραμματισμού εφαρμογών (API) της βιβλιοθήκης χρησιμοποιείται εκτενώς για να παραχθούν αποτελέσματα άλλων δημοφιλών βιβλιοθηκών στην Python όπως Pandas, Matplotlib, SciPy, scikit-image, scikit-learn και επιπρόσθετα εργαλεία συναφή με την ανάλυση δεδομένων. Η βιβλιοθήκη NumPy ήταν σημαντικό τμήμα του λογισμικού της μελέτης για την ανακάλυψη μαγνητικών κυμάτων και της αρχικής απεικόνισης της μαύρης τρύπας (Harris, Millman, & van der Walt, 2020).

Τα δεδομένα που αφορούν τη βιβλιοθήκη αυτή επεξεργάζονται και αποθηκεύονται ως στοιχεία σε δομές πινάκων. Οι πίνακες αυτοί είναι μονοδιάστατοι ή και πολυδιάστατοι και στα στοιχεία τους μπορούν να εφαρμοστούν με ταχύτητα και σχετικά υψηλή απόδοση διάφορες διαδικασίες επεξεργασίας δεδομένων. Ενδεικτικά, κάποιες από αυτές αφορούν μετασχηματισμούς των διαστάσεων του πίνακα, ταξινόμηση στοιχείων, επιλογή στοιχείων, βασικές εφαρμογές γραμμικής άλγεβρας, διακριτός μετασχηματισμός Fourier και πολλές ακόμα. Κάθε δομή πίνακα που χρησιμοποιείται από τη βιβλιοθήκη αποτελείται από έναν δείκτη προς τη μνήμη και κάποια μεταδεδομένα (metadata). Τα μεταδεδομένα περιέχουν κάποιες πολύ σημαντικές πληροφορίες όπως τον τύπο δεδομένων του περιεχομένου, τις διαστάσεις του πίνακα καθώς και το χώρο που καταλαμβάνει κάθε επιμέρους στοιχείο στη μνήμη (Harris, Millman, & van der Walt, 2020).

[7] https://numpy.org/doc/stable/user/absolute_beginners.html

4.1.3 Pandas

Η βιβλιοθήκη Pandas είναι ιδιαίτερα δημοφιλής και χρησιμοποιείται στη γλώσσα Python. Εμφανίζεται πολύ συχνά γιατί προσφέρει μια ευρεία ποικιλία δομών δεδομένων και εργαλεία για την επεξεργασία δομημένων συνόλων δεδομένων. Αν και ήδη προηγούνταν βιβλιοθήκες όπως η NumPy δεν υπήρχε η ίδια ευελιξία ως προς τη διαχείριση πινάκων με ανομοιογενείς τύπους δεδομένων ανά στήλη (McKinney, 2011). Παράλληλα, επιτρέπει ακόμα και τη διαχείριση δεδομένων που αφορούν χρονικά γεγονότα. Η καθιέρωση της συγκεκριμένης βιβλιοθήκης προήλθε από τη διαπίστωση πως γλώσσες βάσεων δεδομένων όπως SQL και στατιστικών αναλύσεων όπως R και SAS παρείχαν δυνατότητες που δεν εμφανίζονταν στην Python (McKinney, 2011). Ειδικότερα, ανάμεσα στις πολλές λειτουργίες της, η βιβλιοθήκη Pandas επιτρέπει τη συνένωση, το διαχωρισμό, το φιλτράρισμα και την ομαδοποίηση των περιεχομένων στους πίνακες δεδομένων.

Οι πιο σημαντικές ιδιότητες που παρέχει η βιβλιοθήκη Pandas είναι η διαχείριση τιμών που απουσιάζουν ή δεν είναι έγκυρες και η χρήση ιεραρχικού ευρετηρίου (hierarchical indexing). Τέλος, αξίζει να αναφερθεί πως προσφέρει πρόσβαση σε δείγματα του συνόλου δεδομένων ακόμα και με χρήση των ετικετών. Το μειονέκτημα της βιβλιοθήκης Pandas είναι πως αν και προσφέρει πλήθος δυνατοτήτων, υπάρχουν περιπτώσεις στις οποίες δεν μπορεί να ολοκληρώσει εργασίες αρκετά αποδοτικά και γρήγορα, οπότε λειτουργεί σε συνεργασία με τη βιβλιοθήκη NumPy (McKinney, 2011).

4.1.4 Seaborn

Η βιβλιοθήκη Seaborn συχνά χρησιμοποιείται στη γλώσσα Python για την απεικόνιση δεδομένων. Η απεικόνιση των δεδομένων είναι πολύ βοηθητική για την κατανόηση των αποτελεσμάτων της ανάλυσης τους και την ευκολότερη μετάδοση των συμπερασμάτων σε τρίτους. Ο λόγος ύπαρξης της συγκεκριμένης βιβλιοθήκης είναι να προσφέρει έναν εύκολο τρόπο για χρήση της matplotlib η οποία αναπτύσσεται τις τελευταίες δύο δεκαετίες και είναι δύσκολο να αξιοποιηθεί λόγω της μεγάλης πολυπλοκότητας και ποικιλίας των δυνατοτήτων που προσφέρει (Waskom, 2021). Η βιβλιοθήκη είναι ιδιαίτερα ευέλικτη, προσφέρει μεγάλο βαθμό ελέγχου της εμφάνισης του τελικού αποτελέσματος και απλοποιεί τις διαδικασίες απεικόνισης δεδομένων.

Η βιβλιοθήκη Seaborn αποδίδει με ικανοποιητικό τρόπο τις απεικονίσεις μιας μεγάλης ποικιλίας συνόλων δεδομένων τα οποία μπορούν να παρουσιαστούν σε μορφή πίνακα (Waskom, 2021). Συγκεκριμένα, μια ιδιαίτερα βοηθητική δυνατότητα είναι η άμεση αντιστοίχιση μεταβλητών του συνόλου δεδομένων σε στοιχεία του εκάστοτε γραφήματος. Ακόμα, η ποσοτική ή ποιοτική αντιστοίχιση στοιχείων απεικόνισης διαφοροποιείται αυτόματα, ανάλογα με την ύπαρξη αριθμητικών ή κατηγορικών δεδομένων. Επιπλέον, η βιβλιοθήκη Seaborn πριν τη δημιουργία των εκάστοτε γραφημάτων μπορεί να εφαρμόσει μετασχηματισμούς στα δεδομένα και να

παρουσιάσει τα αποτελέσματα καθώς και τα περιθώρια σφάλματος των εκτιμήσεων που απεικονίζει (Waskom, 2021).

4.1.5 TensorFlow

Πρόκειται για μια ολοκληρωμένη πλατφόρμα μηχανικής μάθησης ανοικτού κώδικα. Προσφέρει ένα ευέλικτο οικοσύστημα εργαλείων, βιβλιοθηκών αλλά και επιπλέον πόρων, που μπορεί να προκύψουν από το σύνολο του έργου των χρηστών που αξιοποιούν τις δυνατότητες της πλατφόρμας για δικές τους μελέτες. Η συγκεκριμένη πλατφόρμα προσφέρει ένα πολύ σημαντικό πλεονέκτημα, οι αρχιτεκτονικές των μοντέλων και οι υπολογιστικές διαδικασίες που αναπτύσσονται μπορούν να εκτελεστούν με ελάχιστες αλλαγές σε ένα ευρύ σύνολο ετερογενών συστημάτων. Τα συστήματα αυτά μπορεί να αντιστοιχούν σε κινητές συσκευές με περιορισμένες δυνατότητες όπως έξυπνα κινητά έως και κατανεμημένα συστήματα εκατοντάδων συσκευών που αξιοποιούν χιλιάδες μονάδες επεξεργασίας γραφικών (GPUs) (Abadi, και συν., 2015). Το γεγονός αυτό απλοποιεί πολύ τη διαδικασία διάδοσης της χρήσης μεθόδων μηχανικής μάθησης, αφού πλέον δεν υφίσταται ανάγκη διαφορετικής παραμετροποίησης σε συστήματα με πληθώρα υπολογιστικών πόρων έναντι άλλων συστημάτων που μειονεκτούν (Abadi, και συν., 2015). Εκτός από το γεγονός πως προσφέρεται ευελιξία ως προς την επιλογή ανάμεσα σε μια ή σε πολλαπλές συσκευές, επιπλέον παρέχεται η δυνατότητα ακριβούς προσδιορισμού των πόρων που θα χρησιμοποιηθούν ανά συσκευή (π.χ. αριθμός μονάδων επεξεργασίας).

Η διεπαφή προγραμματισμού εφαρμογών (API) της βιβλιοθήκης TensorFlow παρέχει μια πληθώρα εργαλείων. Για παράδειγμα, μεταξύ άλλων υπάρχει η δυνατότητα χρήσης συνελικτικών στρωμάτων 1-διάστασης[8] με υψηλές δυνατότητες παραμετροποίησης. Ακόμα, παρέχεται η δυνατότητα χρήσης πλήρως συνδεδεμένων στρωμάτων[9] (dense layers) και στρωμάτων μακράς βραχύχρονης μνήμης[10] (LSTM layers). Όλα αυτά τα επιμέρους στρώματα καθώς και πολλά ακόμη μπορούν να χρησιμοποιηθούν συνδυαστικά για να συνθέσουν ένα ολοκληρωμένο μοντέλο. Ανάμεσα στις πολλές προσεγγίσεις σύνθεσης μιας αρχιτεκτονικής επιλέχθηκε η ακολουθιακή (Sequential)[11] με βάση την οποία το τελικό μοντέλο αποτυπώνεται ως μια γραμμική στοίβα (linear stack) των επιμέρους στρωμάτων του. Ωστόσο, σε περιπτώσεις παράλληλης φόρτωσης και επεξεργασίας των εισόδων κάτι τέτοιο δεν ήταν εφικτό και επιλέχθηκε η μέθοδος σύνθεσης Model που δίνει περισσότερες σχεδιαστικές ελευθερίες.

[8] https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv1D
[9] https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense
[10] https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM
[11] https://www.tensorflow.org/api_docs/python/tf/keras/Sequential

4.1.6 Transformers

Η βελτίωση των μοντέλων μάθησης με χρήση των μετασχηματιστών (transformers) σε συνδυασμό με την πρόοδο στην προεκπαίδευση των μοντέλων, πρόσφεραν σημαντικά πλεονεκτήματα στους μελετητές. Η βιβλιοθήκη transformers της πρωτοβουλίας Hugging Face[12] παρέχει προεκπαιδευμένα μοντέλα βασισμένα σε μετασχηματιστές, με ενοποιημένο τρόπο, χρησιμοποιώντας διεπαφή προγραμματισμού εφαρμογών (API) (Wolf, και συν., 2020). Τα διαθέσιμα μοντέλα έχουν δημιουργηθεί από μέλη της κοινότητας των ερευνητών και απασχολούν πολλούς διαφορετικούς τομείς του πεδίου της μηχανικής μάθησης με χρήση μηχανισμών προσοχής για επεξεργασία φυσικής γλώσσας. Η δημιουργία νέων αρχιτεκτονικών είναι διαδικασία διαρκής και πολλές από αυτές τις αρχιτεκτονικές χρησιμοποιούνται αποδοτικά στην παραγωγή από διάφορες εταιρείες χάρη στην ευελιξία που παρέχει η βιβλιοθήκη (Wolf, και συν., 2020). Επίσης, εκτός από την χρήση των αρχιτεκτονικών στην παραγωγή (production deployment) υπάρχει ευελιξία στην επιλογή πλατφόρμας για την ανάπτυξη των μοντέλων μάθησης. Τέτοιου είδους αρχιτεκτονικές μπορούν να αναπτυχθούν με χρήση TensorFlow, όμως υπάρχει η δυνατότητα να χρησιμοποιηθεί και η πλατφόρμα PyTorch που είναι εξίσου δημοφιλής (Wolf, και συν., 2020). Η χρήση της βιβλιοθήκης transformers δεν δεσμεύει τον εκάστοτε μελετητή, σε μια πλατφόρμα και σε έναν συγκεκριμένο τρόπο ανάπτυξης της μεθοδολογίας που θα εφαρμοστεί.

Η βιβλιοθήκη transformers είναι σχεδιασμένη έτσι ώστε να συμβαδίζει με την πιο δημοφιλή ροή διαδικασιών σε ζητήματα επεξεργασίας φυσικής γλώσσας με χρήση μηχανικής μάθησης (Wolf, και συν., 2020). Αρχικά, γίνεται επεξεργασία των δειγμάτων, στη συνέχεια η εφαρμογή του μοντέλου και τέλος η εμφάνιση των προβλέψεων. Κάθε μοντέλο της βιβλιοθήκης αποτελείται από τρία θεμελιώδη τμήματα (Wolf, και συν., 2020). Το πρώτο από αυτά τα τμήματα είναι ο μηχανισμός (tokenizer) που μετατρέπει την ακολουθία λέξεων στο αρχικό κείμενο σε λεκτικές μονάδες που εμφανίζονται με τη μορφή κωδικοποιήσεων. Ο μηχανισμός αυτός μπορεί να παραμετροποιηθεί χειροκίνητα αλλά παρέχεται και έτοιμος για χρήση, ανάλογα με το μοντέλο που χρησιμοποιείται. Έπειτα, το δεύτερο τμήμα είναι η εκάστοτε υλοποίηση της αρχιτεκτονικής του μοντέλου μετασχηματιστών. Αν και όλες οι αρχιτεκτονικές αξιοποιούν με τον ίδιο τρόπο το μηχανισμό προσοχής πολλαπλών κεφαλών εμφανίζονται και ουσιαστικές διαφορές. Ένα παράδειγμα είναι ο διαφορετικός τρόπος επεξεργασίας της χωρικής κωδικοποίησης των λεκτικών μονάδων. Τέλος, για τα μοντέλα παρέχονται έτοιμες δομές (heads) οι οποίες τοποθετούνται στην έξοδο και προσφέρουν την αξιοποίηση των εμφυτευμάτων για την υλοποίηση του κάθε διαφορετικού σκοπού (π.χ. inference ή classification).

[12] https://huggingface.co

4.2 Παρουσίαση προηγούμενων μεθοδολογιών

Προτού παρουσιαστεί η μέθοδος που χρησιμοποιήθηκε θα γίνει μια σύντομη αναφορά σε άλλες προσεγγίσεις που ακολουθούν παρόμοια δομή διαδικασιών. Συνήθως, στα αρχικά στάδια των μεθοδολογιών διεξάγεται μια διαδικασία κωδικοποίησης του συνόλου των δειγμάτων κειμένου και στη συνέχεια πραγματοποιείται κάποια διαδικασία μάθησης. Οι πιο δημοφιλείς μέθοδοι και αλγόριθμοι που προηγήθηκαν για ανάλυση συγγραφικού ύφους (style-based) του κειμένου αναφέρονται συνοπτικά παρακάτω:

- **Αλγόριθμος διανυσμάτων υποστήριξης (SVMs):** Η χρήση του οδήγησε στην περιορισμένη ανάγκη για δεδομένα χαρακτηρισμένα με ετικέτες και παρείχε πολύ καλή απόδοση σε πραγματικά δεδομένα παρέχοντας υψηλή ακρίβεια (Demestichas, Remoundou, & Adamopoulou, 2020). Η λειτουργία του αλγορίθμου βασίζεται στην προσαρμογή των διανυσμάτων υποστήριξης με σκοπό τον ορισμό ενός βέλτιστου δυνατού υπερεπιπέδου (ή υπερεπιπέδων), που χωρίζει τα δεδομένα των διαφορετικών κλάσεων. Έτσι τα δείγματα ταξινομούνται σε κατηγορίες.
- **Αλγόριθμος k κοντινότερων γειτόνων (k-NN):** Αποτελεί μια μέθοδο ταξινόμησης κειμένου βασισμένη στην ομοιότητα των κωδικοποιήσεων των δειγμάτων. Το πιο δημοφιλές μέτρο σύγκρισης ομοιότητας είναι η Ευκλείδεια απόσταση. Τέτοιου είδους τεχνικές έχουν αξιοποιηθεί για την ταξινόμηση κειμένων σε πολλαπλές κατηγορίες. Η αποτελεσματικότητα της μεθόδου μπορεί να συγκριθεί με εκείνη των μηχανισμών διανυσμάτων υποστήριξης (Demestichas, Remoundou, & Adamopoulou, 2020).
- **Αλγόριθμος λογιστικής παλινδρόμησης (logistic regression):** Πρόκειται για ένα πολύ χρήσιμο και δημοφιλές ερευνητικό εργαλείο. Στον τομέα επεξεργασίας φυσικής γλώσσας, η λογιστική παλινδρόμηση αποτελεί το μέτρο σύγκρισης σε διαδικασίες επιβλεπόμενης μάθησης που αποσκοπούν στην ταξινόμηση ενός δείγματος ως προς δύο κατηγορίες (Demestichas, Remoundou, & Adamopoulou, 2020). Για την απόδοση της ταξινόμησης σε δύο κατηγορίες στη συγκεκριμένη περίπτωση αξιοποιείται συνήθως η σιγμοειδής συνάρτηση.
- **Μοντέλο ταξινόμησης Naïve-Bayes:** Πρόκειται για μια πιθανοτική προσέγγιση του προβλήματος ταξινόμησης. Η συγκεκριμένη μέθοδος βασίζεται στο θεώρημα δεσμευμένης πιθανότητας του Bayes (Demestichas, Remoundou, & Adamopoulou, 2020). Το μοντέλο αυτό είναι πολύ σαφές ως προς τον τρόπο που λειτουργεί και εντέλει εξάγει τα τελικά αποτελέσματα. Η χρήση του μοντέλου για ταξινόμηση ψευδών ειδήσεων έχει διαπιστωθεί πως αποδίδει ικανοποιητικά σε πολύπλοκα προβλήματα που συναντώνται και σε πραγματικές συνθήκες (Demestichas, Remoundou, & Adamopoulou, 2020).
- **Random Forest:** Αποτελεί μια προηγμένη μέθοδο συγκριτικά με τις προαναφερθείσες. Κατά τη συγκεκριμένη διαδικασία μάθησης δημιουργείται ένας μεγάλος αριθμός τυχαίων δέντρων αποφάσεων (Demestichas, Remoundou, & Adamopoulou, 2020). Στη συνέχεια γίνεται επιλογή εκείνων

με τη βέλτιστη απόδοση. Τα δέντρα αποφάσεων που έχουν επιλεγεί συνδυάζονται με σκοπό την εύρεση της βέλτιστης λύσης.

- **Τεχνητά νευρωνικά δίκτυα (ANN: Artificial Neural Networks):** Ήδη από το 2000 η χρήση τεχνητών νευρωνικών δικτύων παρατηρήθηκε πως προσφέρει πολύ αξιόλογα αποτελέσματα (Demestichas, Remoundou, & Adamopoulou, 2020). Τα δίκτυα αποτελούνται από στρώματα κόμβων, όπου ο κάθε κόμβος δέχεται δείγματα δεδομένων σε συνδυασμό με ορισμένα βάρη και υπολογίζει την έξοδο. Ανάλογα με το είδος των βαρών προσδιορίζεται η σημασία του δείγματος ως προς τον εκάστοτε σκοπό της διαδικασίας μάθησης. Άλλοτε δίνεται αυξημένη βαρύτητα και άλλες φορές συμβαίνει το αντίθετο. Τα βάρη ωστόσο ενημερώνονται σε κάθε επανάληψη με βάση τα δεδομένα και ένα αναμενόμενο αποτέλεσμα. Το σημαντικό μειονέκτημα των ANNs σε ζητήματα ταξινόμησης κειμένου είναι πως δεν υπάρχει κάποιος μηχανισμός μνήμης για τις επιμέρους λέξεις (Demestichas, Remoundou, & Adamopoulou, 2020). Συνεπώς, το πρόβλημα είναι πως δεν γίνεται η επόμενη λέξη να προβλεφθεί ως αποτέλεσμα των πληροφοριών και της επιρροής των όρων που προηγήθηκαν. Για αυτόν τον λόγο, αναζητήθηκε μια αποτελεσματικότερη προσέγγιση.
- **Μοντέλα βαθιάς μάθησης (Deep Learning):** Οι σύγχρονες προσεγγίσεις ανίχνευσης των fake news βασίζονται σε αρχιτεκτονικές βαθιάς μάθησης. Τα συγκεκριμένα μοντέλα είναι βασισμένα σε απλά νευρωνικά δίκτυα όμως αποτελούνται από πολύ περισσότερα στρώματα. Αρχικά, προτάθηκαν τρεις βασικοί τύποι μοντέλων στη βιβλιογραφία. Πρώτο είδος αρχιτεκτονικής που προτάθηκε ήταν τα συνελεκτικά νευρωνικά δίκτυα (Convolutional Neural Networks) (Demestichas, Remoundou, & Adamopoulou, 2020). Έχουν χαμηλές απαιτήσεις ως προς την προεπεξεργασία των δειγμάτων και συμπεριλαμβάνουν μικρό αριθμό υπερπαραμέτρων κατά την εκπαίδευση. Παράγουν ικανοποιητικά αποτελέσματα και δεν υπάρχει τόσο μεγάλη ανάγκη επίβλεψης κατά τη διαδικασία μάθησης όσο στα απλά τεχνητά νευρωνικά δίκτυα (ANNs). Το δεύτερο είδος αρχιτεκτονικής που προτάθηκε ήταν τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks) (Demestichas, Remoundou, & Adamopoulou, 2020). Αυτού του είδους η αρχιτεκτονική παρουσιάζει χρονικά μεταβαλλόμενη συμπεριφορά η οποία επιτρέπει την κατανόηση σημασιολογικών πληροφοριών των λέξεων. Η τρίτη αρχιτεκτονική που προτάθηκε για την ταξινόμηση των fake news ήταν τα δίκτυα ιεραρχικών μηχανισμών προσοχής (Hierarchical Attention Networks). Το συγκεκριμένο είδος δικτύων συλλαμβάνει δύο βασικά χαρακτηριστικά των δειγμάτων που αναλύονται (Demestichas, Remoundou, & Adamopoulou, 2020). Η διαδικασία βασίζεται στη δημιουργία αναπαραστάσεων των προτάσεων ανά δείγμα. Στη συνέχεια οι αναπαραστάσεις αυτές ενώνονται σε μια τελική αναπαράσταση του συνολικού κειμένου του δείγματος. Ταυτόχρονα όμως η μέθοδος εκμεταλλεύεται το γεγονός πως διαφορετικές λέξεις και προτάσεις σε ένα έγγραφο συνεισφέρουν με διαφορετική βαρύτητα στο τελικό νόημα.

Όπως είναι κατανοητό δεν μπορούν όλες οι παραπάνω αρχιτεκτονικές να αποδώσουν με τον βέλτιστο τρόπο σε όλες τις μεθοδολογίες κωδικοποίησης κειμένου. Έως το σημείο αυτό έχει αναλυθεί ο τρόπος δόμησης των εμφυτευμάτων που

δημιουργούνται με γνώμονα το πλαίσιο χρήσης των όρων στα δείγματα. Οι κωδικοποιήσεις αυτές έχουν τη μορφή διανυσμάτων πολλαπλών διαστάσεων και οι απλές μέθοδοι όπως μοντέλα υποστήριξης διανυσμάτων (SVM) πολλές φορές δεν μπορούν να διαχειριστούν αυτές τις αναπαραστάσεις με αποδοτικό τρόπο (Erfani, Rajasegarar, Karunasekera, & Leckie, 2016) είτε λόγω της πολυδιάστατης οργάνωσης της πληροφορίας, είτε λόγω της ανάγκης χρήσης υπερβολικά πολλών υπολογιστικών πόρων.

4.3 Περιγραφή πειραματικής μεθοδολογίας

Στην παρούσα ενότητα, θα πραγματοποιηθεί η περιγραφή της μεθοδολογίας που εφαρμόστηκε στο πειραματικό μέρος της εργασίας. Το πιο σημαντικό και θεμελιώδες πλεονέκτημα τέτοιου είδους μεθόδων, συγκριτικά με παρόμοιες προσεγγίσεις, είναι η αυτόματη εξαγωγή χαρακτηριστικών των δειγμάτων με χρήση βαθιάς μάθησης αξιοποιώντας νευρωνικά δίκτυα που συνήθως περιέχουν συνελικτικά στρώματα (Goswami, Kaliyar, & Narang, 2021). Ο βασικός στόχος της πειραματικής διαδικασίας είναι να γίνει σύγκριση των αποτελεσμάτων ταξινόμησης δειγμάτων των διαφορετικών συνόλων δεδομένων όταν υπάρχουν δύο βαθμίδες αξιοπιστίας (true/fake). Ωστόσο, τα σύνολα δεδομένων έχει εξηγηθεί πως παρέχουν διαφορετικό είδος περιεχομένου και κατηγοριοποιήσεων. Κάποια από τα σύνολα δεδομένων περιέχουν δηλώσεις από ομιλίες και συνεντεύξεις, άλλα περιέχουν δείγματα με δημοσιεύσεις στα μέσα κοινωνικής δικτύωσης και άλλα περιέχουν άρθρα συνδυαστικά με τους τίτλους τους ή απλά σύντομο κείμενο. Σε όσα από τα σύνολα δεδομένων το περιεχόμενο ταξινομείται σε παραπάνω από δύο κατηγορίες, η μεθοδολογία εφαρμόζεται δύο φορές. Στην πρώτη εφαρμογή, αξιολογείται η αποτελεσματικότητα της διαδικασίας στην ταξινόμηση των δειγμάτων σε πολλαπλές κατηγορίες και στη δεύτερη εφαρμογή, ελέγχεται η απόδοση της μεθόδου στην ταξινόμηση των δειγμάτων σε δύο κατηγορίες. Ακόμα, το σώμα των άρθρων και ο τίτλος ανά δείγμα εξετάζονται τόσο ξεχωριστά όσο και συνδυαστικά. Στη συνέχεια, θα γίνει περιγραφή της πειραματικής μεθόδου σε διακριτά βήματα.

1. Στο πρώτο στάδιο της διαδικασίας γίνεται η προετοιμασία του κάθε συνόλου δεδομένων. Για παράδειγμα, υπάρχουν σύνολα δεδομένων χωρισμένα σε δύο μικρότερα υποσύνολα ή περιέχουν δείγματα που δεν έχουν κατηγοριοποιηθεί. Οπότε στις περιπτώσεις που εμφανίζεται το συγκεκριμένο φαινόμενο γίνεται ενοποίηση των υποσυνόλων και στη συνέχεια τυχαία αναδιάταξη των δειγμάτων. Ακόμα, πραγματοποιείται αφαίρεση των δειγμάτων που δεν χαρακτηρίζονται από κάποια ετικέτα. Σε ορισμένα σύνολα δεδομένων που λόγω ανισορροπίας εμφανίζονται προβλήματα απόδοσης είναι ανάγκη να εφαρμοστεί κάποια τεχνική αύξησης των δειγμάτων στην κατηγορία που μειονεκτεί.

2. Στη συνέχεια, με χρήση της βιβλιοθήκης transformers γίνεται εισαγωγή του μηχανισμού κωδικοποίησης του κειμένου σε μορφή αναγνωριστικών (tokenizer). Η έκδοση του tokenizer που αξιοποιείται είναι

παραμετροποιημένη για να αποδώσει βέλτιστα το κείμενο με τρόπο κατανοητό από την έκδοση του μοντέλου DistilBERT που θα χρησιμοποιηθεί. Εκτός από τη μετατροπή του κειμένου σε μονάδες αναγνωριστικών, προηγείται και προσθήκη ειδικών ετικετών που σηματοδοτούν την αρχή του κειμένου και το διαχωρισμό των προτάσεων.

3. Στη συνέχεια, με χρήση της βιβλιοθήκης των transformers γίνεται εισαγωγή του DistilBERT σε μορφή προσαρμοσμένη για εργασίες ταξινόμησης κειμένου. Η συγκεκριμένη έκδοση αποτελείται από την αρχιτεκτονική του παραδοσιακού μοντέλου κωδικοποίησης DistilBERT με την προσθήκη μιας δομής ταξινομητή στην έξοδο του, η αρχιτεκτονική του ταξινομητή διαφοροποιείται ανάλογα με τον αριθμό κατηγοριών του συνόλου δεδομένων. Το συγκεκριμένο μοντέλο περιέχει 66.955.010 παραμέτρους και παρέχεται με προεκπαίδευση σε μεγάλα σύνολα δεδομένων. Το αποτέλεσμα είναι οι λέξεις να έχουν αποδοθεί σε πλαίσιο χρήσης που διαφέρει από τον τρόπο που αξιοποιούνται οι ίδιοι όροι σε δείγματα fake news. Για αυτόν τον λόγο, επιτρέποντας την εκπαίδευση όλων των παραμέτρων του μοντέλου πραγματοποιείται από την αρχή μια διαδικασία μάθησης. Ακολουθώντας αυτή τη διαδικασία, τα εμφυτεύματα που προκύπτουν σε επόμενο στάδιο αποτυπώνουν όσο το δυνατό καλύτερα τις λέξεις ανάλογα με τον τρόπο που χρησιμοποιούνται. Το μοντέλο που έχει επιλεχθεί βασίζει τους υπολογισμούς του στους τανυστές, στην κυρίαρχη δομή υπολογισμών της βιβλιοθήκης TensorFlow.

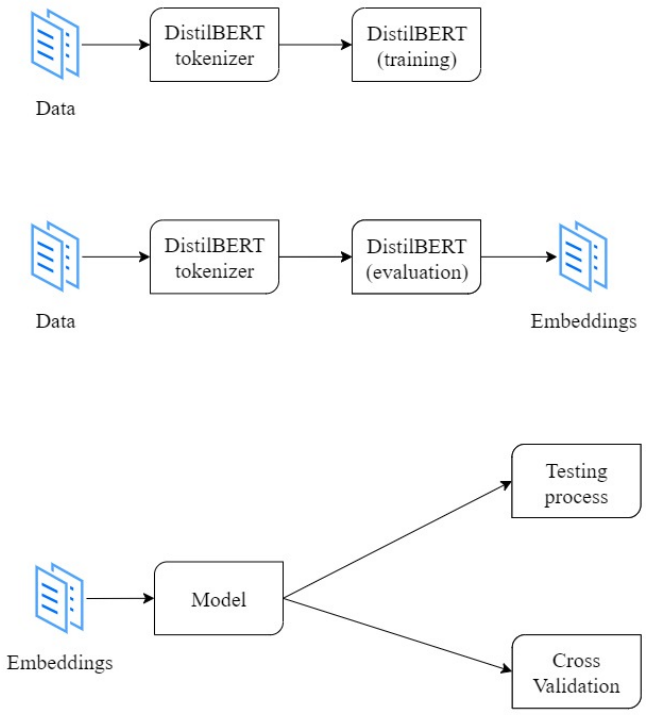
4. Αφού έχει γίνει η προεκπαίδευση του μοντέλου πραγματοποιείται αξιολόγηση του κάθε δείγματος από το DistilBERT και εξάγεται η κωδικοποίηση. Για κάθε μια μονάδα του δείγματος κωδικοποιούνται 768 χαρακτηριστικά στην αρχιτεκτονική που προσφέρεται από τη βιβλιοθήκη. Ωστόσο, το μήκος του κάθε δείγματος σε λεκτικές μονάδες είναι διαφορετικό (x,768). Υπάρχουν δύο επιλογές ως προς τα στοιχεία των τελικών κωδικοποιήσεων που μπορούν να χρησιμοποιηθούν. Για την ταξινόμηση μπορεί να χρησιμοποιηθεί η κωδικοποίηση της πρώτης ετικέτας ([CLS]) ανά δείγμα που προστίθεται από τον tokenizer και παρέχει την απαραίτητη πληροφορία για σκοπούς ταξινόμησης. Όμως, στην παρούσα εργασία χρησιμοποιήθηκαν εναλλακτικά τα εμφυτεύματα για να προκύψει μια όσο το δυνατό πιο αντιπροσωπευτική αναπαράσταση της πληροφορίας του δείγματος προς ταξινόμηση. Συγκεκριμένα, υπολογίζεται ο μέσος όρος των χαρακτηριστικών των λεκτικών μονάδων. Έτσι, η τελική διάσταση του εμφυτεύματος κάθε δείγματος είναι (1,768). Η ίδια διαδικασία εφαρμόζεται για τα δεδομένα επαλήθευσης αποτελεσματικότητας των αρχιτεκτονικών που θα χρησιμοποιηθούν αργότερα.

5. Στη συνέχεια, το σύνολο των εμφυτευμάτων εκπαίδευσης τροφοδοτείται στις 7 επιμέρους αρχιτεκτονικές που δημιουργήθηκαν στο πλαίσιο της εργασίας. Κάθε εμφύτευμα πριν προωθηθεί ως είσοδος θα μετασχηματιστεί σε διαστάσεις (768,1) για να είναι συμβατό με τις δομές των αρχιτεκτονικών.

Μετά την ολοκλήρωση της εκπαίδευσης που διαρκεί 10 εποχές γίνεται επαλήθευση της απόδοσης του μοντέλου για να διαπιστωθεί η ικανότητα ταξινόμησης σε ένα μικρό σύνολο δεδομένων που δεν έχει συναντήσει ξανά.

6. Έπειτα πραγματοποιείται η διαδικασία της διασταυρωμένης επικύρωσης (cross validation). Η μέθοδος αυτή δίνει χρήσιμες πληροφορίες για την ικανότητα του μοντέλου να προβλέπει την ταξινόμηση δειγμάτων που δεν έχει επεξεργαστεί ξανά. Συγκεκριμένα, τα δεδομένα εκπαίδευσης χωρίζονται σε 5 όμοια υποσύνολα και επαναλαμβάνεται από την αρχή η διαδικασία εκπαίδευσης 5 φορές. Κάθε φορά χρησιμοποιούνται 4 από τα 5 υποσύνολα για εκπαίδευση και εξετάζεται η απόδοση της αρχιτεκτονικής στο 1 υποσύνολο που απέμεινε. Στο τέλος, οι μετρικές αξιολόγησης της αρχιτεκτονικής για τις 5 διαφορετικές διαδικασίες συνδυάζονται για τον υπολογισμό ενός μέσου όρου ανά μετρική αλλά και της απόκλισης που εμφανίζεται.

Θεωρώντας πως έχει γίνει η εισαγωγή των απαραίτητων εργαλείων και αρχιτεκτονικών στο εκάστοτε προγραμματιστικό περιβάλλον, στη συνέχεια παρέχεται ένα σχεδιάγραμμα που περιγράφει τη διαδικασία που αναλύθηκε παραπάνω.



Εικόνα 27 - Παρουσίαση της πειραματικής διαδικασίας.

Το όριο της εισόδου που μπορεί να δεχτεί για κωδικοποίηση το μοντέλο DistilBERT είναι μόλις 512 λεκτικές μονάδες. Αυτός ο περιορισμός προκύπτει από το ανώτατο όριο που μπορούν να διαχειριστούν οι μηχανισμοί προσοχής του μοντέλου για να είναι αποτελεσματικοί. Όταν το δείγμα υπερβεί σε μήκος λέξεων τον περιορισμό τότε λαμβάνονται υπόψη μόνο οι πρώτες 512 λεκτικές μονάδες. Οπότε σε κείμενα μεγάλου μήκους μια πιθανή λύση είναι να υπολογιστεί ποιες προτάσεις είναι καταλληλότερες για μελέτη και στη συνέχεια να συντεθεί ένα κείμενο με 512 λέξεις

αποτελούμενο από τις πιο αξιόλογες προτάσεις. Για αυτό τον λόγο χρησιμοποιήθηκε η διεπαφή προγραμματισμού εφαρμογών του Innovative Data Intelligence Research Laboratory[13] του University of Texas at Arlington. Η συγκεκριμένη διεπαφή προγραμματισμού[14] μεταξύ άλλων λειτουργιών παρέχει κατάταξη των προτάσεων ενός κειμένου ανάλογα με το βαθμό σημαντικότητας των ισχυρισμών τους. Έτσι στη συνέχεια διατηρώντας τις σημαντικότερες προτάσεις και συνδυάζοντάς τις με τη σειρά που είχαν στο αρχικό κείμενο συντίθεται ένα νέο δείγμα προς μελέτη. Η τεχνική αυτή προτάθηκε στη μελέτη με τίτλο Fake news detection using parallel BERT deep neural networks (Farokhian, Rafe, & Veisi, 2022) όπου δημιουργήθηκε παρόμοιο πρόβλημα σε ταξινόμηση fake news με χρήση του BERT.

Στο σύνολο δεδομένων ISOT η διαδικασία που αναλύθηκε παραπάνω εκτελέστηκε τρεις φορές. Δηλαδή, η μέθοδος εφαρμόστηκε ξεχωριστά στους τίτλους, στα άρθρα και στη συμπυκνωμένη μορφή των άρθρων. Στη συνέχεια, εξετάστηκε ο συνδυασμός δειγμάτων τίτλων μαζί με τα αντίστοιχα άρθρα αξιοποιώντας ένα επιπλέον είδος αρχιτεκτονικής που αναπτύχθηκε συγκεκριμένα για το σκοπό αυτό. Το ίδιο συνέβη και για τα ζεύγη τίτλων και δειγμάτων που προέκυψαν από την τεχνική της σύμπτυξης των άρθρων. Ο κώδικας εκτέλεσης της διαδικασίας είναι διαθέσιμος μέσω της υπερσύνδεσης στο Παράρτημα **A.1**.

Στα σύνολα δεδομένων LIAR και PHEME περιέχεται μόνο ένας τύπος δείγματος. Συγκεκριμένα, στην πρώτη περίπτωση παρέχονται σύντομες δηλώσεις διαφόρων ομιλητών, ενώ στη δεύτερη δημοσιεύσεις στα μέσα κοινωνικής δικτύωσης. Όμως, τα δείγματα στο σύνολο δεδομένων LIAR ταξινομούνται σε έξι βαθμίδες αξιοπιστίας και στο σύνολο PHEME σε τρεις. Η μεθοδολογία που αναλύθηκε παραπάνω εφαρμόζεται και στις δύο περιπτώσεις χωρίς κάποια απόκλιση, με τη διαφορά πως η συνάρτηση ενεργοποίησης του τελικού στρώματος ταξινόμησης των εμπλεκόμενων μοντέλων αντί για σιγμοειδής είναι softmax. Όμως, στη συνέχεια ενοποιώντας συγκεκριμένες κατηγορίες στο σύνολο LIAR και αντίστοιχα αφαιρώντας δείγματα που δεν έχει προσδιοριστεί η εγκυρότητά τους στο σύνολο PHEME εφαρμόζεται η μεθοδολογία για μια ακόμη φορά. Ωστόσο, κατά τη δεύτερη εφαρμογή της μεθοδολογίας το κάθε σύνολο δεδομένων περιέχει μόνο δύο βαθμίδες αξιοπιστίας για κάθε δείγμα. Ο κώδικας εκτέλεσης της διαδικασίας για τα σύνολα δεδομένων LIAR και PHEME αντίστοιχα είναι διαθέσιμος στο Παράρτημα **A.2** και Παράρτημα **A.3**.

Στο σύνολο δεδομένων FakeNewsNet τα δείγματα είναι απλοί τίτλοι άρθρων. Το συγκεκριμένο σύνολο δεδομένων χρησιμοποιήθηκε για τον προσδιορισμό περιπτώσεων fake news με μικρό μέσο μήκος λέξεων. Το πρόβλημα με αυτό το σύνολο δεδομένων είναι πως υπάρχει έντονη ανισορροπία, μεταξύ των κατηγοριών του. Για την επίλυση του προβλήματος, χρησιμοποιήθηκε η βιβλιοθήκη nltk με στόχο την αύξηση των δειγμάτων της κατηγορίας που μειονεκτεί. Κάθε δείγμα του αντίστοιχου υποσυνόλου αναλύθηκε και επαναδιατυπώθηκε με συνώνυμες λέξεις. Έτσι, το τελικό νόημα του εκάστοτε νέου δείγματος διατηρείται. Τα δείγματα που

[13] https://idir.uta.edu/index.html
[14] https://idir.uta.edu/claimbuster/api/

παράχθηκαν προστέθηκαν στο σύνολο δεδομένων και μειώθηκε η ανισορροπία. Έπειτα, εφαρμόστηκε η μεθοδολογία που αναλύθηκε παραπάνω στα δείγματα. Ο κώδικας εκτέλεσης της διαδικασίας είναι διαθέσιμος μέσω της υπερσύνδεσης στο Παράρτημα **A.4**.

Σχετικά με το σύνολο FakeNewsChallenge, η μεθοδολογία εφαρμόστηκε συνδυάζοντας τις προσεγγίσεις των LIAR, PHEME και ISOT. Αυτό συμβαίνει, γιατί το σύνολο δεδομένων περιέχει τέσσερις κατηγορίες ταξινόμησης και τα δείγματα αποτελούνται από τους τίτλους και τα αντίστοιχα άρθρα τους. Ωστόσο, υπάρχουν κάποιες ιδιαιτερότητες, όπως το γεγονός πως το συγκεκριμένο σύνολο δεδομένων βασίζεται κυρίως στην αλληλεπίδραση μεταξύ τίτλου και κειμένου όπως αναλύθηκε στο τρίτο κεφάλαιο. Συνεπώς, δεν θα υπήρχαν χρήσιμα αποτελέσματα αν οι τίτλοι εξετάζονταν σαν ξεχωριστή κατηγορία δειγμάτων όπως συνέβη στο σύνολο ISOT. Η προτεινόμενη διαδικασία, εφαρμόστηκε ξεχωριστά για τα άρθρα, για τα συμπυκνωμένα άρθρα αλλά και συνδυαστικά με τους τίτλους στην περίπτωση ύπαρξης δειγμάτων τεσσάρων βαθμίδων ταξινόμησης. Το ίδιο συνέβη έπειτα για την πραγματοποίηση της δυαδικής ταξινόμησης ενώνοντας τις κατηγορίες σε ζεύγη όπως αναφέρθηκε και κατά την ανάλυση του συνόλου δεδομένων. Μια ακόμα παρατήρηση που αφορά τη μεθοδολογία στο συγκεκριμένο σύνολο δεδομένων είναι πως υπήρξαν πολύ μικρές διαφορές όταν χρησιμοποιήθηκε μεταφορά μάθησης (transfer learning) συγκριτικά με την εκπαίδευση του μοντέλου κωδικοποίησης από την αρχή. Για να επιτευχθεί εξοικονόμηση πόρων και μείωση του απαιτούμενου χρόνου εκπαίδευσης, έγινε η επιλογή να διατηρηθούν εκπαιδεύσιμα μόνο τα τρία ανώτερα στρώματα, που αποτελούν τον ταξινομητή στην έξοδο του μοντέλου, κατά το στάδιο εκπαίδευσης του DistilBERT. Ο κώδικας εκτέλεσης της διαδικασίας είναι διαθέσιμος μέσω της υπερσύνδεσης στο Παράρτημα **A.5**.

Για να διαπιστωθεί η επιτυχία της μεθόδου δεν αρκεί η αξιολόγηση μόνο ως προς μια μοναδική μετρική. Ο προσδιορισμός μιας επιτυχημένης διαδικασίας μάθησης για εργασίες ταξινόμησης κειμένου πραγματοποιείται από την αξιολόγηση των μετρικών accuracy, precision, recall και F1 (Alghamdi, Lin, & Luo, 2022). Κάθε μια από τις προαναφερθείσες μετρικές αντιπροσωπεύει ένα διαφορετικό δείκτη επίδοσης που υπολογίζεται με βάση την ταξινόμηση των δειγμάτων επαλήθευσης από το μοντέλο, συνδυαστικά με την πραγματική ετικέτα τους στο εκάστοτε υποσύνολο. Έτσι τα δείγματα χωρίζονται ως αληθώς θετικά (TP: true positive), αληθώς αρνητικά (TF: true negative), ψευδώς θετικά (FP: false positive) και ψευδώς αρνητικά (FN: false negative). Η μετρική accuracy εκφράζει την ικανότητα του μοντέλου να κατατάσσει τα δείγματα ορθά ως αληθή ή ψευδή και υπολογίζεται ως εξής (Alghamdi, Lin, & Luo, 2022):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Η μετρική precision είναι ο λόγος των αληθώς θετικών δειγμάτων προς τα συνολικά δείγματα που έχουν ταξινομηθεί στη θετική κατηγορία (αληθώς θετικά και ψευδώς θετικά). Η συγκεκριμένη μετρική εκφράζει την ακρίβεια του μοντέλου να εντοπίζει θετικά δείγματα και υπολογίζεται ως εξής (Alghamdi, Lin, & Luo, 2022):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Η μετρική recall εκφράζει το βαθμό κατά τον οποίο καταφέρνει μια αρχιτεκτονική να εντοπίζει όλα τα θετικά δείγματα. Υπολογίζεται ως ο λόγος των αληθώς θετικών δειγμάτων προς το σύνολο των δειγμάτων που είναι όντως θετικά (αληθώς θετικά και ψευδώς αρνητικά δείγματα) όπως φαίνεται παρακάτω (Alghamdi, Lin, & Luo, 2022):

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Όλες οι μετρικές που αναφέρθηκαν καθώς και η μετρική F1 υπολογίζονται όχι μόνο για προβλήματα δυαδικής ταξινόμησης αλλά και σε προβλήματα ταξινόμησης περισσότερων κατηγοριών. Η βιβλιοθήκη TensorFlow παρέχει τη δυνατότητα υπολογισμού των μετρικών με βάση το σύνολο επαλήθευσης για κάθε επιμέρους κατηγορία. Η μετρική F1-score είναι ο σταθμισμένος μέσος όρος των μετρικών recall και precision (Alghamdi, Lin, & Luo, 2022). Υπολογίζεται όπως φαίνεται παρακάτω:

$$\text{F1} = \frac{2 * \text{recall} * \text{precision}}{\text{precision} + \text{recall}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}$$

Πριν γίνει παρουσίαση των αρχιτεκτονικών πρέπει να διευκρινιστεί μια ακόμα πληροφορία για το μοντέλο DistilBERT που είναι υπεύθυνο για την κωδικοποίηση του κειμένου. Μια από τις πιο σημαντικές παραμέτρους έπειτα από δοκιμές είναι ο βαθμός μάθησης (learning rate). Η συγκεκριμένη παράμετρος θέτει το ρυθμό με τον οποίο μεταβάλλονται τα βάρη του μοντέλου ανά εποχή εκπαίδευσης. Δηλαδή προσδιορίζει πόσο γρήγορα ή αργά συγκλίνει το μοντέλο (Zeiler, 2012). Για αυτό το λόγο έγιναν δοκιμές σε εύρος τιμών 2e-5 έως 4e-5 με σκοπό να διαπιστωθεί πότε εμφανίζεται η καλύτερη απόδοση της εκπαίδευσης του μοντέλου στα δεδομένα προς κωδικοποίηση.

Πίνακας 3 - Βέλτιστοι βαθμοί μάθησης DistilBERT ανά σύνολο δεδομένων

Σύνολο δεδομένων	Βαθμός μάθησης (learning rate)
ISOT	3e-5
PHEME	3e-5
PHEME (Binary)	3e-5
LIAR	2e-5
LIAR (Binary)	3e-5
FakeNewsNet	3e-5
FakeNewsChallenge	3e-5
FakeNewsChallenge (Binary)	3e-5

4.4 Περιγραφή αρχιτεκτονικών

4.4.1 Περιγραφή λειτουργίας δομικών στοιχείων των αρχιτεκτονικών

Η δημιουργία των αρχιτεκτονικών βασίστηκε κυρίως στη μελέτη με τίτλο FakeBERT: Fake news detection in social media with a BERT-based deep learning approach (Goswami, Kaliyar, & Narang, 2021). Κάθε δείγμα του συνόλου δεδομένων μετά τη διαδικασία κωδικοποίησης που αναλύθηκε στην προηγούμενη ενότητα αντιστοιχεί σε ένα εμφύτευμα. Το κάθε εμφύτευμα του συνόλου εκπαίδευσης στη συνέχεια δίνεται ως είσοδος στα μοντέλα που αναπτύχθηκαν. Σε αρκετές μελέτες το πρόβλημα της ταξινόμησης των fake news έχει επιλυθεί με τη χρήση κάποιου μηχανισμού κωδικοποίησης δειγμάτων μιας κατεύθυνσης (unidirectional word embeddings) (Goswami, Kaliyar, & Narang, 2021). Οι αναπαραστάσεις αυτές στη συνέχεια τροφοδοτούν ένα νευρωνικό δίκτυο αποτελούμενο από στρώματα ζευγών συνελικτικών στρωμάτων μιας διάστασης (1D-convolutional layer) και στρωμάτων μέγιστης συσσώρευσης (max-pooling layer) (Goswami, Kaliyar, & Narang, 2021). Στην συγκεκριμένη εργασία, τα νευρωνικά δίκτυα που χρησιμοποιήθηκαν μετά την κωδικοποίηση αξιοποιούν παρόμοιες δομές. Στη συνέχεια, θα παρουσιαστούν σύντομα κάποιες από τις συχνότερα εμφανιζόμενες στην πειραματική διαδικασία.

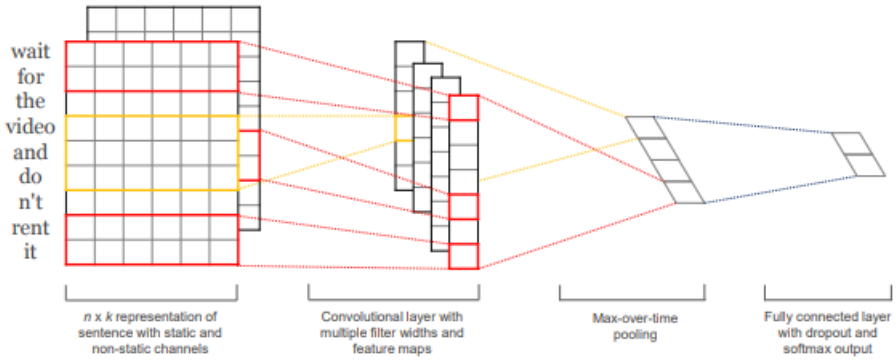
Τα συνελικτικά στρώματα μιας διάστασης βασίζονται στην πράξη της συνέλιξης όπως άλλωστε δηλώνει και το όνομά τους. Παραδοσιακά τα στρώματα αυτά βασίζονται στη χρήση φίλτρων με πυρήνες συγκεκριμένου μεγέθους που χρησιμοποιούνται για την εξαγωγή των σημασιολογικών αναπαραστάσεων λέξεων διαφορετικού μήκους (Goswami, Kaliyar, & Narang, 2021). Η βασική λειτουργία τους περιλαμβάνει πολλαπλασιασμό πολυδιάστατων πινάκων (μη γραμμική διαδικασία) και στη συνέχεια χρήση κάποιας συνάρτησης ενεργοποίησης για την παραγωγή της τελικής εξόδου (Goswami, Kaliyar, & Narang, 2021). Στο τέλος, παράγονται αναπαραστάσεις των χαρακτηριστικών σε μεγαλύτερες διαστάσεις για παράδειγμα μια είσοδος (768,1) μετατρέπεται σε (764,128).

Για να γίνει πιο κατανοητή η διαδικασία λειτουργίας τους θα δοθεί ένα παράδειγμα. Έστω ένα εμφύτευμα με διαστάσεις (768,1) που θα τροφοδοτήσει ένα συνελικτικό στρώμα με 128 φίλτρα, μέγεθος πυρήνα 5 στοιχείων και συνάρτηση ενεργοποίησης ReLU. Επειδή το μέγεθος του πυρήνα είναι 5 θα γίνει εφαρμογή των φίλτρων στα χαρακτηριστικά των θέσεων 1 έως 5, στη συνέχεια στα χαρακτηριστικά στις θέσεις 2 έως 6 και ούτω καθεξής. Επειδή υπάρχουν 128 φίλτρα που συλλαμβάνουν διαφορετικά χαρακτηριστικά, στο τέλος κάθε έξοδος θα περιέχει 128 στοιχεία. Έπειτα με χρήση της συνάρτησης ενεργοποίησης τα αρνητικά στοιχεία μηδενίζονται και τα θετικά διατηρούνται ως έχουν. Τελικά παράγεται μια αναπαράσταση του δείγματος με διαστάσεις (764,128). Ο αριθμός των εξόδων που παράγονται (764) υπολογίζεται ως εξής με βάση τα όσα αναφέρθηκαν:

$$\text{output length} = \text{input length} - \text{kernel size} + 1$$

Στη συνέχεια, οι έξοδοι αυτές τροφοδοτούν στρώματα μέγιστης συσσώρευσης (max-pooling). Τα στρώματα αυτά με βάση ένα συγκεκριμένο μέγεθος πυρήνα k,

επιλέγουν το μέγιστο, ανά k στοιχεία, των 128 χαρακτηριστικών των 764 εξόδων. Η διαδικασία αυτή λειτουργεί ως ένας τρόπος υποδειγματοληψίας, για να μειωθεί το υπολογιστικό κόστος της επεξεργασίας της εξόδου των συνελικτικών στρωμάτων (Goswami, Kaliyar, & Narang, 2021). Όταν η δομή εμφανίζεται σε μια αρχιτεκτονική επαναληπτικά τότε προκαλείται προοδευτική μείωση των διαστάσεων με σκοπό την σταδιακή αποσυμφόρηση του υπολογιστικού φόρτου (Goswami, Kaliyar, & Narang, 2021). Παρακάτω δίνεται μια απεικόνιση της διαδικασίας που αναλύθηκε και περιλαμβάνει τα δύο είδη στρωμάτων που αναφέρθηκαν.



Εικόνα 28 - Παρουσίαση συνελικτικού στρώματος ακολουθούμενο από στρώμα μέγιστης συσσώρευσης και πλήρως συνδεδεμένου στρώματος. *(Yoon, 2014)*

Ωστόσο, στη συγκεκριμένη μεθοδολογία πρέπει να επισημανθεί πως το κάθε εμφύτευμα περιγράφει το μέσο όρο των 768 χαρακτηριστικών των λέξεων του εκάστοτε δείγματος (1,768). Οπότε, η παραπάνω διαδικασία δεν θα εφαρμοστεί σε λέξεις αλλά θα εφαρμοστεί σε 768 στοιχεία με μια διάσταση (768,1) όπως αναφέρθηκε και στην ανάλυση της μεθοδολογίας.

Επιπροσθέτως, τα στρώματα μακράς βραχύχρονης μνήμης (LSTM:Long Short Term Memory) υπήρξαν πολύ σημαντικά για τη δόμηση των μοντέλων που προτείνονται στην παρούσα εργασία. Τα στρώματα μακράς βραχύχρονης μνήμης αποτελούν ένα είδος αναδρομικού δικτύου το οποίο έχει σχεδιαστεί, για να διαχειρίζεται με αποτελεσματικό τρόπο δεδομένα που αποτελούνται από ακολουθίες (π.χ. κείμενο) (Hochreiter & Schmidhuber, 1997). Η κεντρική ιδέα της λειτουργίας τους αφορά τη χρήση πυλών (gates) και κυττάρων μνήμης (memory cells) (Hochreiter & Schmidhuber, 1997). Οι δομές αυτές δίνουν τη δυνατότητα επιλεκτικής συγκράτησης πληροφορίας. Με αυτόν τον τρόπο, μπορούν να εντοπιστούν ακριβέστερα οι εξαρτήσεις μεταξύ στοιχείων μιας ακολουθίας που εμφανίζονται σε μεγάλη απόσταση μεταξύ τους (π.χ. απομακρυσμένες λέξεις σε ένα κείμενο) συγκριτικά με άλλου είδους αναδρομικά δίκτυα (RNNs: Recurrent Neural Networks).

Συνεχίζοντας με επιπλέον πληροφορίες για τα στρώματα που χρησιμοποιούνται θα γίνει αναφορά στα Flatten layers. Όπως έχει γίνει κατανοητό έως αυτό το σημείο οι προαναφερθείσες δομές παράγουν συνήθως πολυδιάστατα αποτελέσματα. Ο ρόλος των στρωμάτων αυτών είναι να μετατρέψουν τα αποτελέσματα αυτά σε μονοδιάστατες αναπαραστάσεις που θα τροφοδοτήσουν στη συνέχεια τα πλήρως συνδεδεμένα στρώματα (Goswami, Kaliyar, & Narang, 2021).

Τα πλήρως συνδεδεμένα στρώματα αποτελούνται από ένα σύνολο νευρώνων. Κάθε νευρώνας δέχεται πληροφορία από όλους τους νευρώνες του προηγούμενου στρώματος και έτσι άλλωστε προκύπτει η ονομασία της δομής. Οι έξοδοι των νευρώνων υπολογίζονται με χρήση ενός μητρώου βαρών (weight matrix), ενός διανύσματος προκαθορισμένων τιμών (bias vector) και των αποτελεσμάτων της συνάρτησης ενεργοποίησης του προηγούμενου στρώματος (Goswami, Kaliyar, & Narang, 2021). Συνήθως χρησιμοποιούνται για την αποφυγή του κορεσμού της διαδικασίας μάθησης (overfitting) (Goswami, Kaliyar, & Narang, 2021).

Ένα επιπλέον είδος μηχανισμού που αξιοποιήθηκε ονομάζεται Dropout. Πρόκειται για μια τεχνική κανονικοποιήσης όπου ένα καθορισμένο ποσοστό τυχαία επιλεγμένων νευρώνων αγνοούνται κατά τη διαδικασία εκπαίδευσης και συνεπώς κατά τη διαδικασία ενημέρωσης βαρών (Goswami, Kaliyar, & Narang, 2021). Στις αρχιτεκτονικές της παρούσας εργασίας χρησιμοποιήθηκε ο μηχανισμός αυτός στα πλήρως συνδεδεμένα στρώματα.

Στη συνέχεια θα γίνει περιγραφή της συνάρτησης ενεργοποίησης ReLU (Rectified Linear Unit). Η συνάρτηση αυτή είναι από τις συχνότερα χρησιμοποιούμενες για τα αποτελέσματα των συνελικτικών στρωμάτων. Το βασικό της πλεονέκτημα είναι πως δεν ενεργοποιούνται όλοι οι νευρώνες ταυτόχρονα. Η συνάρτηση είναι μη γραμμική όπως για παράδειγμα η σιγμοειδής συνάρτηση και χρησιμοποιείται μετά τη συνέλιξη (Goswami, Kaliyar, & Narang, 2021). Περιγράφεται ως εξής για είσοδο z:

$$\sigma = \max(0, z)$$

Το κριτήριο τερματισμού της εκπαίδευσης (loss function) είναι η διασταυρούμενη εντροπία (cross entropy). Η οποία προκύπτει από τη σύγκριση της ταξινόμησης δειγμάτων του μοντέλου με τις ετικέτες που τους έχουν αποδοθεί στο σύνολο δεδομένων, οι οποίες αναπαριστούν την πραγματική κατανομή πιθανότητας για την ταξινόμηση. Όσο μειώνεται η εντροπία τόσο πιο ακριβές θεωρείται το μοντέλο. Σε πρόβλημα δυαδικής ταξινόμησης με μέτρο πιθανότητας p και τυχαία μεταβλητή πειράματος y ο τύπος υπολογισμού της δυαδικής διασταυρούμενης εντροπίας είναι ο παρακάτω. Ωστόσο η μετρική μπορεί να χρησιμοποιηθεί και με παραλλαγές για ταξινόμηση σε περισσότερες κατηγορίες.

$$L = -(y\log(p) + (1-y)\log(1-p))$$

4.4.2 CNN

Έπειτα από κάποιες προσαρμογές συγκριτικά με τη μελέτη FakeBERT: Fake news detection in social media with a BERT-based deep learning approach (Goswami, Kaliyar, & Narang, 2021), προκύπτει μια αρχιτεκτονική που αποτελείται κυρίως από τρεις δομές ζευγών συνελικτικών στρωμάτων και στρωμάτων μέγιστης συσσώρευσης. Τα συνελικτικά στρώματα έχουν 128 φίλτρα και μέγεθος πυρήνα (kernel size) ίσο με 5. Επίσης, τα στρώματα μέγιστης συσσώρευσης έχουν μέγεθος πυρήνα ίσο με 5. Μόλις ολοκληρωθεί η επεξεργασία στις τρεις διαδοχικές δομές ακολουθεί μια δομή Flatten η οποία στοχεύει στη μονοδιάστατη αναπαράσταση της εισόδου που δέχεται. Έπειτα, ακολουθούν δύο δομές πλήρως συνδεδεμένων

στρωμάτων με 128 και 64 νευρώνες και χρησιμοποιείται Dropout σε ποσοστό 50%. Τέλος, ένα πλήρως συνδεδεμένο στρώμα με σιγμοειδή συνάρτηση ενεργοποίησης (sigmoid) υπολογίζει την ταξινόμηση του δείγματος. Σε περιπτώσεις ταξινόμησης σε παραπάνω από δύο κατηγορίες χρησιμοποιείται συνάρτηση ενεργοποίησης softmax. Η απεικόνιση της αρχιτεκτονικής αποδίδεται στο Παράρτημα **B.1**. Στη συνέχεια θα παρουσιαστούν με συντομία οι παράμετροι ανά στρώμα και οι διαστάσεις εισόδου και εξόδου. Συνολικά το μοντέλο έχει 255.233 παραμέτρους που είναι όλες εκπαιδεύσιμες.

Πίνακας 4 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN ανά στρώμα

	Input	Output	Parameters
Conv1D	(768,1)	(764,128)	768
Max Pooling	(764,128)	(152,128)	0
Conv1D	(152,128)	(148,128)	82048
Max Pooling	(148,128)	(29,128)	0
Conv1D	(29,128)	(25,128)	82048
Max Pooling	(25,128)	(5,128)	0
Flatten	(5,128)	(1,640)	0
Dense	(640)	(128)	82048
Dropout	(128)	(128)	0
Dense	(128)	(64)	8256
Dropout	(64)	(64)	0
Dense	(64)	(1)	65

Οι αλλαγές που έγιναν συγκριτικά με την αρχιτεκτονική που προτάθηκε στη μελέτη είναι πως πλέον αλλάζει ο αριθμός των νευρώνων των πλήρως συνδεδεμένων στρωμάτων συνδυαστικά με μεγαλύτερο ποσοστό επιρροής στην τεχνική Dropout. Ειδικότερα, το δεύτερο πλήρως συνδεδεμένο στρώμα αποτελείται από 64 νευρώνες αντί για 128. Ακόμα, το ποσοστό των νευρώνων που επηρεάζονται από την τεχνική Dropout αυξάνεται από 20% σε 50%.

4.4.3 Bidirectional LSTM

Η αρχιτεκτονική προέκυψε από τη μελέτη FakeBERT: Fake news detection in social media with a BERT-based deep learning approach (Goswami, Kaliyar, & Narang, 2021) έπειτα από ορισμένες προσαρμογές. Συγκεκριμένα, αποτελείται από δύο διαδοχικές δομές ζευγών συνελικτικών στρωμάτων που ακολουθούνται από στρώματα μέγιστης συσσώρευσης. Τα συνελικτικά στρώματα έχουν 128 φίλτρα και μέγεθος πυρήνα ίσο με 5. Ενώ, τα στρώματα μέγιστης συσσώρευσης έχουν μέγεθος πυρήνα ίσο με 2. Οι έξοδοι στη συνέχεια τροφοδοτούν ένα στρώμα μακράς βραχύχρονης μνήμης δύο κατευθύνσεων με 64 κρυφά στρώματα. Στη συνέχεια, υπάρχει ένα στρώμα κανονικοποιήσης συνόλων εκπαίδευσης (Batch Normalization). Ο συγκεκριμένος μηχανισμός συνεισφέρει σε βελτιωμένους χρόνους εκπαίδευσης, γρηγορότερη σύγκλιση και καλύτερα αποτελέσματα γενίκευσης (Ioffe & Szegedy, 2015). Έπειτα, ακολουθούν τρία διαδοχικά πλήρως συνδεδεμένα στρώματα με 128,

64 και 32 νευρώνες αντίστοιχα που χρησιμοποιούν συνάρτηση ενεργοποίησης ReLU και τεχνική κανονικοποίησης L2. Η μέθοδος αυτή ονομάζεται εναλλακτικά τεχνική αποσάθρωσης - ελάττωσης βάρους (weight decay). Η λειτουργία της τεχνικής βασίζεται στον περιορισμό της μεταβολής των βαρών κατά τη διάρκεια της διαδικασίας εκπαίδευσης (Hinton, 2012). Έπειτα, ακολουθεί μια δομή Flatten η οποία μετατρέπει τις πολυδιάστατες αναπαραστάσεις σε μονοδιάστατες και τελικά ένα πλήρως συνδεδεμένο στρώμα που αποτελείται από έναν μοναδικό νευρώνα με σιγμοειδή συνάρτηση ενεργοποίησης. Μια απεικόνιση των επιμέρους στρωμάτων της αρχιτεκτονικής και του τρόπου διασύνδεσής τους παρέχεται στο Παράρτημα **B.2**. Παρακάτω δίνεται μια σύντομη επισκόπηση των στρωμάτων ως προς τις διαστάσεις εισόδου, εξόδου και τις παραμέτρους τους. Συνολικά περιέχονται 215.041 εκπαιδεύσιμες παράμετροι.

Πίνακας 5 - Αριθμός παραμέτρων της αρχιτεκτονικής Bidirectional LSTM ανά στρώμα

	Input	Output	Parameters
Conv1D	(768,1)	(764,128)	768
Max Pooling	(764,128)	(382,128)	0
Conv1D	(382,128)	(378,128)	82048
Max Pooling	(378,128)	(189,128)	0
Bidirectional LSTM	(189,128)	(189,128)	98816
Batch Normalization	(189,128)	(189,128)	512
Dense	(189,128)	(189,128)	16512
Dense	(189,128)	(189,64)	8256
Dense	(189,64)	(189,32)	2080
Flatten	(189,32)	(6048)	0
Dense	(6048)	(1)	6049

4.4.4 FakeBERT

Η συγκεκριμένη αρχιτεκτονική αξιοποιεί την παράλληλη επεξεργασία των δεδομένων εισόδου. Είναι αποτέλεσμα της μελέτης FakeBERT: Fake news detection in social media with a BERT-based deep learning approach (Goswami, Kaliyar, & Narang, 2021). Τα δεδομένα εισόδου τροφοδοτούν παράλληλα τρεις δομές ζευγών συνελικτικών στρωμάτων και στρωμάτων μέγιστης συσσώρευσης. Τα συνελικτικά στρώματα έχουν 128 φίλτρα το καθένα. Όμως σε κάθε δομή υπάρχει διαφορετικό μέγεθος πυρήνα των επιμέρους στρωμάτων που κυμαίνεται από 3 έως 5. Στη συνέχεια, τα αποτελέσματα των δομών συνδυάζονται. Έπειτα, ακολουθεί άλλη μια δομή με ζεύγος συνελικτικού στρώματος και στρώματος μέγιστης συσσώρευσης με 128 φίλτρα και μέγεθος πυρήνα 5. Μετά ακολουθούν δύο πλήρως συνδεδεμένα στρώματα με το πρώτο να περιέχει 384 νευρώνες και το δεύτερο 128. Παράλληλα, έχει εφαρμοστεί τεχνική Dropout και στα δύο προαναφερθέντα στρώματα σε ποσοστό 50%. Τελικά, ακολουθεί ένα πλήρως συνδεδεμένο στρώμα ενός νευρώνα με σιγμοειδή συνάρτηση ενεργοποίησης. Η συγκεκριμένη αρχιτεκτονική περιέχει 4.033.153 εκπαιδεύσιμες παραμέτρους. Η απεικόνισή του γίνεται στο Παράρτημα **B.3** και στη συνέχεια παρουσιάζονται κάποια στοιχεία σχετικά με τις διαστάσεις εισόδου, εξόδου και των αριθμό παραμέτρων για κάθε στρώμα του.

Πίνακας 6 - Αριθμός παραμέτρων της αρχιτεκτονικής FakeBERT ανά στρώμα

	Input	Output	Parameters
Conv1D	(768,1)	(768,128)	512
Conv1D	(768,1)	(768,128)	640
Conv1D	(768,1)	(768,128)	768
Max Pooling	(768,128)	(153,128)	0
Max Pooling	(768,128)	(153,128)	0
Max Pooling	(768,128)	(153,128)	0
Concatenate	[(153,128), (153,128), (153,128)]	(153,384)	0
Conv1D	(153,384)	(153,128)	245888
Max Pooling	(153,128)	(76,128)	0
Flatten	(76,128)	(9728)	0
Dense	(9728)	(384)	3735936
Dropout	(384)	(384)	0
Dense	(384)	(128)	49280
Dropout	(128)	(128)	0
Dense	(128)	(1)	129

4.4.5 CNN-L2 Regularization

Πρόκειται για αρχιτεκτονική η οποία έχει ακριβώς την ίδια διαστρωμάτωση και παραμετροποίηση με την CNN αρχιτεκτονική, που παρουσιάστηκε παραπάνω. Ωστόσο, η βασική διαφορά είναι η χρήση της τεχνικής κανονικοποιήσης ελάττωσης βαρών στα πλήρως συνδεδεμένα στρώματα. Αρχικά, υπάρχουν τρεις διαδοχικές δομές που αποτελούνται από ζεύγη συνελικτικών στρωμάτων ακολουθούμενα από στρώματα μέγιστης συσσώρευσης. Στη συνέχεια, γίνεται η διαμόρφωση της αναπαράστασης έτσι ώστε να αποδοθεί μονοδιάστατα. Έπειτα, υπάρχουν δύο πλήρως συνδεδεμένα στρώματα στα οποία εφαρμόζεται τεχνική Dropout και έχουν συνάρτηση ενεργοποίησης ReLU. Τέλος, χρησιμοποιείται ένα πλήρως συνδεδεμένο στρώμα αποτελούμενο από έναν νευρώνα και σιγμοειδή συνάρτηση ενεργοποίησης. Μια απεικόνιση της αρχιτεκτονικής που αναλύθηκε, δίνεται στο Παράρτημα **B.4**. Συνολικά υπάρχουν 255.233 εκπαιδεύσιμες παράμετροι και στη συνέχεια παρατίθεται μια σύντομη περιγραφή των στρωμάτων ως προς τις διαστάσεις εισόδου, εξόδου και αριθμού παραμέτρων.

Πίνακας 7 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN-L2 Regularization ανά στρώμα

	Input	Output	Parameters
Conv1D	(768,1)	(764,128)	768
Max Pooling	(764,128)	(152,128)	0
Conv1D	(152,128)	(148,128)	82048
Max Pooling	(148,128)	(29,128)	0
Conv1D	(29,128)	(25,128)	82048
Max Pooling	(25,128)	(5,128)	0
Flatten	(5,128)	(640)	0
Dense	(640)	(128)	82048
Dense	(128)	(64)	8256
Dense	(64)	(1)	65

4.4.6 LSTM

Η συγκεκριμένη αρχιτεκτονική αποτελεί μια απλοποιημένη μορφή του μοντέλου Bidirectional LSTM. Η διαστρωμάτωση των επιμέρους δομών του μοντέλου είναι όμοια και η παραμετροποίηση παραμένει ίδια. Ωστόσο, υπάρχουν δύο θεμελιώδεις διαφορές. Η πρώτη διαφορά είναι πως το στρώμα μακράς βραχύχρονης μνήμης (Long Short Term Memory) είναι μονής κατεύθυνσης, αυτό σημαίνει ότι εντοπίζονται εξαρτήσεις μόνο ως προς μία κατεύθυνση των δεδομένων. Η δεύτερη διαφορά είναι πως στα πλήρως συνδεδεμένα στρώματα δεν χρησιμοποιείται τεχνική ελάττωσης βαρών. Μια απεικόνιση της αρχιτεκτονικής παρουσιάζεται στο Παράρτημα **B.5** και παρακάτω παρουσιάζονται με συντομία οι διαστάσεις εισόδου, εξόδου και ο αριθμός παραμέτρων για τα επιμέρους στρώματα. Οι εκπαιδεύσιμες παράμετροι της αρχιτεκτονικής είναι 157.185.

Πίνακας 8 - Αριθμός παραμέτρων της αρχιτεκτονικής LSTM ανά στρώμα

	Input	Output	Parameter
Conv1D	(768,1)	(764,128)	768
Max Pooling	(764,128)	(382,128)	0
Conv1D	(382,128)	(378,128)	82048
Max Pooling	(378,128)	(189,128)	0
LSTM	(189,128)	(189,64)	49408
Batch Normalization	(189,64)	(189,64)	256
Dense	(189,64)	(189,128)	8320
Dense	(189,128)	(189,64)	8256
Dense	(189,64)	(189,32)	2080
Flatten	(189,32)	(6048)	0
Dense	(6048)	(1)	6049

4.4.7 CNN Light

Αναφορικά με την οργάνωση του μοντέλου, αρχικά υπάρχει μια δομή με ζεύγος ενός συνελικτικού στρώματος συνδυαστικά με ένα στρώμα μέγιστης

συσσώρευσης. Το συνελικτικό στρώμα έχει συνολικά 32 φίλτρα και μέγεθος πυρήνα ίσο με 3 και το μέγεθος πυρήνα του στρώματος μέγιστης συσσώρευσης είναι ίσο με 2. Έπειτα, γίνεται μετασχηματισμός της πολυδιάστατης αναπαράστασης σε μονοδιάστατη μορφή. Έπειτα, ακολουθεί ένα πλήρως συνδεδεμένο στρώμα με 64 νευρώνες και συνάρτηση ενεργοποίησης ReLU. Τέλος, η έξοδος παράγεται από πλήρως συνδεδεμένο στρώμα ενός νευρώνα με σιγμοειδή συνάρτηση ενεργοποίησης. Η απεικόνιση του μοντέλου δίνεται στο Παράρτημα **B.6** και παρακάτω δίνεται μια σύντομη παρουσίαση των διαστάσεων εισόδου, εξόδου και του αριθμού παραμέτρων ανά στρώμα. Συνολικά η αρχιτεκτονική έχει 784.641 εκπαιδεύσιμες παραμέτρους.

Πίνακας 9 - Αριθμός παραμέτρων της αρχιτεκτονικής CNN-Light ανά στρώμα

	Input	Output	Parameter
Conv1D	(768,1)	(766,32)	128
Max Pooling	(766,32)	(383,32)	0
Flatten	(383,32)	(12256)	0
Dense	(12256)	(64)	784448
Dense	(64)	(1)	65

4.4.8 Title – Text

Πρόκειται για μια αρχιτεκτονική που χρησιμοποιήθηκε για σύνολα δεδομένων που παρέχουν τίτλους μαζί με το περιεχόμενο των άρθρων. Είναι μια χρήσιμη προσέγγιση, για να μελετηθεί το αποτέλεσμα των διαδικασιών μάθησης όταν συνδυάζονται οι πληροφορίες δύο ειδών δειγμάτων. Η αρχιτεκτονική αναπτύχθηκε με βάση τη μελέτη με τίτλο Fake news detection using parallel BERT deep neural networks (Farokhian, Rafe, & Veisi, 2022). Τα εμφυτεύματα μιας διάστασης με 768 στοιχεία το καθένα εισάγονται σε ένα πλήρως συνδεδεμένο στρώμα 768 νευρώνων. Στη συνέχεια, ενώνονται σε ένα ενιαίο διάνυσμα μιας διάστασης με 1.536 στοιχεία. Έπειτα υπάρχει ένα στρώμα που εφαρμόζει τεχνική Dropout σε ποσοστό 50% και ένα τελευταίο πλήρως συνδεδεμένο στρώμα με έναν νευρώνα και σιγμοειδή συνάρτηση ενεργοποίησης. Παρακάτω, παρουσιάζονται οι διαστάσεις εισόδου, εξόδου και οι παράμετροι κάθε στρώματος. Συνολικά, η αρχιτεκτονική περιέχει 592.129 εκπαιδεύσιμες παραμέτρους και η απεικόνιση του παρουσιάζεται στο Παράρτημα **B.7**.

Πίνακας 10 - Αριθμός παραμέτρων της αρχιτεκτονικής Title-Text ανά στρώμα

	Input	Output	Parameters
Dense	(768)	(768)	590592
Concatenate	[(768), (768)]	(1536)	0
Dropout	(1536)	(1536)	0
Dense	(1536)	(1)	1537

5 Ανάλυση αποτελεσμάτων

Στο παρόν κεφάλαιο, για κάθε ξεχωριστό σύνολο δεδομένων, θα παρουσιαστούν τα αποτελέσματα των μοντέλων μετά τον έλεγχο της ικανότητάς τους για γενίκευση. Κάθε σύνολο δεδομένων, έχει εξεταστεί ως προς την περίπτωση ταξινόμησης σε δύο βαθμίδες αξιοπιστίας, έτσι ώστε να είναι δυνατή η σύγκριση της ανταπόκρισής τους στην προτεινόμενη μεθοδολογία.

5.1 ISOT

Παρακάτω, θα παρουσιαστούν αναλυτικά τα αποτελέσματα για κάθε αρχιτεκτονική στις διαφορετικές κατηγορίες δειγμάτων, που παρέχει το σύνολο δεδομένων ISOT. Υπενθυμίζεται πως το σύνολο δεδομένων περιέχει τίτλους, τα αντίστοιχα άρθρα και την σύντομη αναπαράσταση των άρθρων, για αντιμετώπιση περιπτώσεων που υπερβαίνουν το όριο λέξεων, που μπορεί να διαχειριστεί το DistilBERT σαν είσοδο.

5.1.1 Τίτλοι

Για κάθε αρχιτεκτονική παρουσιάζονται οι μετρικές με βάση τις οποίες αξιολογείται το μοντέλο μετά τη διαδικασία διασταυρωμένης επικύρωσης. Οι μετρικές αυτές είναι οι accuracy, recall, precision και f1 score. Παράλληλα, σε κάθε μια από τις περιπτώσεις παρουσιάζεται η απόκλιση μεταξύ των διαφορετικών αποτελεσμάτων που προέκυψαν, κατά τις επαναλήψεις της διαδικασίας επικύρωσης.

Πίνακας 11 - Αποτελέσματα εφαρμογής της μεθοδολογίας στους τίτλους του συνόλου δεδομένων ISOT

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,99 (+/- 0,00)	1,00 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)
Bidirectional	0,99 (+/- 0,00)	1,00 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)
FakeBERT	0,99 (+/- 0,00)	0,99 (+/- 0,01)	0,99 (+/- 0,01)	0,99 (+/- 0,00)
CNN-L2	0,99 (+/- 0,00)	1,00 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)
LSTM	0,99 (+/- 0,00)	1,00 (+/- 0,01)	0,99 (+/- 0,00)	0,99 (+/- 0,00)
CNN-Light	0,99 (+/- 0,00)	1,00 (+/- 0,00)	0,99 (+/- 0,01)	0,99 (+/- 0,00)

Από τον παραπάνω πίνακα φαίνεται πως υπάρχουν παρόμοια αποτελέσματα σε κάθε αρχιτεκτονική ως προς την απόδοση και τη δυνατότητα γενίκευσης.

5.1.2 Άρθρα

Στη συνέχεια, θα παρουσιαστεί η απόδοση των αρχιτεκτονικών, όταν πλέον η είσοδός τους είναι οι πρώτοι όροι των άρθρων, έως το σημείο που θα συμπληρωθεί το όριο των 512 λέξεων που τίθεται από το DistilBERT.

Πίνακας 12 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα άρθρα του συνόλου δεδομένων ISOT

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)
Bidirectional	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)
FakeBERT	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)
CNN-L2	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)
LSTM	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)
CNN-Light	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)	1,00 (+/- 0,00)

Παρατηρείται, πως εμφανίζονται καλύτερα αποτελέσματα συγκριτικά με την τροφοδότηση των αρχιτεκτονικών αποκλειστικά με δείγματα τίτλων. Το φαινόμενο αυτό εξηγείται λόγω της διαπίστωσης, πως τα άρθρα έχουν πολύ μεγαλύτερο μέσο μήκος από τους τίτλους και προσφέρουν καλύτερες αναπαραστάσεις προς μελέτη.

5.1.3 Συμπυκνωμένα άρθρα

Στα αποτελέσματα που παρουσιάζονται παρακάτω, έχει χρησιμοποιηθεί διαφορετική κατηγορία δειγμάτων. Πρόκειται για τα άρθρα, όπως προκύπτουν μετά από επεξεργασία του συνόλου δεδομένων. Συγκεκριμένα, έχει επιχειρηθεί να αφαιρεθούν οι περιττές προτάσεις και να διατηρηθούν μόνο εκείνες, που έχει αξία να αναλυθούν για να διαπιστωθεί η εγκυρότητα. Οι προτάσεις που τελικά επιλέγονται, ενώνονται με τη σειρά που εμφανίζονται στο κείμενο και συνθέτουν μια νέα κατηγορία δείγματος.

Πίνακας 13 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα συμπυκνωμένα άρθρα του συνόλου δεδομένων ISOT

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,99 (+/- 0,00)	0,98 (+/- 0,01)	1,00 (+/- 0,00)	0,99 (+/- 0,00)
Bidirectional	0,99 (+/- 0,00)	0,98 (+/- 0,01)	0,99 (+/- 0,01)	0,99 (+/- 0,00)
FakeBERT	0,99 (+/- 0,00)	0,98 (+/- 0,01)	1,00 (+/- 0,00)	0,99 (+/- 0,00)
CNN-L2	0,99 (+/- 0,00)	0,98 (+/- 0,01)	0,99 (+/- 0,01)	0,99 (+/- 0,00)
LSTM	0,99 (+/- 0,00)	0,98 (+/- 0,01)	1,00 (+/- 0,00)	0,99 (+/- 0,00)
CNN-Light	0,99 (+/- 0,00)	0,99 (+/- 0,01)	0,99 (+/- 0,01)	0,99 (+/- 0,00)

Από τα παραπάνω στοιχεία, είναι φανερό πως η απόδοση είναι αρκετά υψηλή, αλλά υστερεί ελάχιστα ως προς τις μετρικές recall και precision, συγκριτικά με την προηγούμενη κατηγορία δειγμάτων. Το ίδιο παρατηρείται και για τις υπόλοιπες μετρικές, όμως υπάρχει ομοιογένεια ως προς τις τιμές τους για κάθε διαφορετική αρχιτεκτονική (π.χ. accuracy:0.99 και f1-score:0.99). Ο λόγος που οι μετρικές είναι

μειωμένες εντοπίζεται στο γεγονός πως τα δείγματα περιγράφουν πολύ περισσότερη πληροφορία ενώ παράλληλα τα κείμενα έχουν αλλοιωθεί ως προς τη σύνταξή τους. Επίσης, τα δείγματα πλέον αποτελούν ένα σύνολο νοηματικά ασύνδετων προτάσεων έχοντας διατηρήσει μόνο τις σημαντικότερες και πιο χρήσιμες ακολουθίες.

5.2 FakeNewsNet

Στην παρούσα ενότητα, θα αναλυθούν τα αποτελέσματα των αρχιτεκτονικών στις κωδικοποιήσεις των δειγμάτων του συνόλου δεδομένων FakeNewsNet. Το συγκεκριμένο σύνολο δεδομένων με βάση την ανάλυση του προηγούμενου κεφαλαίου έχει υποστεί αύξηση των δειγμάτων στην κατηγορία μειοψηφίας.

Πίνακας 14 - Αποτελέσματα εφαρμογής της μεθοδολογίας στους τίτλους του συνόλου δεδομένων FakeNewsNet

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,97 (+/- 0,00)	0,98 (+/- 0,01)	0,97 (+/- 0,01)	0,97 (+/- 0,00)
Bidirectional	0,97 (+/- 0,01)	0,98 (+/- 0,01)	0,97 (+/- 0,01)	0,97 (+/- 0,01)
FakeBERT	0,97 (+/- 0,00)	0,97 (+/- 0,02)	0,97 (+/- 0,01)	0,97 (+/- 0,00)
CNN-L2	0,97 (+/- 0,00)	0,98 (+/- 0,01)	0,96 (+/- 0,01)	0,97 (+/- 0,00)
LSTM	0,97 (+/- 0,01)	0,98 (+/- 0,01)	0,97 (+/- 0,02)	0,97 (+/- 0,01)
CNN-Light	0,97 (+/- 0,00)	0,98 (+/- 0,01)	0,96 (+/- 0,01)	0,97 (+/- 0,00)

Από τα παραπάνω αποτελέσματα, φαίνεται πως όλες οι αρχιτεκτονικές που χρησιμοποιούνται συγκλίνουν ως προς τις τιμές των μετρικών τους. Τα αποτελέσματα δείχνουν, πως τα μοντέλα έχουν ικανοποιητική δυνατότητα γενίκευσης.

5.3 FakeNewsChallenge

Το σύνολο δεδομένων περιέχει τίτλους καθώς και τα άρθρα που τους συνοδεύουν. Ωστόσο, ο χαρακτηρισμός των δειγμάτων από τους δημιουργούς έχει προκύψει αξιολογώντας ταυτόχρονα τον τίτλο με το περιεχόμενο του άρθρου, ένα φαινόμενο που δεν εντοπίζεται στο ISOT. Άρα, δεν θα γίνει προσπάθεια να ταξινομηθούν τα δείγματα με βάση τους τίτλους μεμονωμένα, όπως συνέβη σε προηγούμενο σύνολο δεδομένων.

Συνδυάζοντας τις τέσσερις βαθμίδες αξιοπιστίας σε ζεύγη, προκύπτουν δύο βαθμίδες αξιοπιστίας. Συγκεκριμένα, ως αληθή θεωρούνται τα δείγματα των κατηγοριών agree και discuss, ενώ ως αναληθή τα δείγματα των κατηγοριών disagree και unrelated. Ωστόσο, η ενέργεια αυτή επιδεινώνει το έντονο πρόβλημα ανισορροπίας στην κατανομή των δειγμάτων σε κατηγορίες. Για κάποιους μελετητές, η έλλειψη συνάφειας μεταξύ τίτλου και άρθρου (unrelated), δεν υποδεικνύει ψευδή δείγματα, όμως αυτό είναι αμφιλεγόμενη πρακτική με βάση το θεωρητικό υπόβαθρο.

5.3.1 Άρθρα

Παρακάτω, παρουσιάζονται τα αποτελέσματα της προσέγγισης που χρησιμοποιήθηκε, με εφαρμογή στο περιεχόμενο των άρθρων.

Πίνακας 15 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα άρθρα του συνόλου δεδομένων FakeNewsChallenge ***δύο*** *βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0, 80(+/- 0,01)	0,34 (+/- 0,09)	0,70 (+/- 0,08)	0,45 (+/- 0,07)
Bidirectional	0,79 (+/- 0,03)	0,31 (+/- 0,13)	0,68 (+/- 0,06)	0,43 (+/- 0,14)
FakeBERT	0,81 (+/- 0,01)	0,34 (+/- 0,06)	0,81 (+/- 0,05)	0,47 (+/- 0,06)
CNN-L2	0,76 (+/- 0,02)	0,15 (+/- 0,28)	0,34 (+/- 0,56)	0,21 (+/- 0,36)
LSTM	0,79 (+/- 0,02)	0,43 (+/- 0,12)	0,63 (+/- 0,08)	0,51 (+/- 0,09)
CNN-Light	0,81 (+/- 0,01)	0,42 (+/- 0,07)	0,71 (+/- 0,08)	0,53 (+/- 0,04)

Παρατηρείται, πως σε όλα τα μοντέλα υπάρχουν μεγάλες τιμές accuracy και precision συγκριτικά με τις άλλες δύο μετρικές. Ωστόσο, οι χαμηλές τιμές recall και αντίστοιχα f1-score, δείχνουν πως λόγω της μεγάλης ανισότητας στην κατανομή των δειγμάτων, τα μοντέλα μπορούν να προβλέψουν με μεγάλη επιτυχία κυρίως την κατηγορία που υπερτερεί. Λόγω του συνδυασμού των κατηγοριών, η ανισότητα στην κατανομή των δειγμάτων εντείνεται. Ακόμα και με την εφαρμογή τεχνικών αύξησης των δειγμάτων της κατηγορίας μειοψηφίας, η ανισότητα δεν μπορεί να γεφυρωθεί αρκετά, ώστε να υπάρξουν ικανοποιητικά αποτελέσματα όπως συνέβη στην περίπτωση του συνόλου FakeNewsNet.

5.3.2 Συμπυκνωμένα άρθρα

Τα αποτελέσματα που παρουσιάζονται στη συνέχεια, αφορούν τα αρχικά άρθρα σε συμπυκνωμένη μορφή. Δεν παρατηρείται κάποια αξιόλογη μεταβολή συγκριτικά με την περίπτωση των μη συμπυκνωμένων άρθρων, εκτός από μικρές αλλαγές στις τιμές των μετρικών και των αντίστοιχων αποκλίσεων.

Πίνακας 16 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα συμπυκνωμένα άρθρα του συνόλου δεδομένων FakeNewsChallenge ***δύο*** *βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,80 (+/- 0,01)	0,36 (+/- 0,09)	0,70 (+/- 0,08)	0,47 (+/- 0,06)
Bidirectional	0,79 (+/- 0,03)	0,44 (+/- 0,18)	0,63 (+/- 0,12)	0,51 (+/- 0,10)
FakeBERT	0,80 (+/- 0,03)	0,29 (+/- 0,16)	0,78 (+/- 0,07)	0,42 (+/- 0,18)
CNN-L2	0,76 (+/- 0,02)	0,29 (+/- 0,32)	0,46 (+/- 0,46)	0,34 (+/- 0,36)
LSTM	0,80 (+/- 0,02)	0,36 (+/- 0,07)	0,70 (+/- 0,15)	0,47 (+/- 0,05)
CNN-Light	0,81 (+/- 0,01)	0,35 (+/- 0,07)	0,74 (+/- 0,05)	0,48 (+/- 0,06)

5.4 LIAR

Το συγκριμένο σύνολο δεδομένων όπως αναλύθηκε και σε προηγούμενο κεφάλαιο περιέχει έξι διαφορετικές βαθμίδες αξιοπιστίας. Ωστόσο, ο αριθμός των δειγμάτων του είναι ιδιαίτερα περιορισμένος, γεγονός που λειτουργεί αρνητικά ως προς τα αναμενόμενα αποτελέσματα.

Συνδυάζοντας σε μια κατηγορία τις δύο βαθμίδες που αφορούν τα εγκυρότερα δείγματα και σε μια δεύτερη τις υπόλοιπες τέσσερις, που αφορούν τα λιγότερο αξιόπιστα δείγματα, δίνεται η δυνατότητα για δυαδική ταξινόμηση. Τα αποτελέσματα της διαδικασίας φαίνονται παρακάτω.

*Πίνακας 17 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα δείγματα του συνόλου δεδομένων LIAR **δύο** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,80 (+/- 0,01)	0,68 (+/- 0,10)	0,73 (+/- 0,03)	0,70 (+/- 0,04)
Bidirectional	0,80 (+/- 0,02)	0,74 (+/- 0,10)	0,71 (+/- 0,05)	0,72 (+/- 0,02)
FakeBERT	0,80 (+/- 0,02)	0,68 (+/- 0,09)	0,73 (+/- 0,03)	0,70 (+/- 0,04)
CNN-L2	0,80 (+/- 0,01)	0,75 (+/- 0,13)	0,70 (+/- 0,07)	0,72 (+/- 0,04)
LSTM	0,79 (+/- 0,03)	0,67 (+/- 0,09)	0,72 (+/- 0,09)	0,69 (+/- 0,03)
CNN-Light	0,79 (+/- 0,02)	0,61 (+/- 0,13)	0,77 (+/- 0,06)	0,67 (+/- 0,07)

Τα παραπάνω αποτελέσματα είναι ικανοποιητικά ως προς τη δυνατότητα του μοντέλου σε γενικεύσεις, όταν χρειάζεται να αντιμετωπίσει νέα δεδομένα. Οι τιμές των μετρικών αξιολόγησης είναι πολύ ικανοποιητικές και δεν εμφανίζονται πολύ υψηλές διακυμάνσεις .

5.5 PHEME

Το σύνολο δεδομένων PHEME περιέχει τρεις κατηγορίες ως προς τις οποίες ταξινομεί το περιεχόμενο δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης. Παρακάτω, θα γίνει παρουσίαση των αποτελεσμάτων στην περίπτωση που τα μοντέλα ταξινομούν τα δείγματα σε δύο κατηγορίες.

Διαγράφοντας την μια από τις τρεις βαθμίδες αξιοπιστίας, που αντιπροσωπεύει δείγματα για τα οποία δεν είναι δυνατό να προσδιοριστεί η εγκυρότητά τους, περισσεύουν δύο κατηγορίες με αληθή και ψευδή δείγματα. Τα αποτελέσματα της διαδικασίας επικύρωσης σε κάθε αρχιτεκτονική παρουσιάζονται παρακάτω.

*Πίνακας 18 - Αποτελέσματα εφαρμογής της μεθοδολογίας στα δείγματα του συνόλου δεδομένων PHEME **δύο** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,98 (+/- 0,02)	0,99 (+/- 0,01)	0,97 (+/- 0,03)	0,98 (+/- 0,02)
Bidirectional	0,97 (+/- 0,01)	0,96 (+/- 0,04)	0,99 (+/- 0,02)	0,97 (+/- 0,01)
FakeBERT	0,98 (+/- 0,01)	0,99 (+/- 0,01)	0,98 (+/- 0,02)	0,98 (+/- 0,01)
CNN-L2	0,98 (+/- 0,02)	0,98 (+/- 0,03)	0,98 (+/- 0,02)	0,98 (+/- 0,01)
LSTM	0,98 (+/- 0,02)	0,98 (+/- 0,05)	0,98 (+/- 0,02)	0,98 (+/- 0,02)
CNN-Light	0,98 (+/- 0,02)	0,99 (+/- 0,01)	0,97 (+/- 0,03)	0,98 (+/- 0,02)

Τα αποτελέσματα είναι πολύ ενθαρρυντικά και δεν εμφανίζονται ιδιαίτερα μεγάλες διακυμάνσεις στις τιμές των μετρικών ανά μοντέλο. Ακόμα, τα μοντέλα φαίνεται πως παρέχουν ικανοποιητική δυνατότητα γενίκευσης.

5.6 Σύγκριση βέλτιστων αποτελεσμάτων ανά σύνολο δεδομένων

Το σύνολο δεδομένων ISOT είχε την καλύτερη ανταπόκριση στις προτεινόμενες αρχιτεκτονικές κατά τη διαδικασία δυαδικής ταξινόμησης, όπως φαίνεται και στον παρακάτω πίνακα. Τα σύνολα δεδομένων που έπονται, ως προς τα τελικά αποτελέσματα, είναι τα PHEME και FakeNewsNet. Τα δύο τελευταία σύνολα δεδομένων φαίνεται πως έχουν πολύ παρεμφερείς τιμές μετρικών, παρά την μικρή ανισορροπία που μπορεί να εμφανίζουν στην κατανομή των δειγμάτων σε κατηγορίες.

Επιπλέον, ενώ υπήρξαν ενθαρρυντικές τιμές στις μετρικές των περισσότερων πειραμάτων, στα σύνολα δεδομένων FakeNewsChallenge και LIAR αναδείχθηκε ένα βασικό μειονέκτημα μεθόδων κωδικοποίησης κειμένου όμοιων του DistilBERT. Το μειονέκτημα, αφορά την μεγάλη εξάρτηση της κωδικοποίησης από τα χαρακτηριστικά του συνόλου δεδομένων. Για παράδειγμα, κατά τη διαδικασία ταξινόμησης σε δύο κατηγορίες, για το σύνολο δεδομένων FakeNewsChallenge, η έντονη ανισορροπία που υπήρχε επιδεινώνεται. Οι συνέπειες, εκδηλώνονται ως μειωμένες τιμές μετρικών επίδοσης και αυξημένες τιμές απόκλισης. Το φαινόμενο αυτό, προκύπτει λόγω της μεγάλης διαφοράς του αριθμού δειγμάτων ανάμεσα στην κατηγορία πλειοψηφίας και μειοψηφίας, μετά από επεξεργασία του ήδη ιδιόμορφου αρχικού συνόλου δεδομένων FakeNewsChallenge.

Η παρουσίαση των παραπάνω τιμών είχε δύο σκοπούς. Αρχικά, εξετάστηκε η αποτελεσματική λειτουργία της προτεινόμενης μεθόδου. Δηλαδή, αποδείχτηκε πως η δημιουργία εμφυτευμάτων με το DistilBERT και η επακόλουθη επεξεργασία τους από τις επιμέρους αρχιτεκτονικές όντως προσφέρει πολύ ικανοποιητική απόδοση. Ο δεύτερος σκοπός ήταν να εξεταστεί η ανταπόκριση των συνόλων δεδομένων στην προτεινόμενη μέθοδο. Πιο συγκεκριμένα, παρακάτω γίνεται σύγκριση των βέλτιστων αποτελεσμάτων στα σύνολα δεδομένων ανεξαρτήτως τύπου δείγματος, αρχιτεκτονικής και με ταξινόμηση δειγμάτων ως προς δύο βαθμίδες αξιοπιστίας.

Πίνακας 19 - Σύγκριση βέλτιστων αποτελεσμάτων ανά σύνολο δεδομένων ανεξαρτήτως του τύπου δείγματος

Dataset	**Αρχιτεκτονική**	**Τύπος δείγματος**	**Accuracy**	**Recall**	**Precision**	**F1-score**
ISOT	FakeBERT	Άρθρα	**1,00 (+/- 0,00)**	**1,00 (+/- 0,00)**	**1,00 (+/- 0,00)**	**1,00 (+/- 0,00)**
FakeNewsNet	CNN	Τίτλοι	0,97 (+/- 0,00)	0,98 (+/- 0,01)	0,97 (+/- 0,01)	0,97 (+/- 0,00)
FNC	FakeBERT	Άρθρα	0,81 (+/- 0,01)	0,34 (+/- 0,06)	0,81 (+/- 0,05)	0,47 (+/- 0,06)
LIAR	FakeBERT	Σύντομες Δηλώσεις	0,80 (+/- 0,02)	0,68 (+/- 0,09)	0,73 (+/- 0,03)	0,70 (+/- 0,04)
PHEME	FakeBERT	Δημοσιεύσεις Twitter	0,98 (+/- 0,01)	0,99 (+/- 0,01)	0,98 (+/- 0,02)	0,98 (+/- 0,01)

Εκτός των όσων συμπερασμάτων αναφέρθηκαν στην παρούσα ενότητα, από τον παραπάνω πίνακα προκύπτει άλλη μια διαπίστωση. Ειδικότερα, συγκρίνοντας τα βέλτιστα αποτελέσματα, παρατηρείται πως συνήθως προκύπτουν από το μοντέλο FakeBERT. Ως βέλτιστα, θεωρούνται τα αποτελέσματα τα οποία προσφέρουν τις μέγιστες τιμές μετρικών με τη μικρότερη δυνατή απόκλιση. Στο παρόν κεφάλαιο, δεν έγινε αναφορά στη χρήση της αρχιτεκτονικής Title-Text, καθώς δεν παρέχουν όλα τα σύνολα δεδομένων τίτλους αλλά και άρθρα.

6 Σύγκριση αποτελεσμάτων και σχολιασμός

Στο προηγούμενο κεφάλαιο, παρουσιάστηκαν τα αποτελέσματα που αφορούν μόνο δύο βαθμίδες αξιοπιστίας (Fake/True) ανά σύνολο δεδομένων και μελετήθηκε κάθε τύπος δείγματος ξεχωριστά. Η διαδικασία αυτή πραγματοποιήθηκε, για να εξεταστεί η απόδοση της εφαρμογής της μεθόδου, όταν σε κάθε περίπτωση γίνεται διάκριση αποκλειστικά ανάμεσα σε ψευδή και αληθή δείγματα.

Στο παρόν κεφάλαιο, θα γίνει προσπάθεια σύγκρισης των αποτελεσμάτων της προτεινόμενης προσέγγισης με αποτελέσματα άλλων μελετών, που αξιοποιούν όμως και συμβατικές μεθόδους κωδικοποίησης κειμένου. Για να έχει ουσιαστικό νόημα η οποιαδήποτε σύγκριση, είναι αναγκαίο να μελετηθεί η απόδοση της προτεινόμενης μεθοδολογίας, υπό τις συνθήκες που ορίζονται από την εκάστοτε έρευνα. Στη συνέχεια του κεφαλαίου, θα παρουσιαστούν τα συμπεράσματα που προκύπτουν από την παρούσα εργασία καθώς και κάποιες μελλοντικές προεκτάσεις.

6.1 Σύγκριση με προσεγγίσεις συμβατικών τεχνικών κωδικοποίησης

Είναι πιθανό μια μελέτη να εξετάζει ένα σύνολο δεδομένων με ταυτόχρονη επεξεργασία τίτλων και άρθρων. Για να είναι αποδοτική η σύγκριση, τα δείγματα δεν θα πρέπει να εξεταστούν μεμονωμένα όπως συνέβη στο προηγούμενο κεφάλαιο. Οπότε, είναι αναγκαίο να παρουσιαστούν τα αποτελέσματα της προτεινόμενης μεθόδου όταν τα εμφυτεύματα τίτλων και άρθρων τροφοδοτούν μια αρχιτεκτονική που επεξεργάζεται παράλληλα τις δύο κατηγορίες δειγμάτων. Από τις προτεινόμενες αρχιτεκτονικές του 4^{ου} κεφαλαίου, εκείνη που πετυχαίνει τη συνδυαστική μελέτη τίτλων και άρθρων είναι η Title-Text.

Εναλλακτικά, αν μια μελέτη εξετάζει την ταξινόμηση δειγμάτων σε πολλαπλές κατηγορίες είναι ανάγκη να συμβεί το ίδιο και για την μέθοδο που προτείνεται στο πλαίσιο της παρούσας εργασίας. Στο προηγούμενο κεφάλαιο, τα σύνολα δεδομένων PHEME, FakeNewsChallenge και LIAR υπέστησαν τροποποιήσεις, ώστε να διακρίνονται τα δείγματα αποκλειστικά ως αληθή ή ψευδή. Στο παρόν κεφάλαιο, θα εξεταστεί η εφαρμογή της προτεινόμενης μεθοδολογίας όταν τα δείγματα ταξινομούνται στις βαθμίδες που προϋπήρχαν.

6.1.1 ISOT

Ξεκινώντας με το σύνολο δεδομένων ISOT, θα παρουσιαστούν κάποια ενδεικτικά αποτελέσματα προσεγγίσεων με χρήση συμβατικών μεθόδων κωδικοποίησης κειμένου. Ειδικότερα, θα χρησιμοποιηθούν τα αποτελέσματα της μελέτης με τίτλο Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques (Ahmed, Traore, & Saad, 2017). Στην προαναφερθείσα έρευνα, χρησιμοποιώντας συνδυαστικά τα άρθρα μαζί με τον τίτλο, έγινε

κωδικοποίηση με τη μέθοδο TF-IDF. Στη συνέχεια, μεταξύ άλλων μηχανισμών έγινε χρήση του αλγορίθμου k κοντινότερων γειτόνων (kNN), του μηχανισμού διανυσμάτων υποστήριξης (SVM) και του αλγορίθμου γραμμικής λογιστικής παλινδρόμησης (logistic regression).

Τα αποτελέσματα της μελέτης που αναφέρθηκε, αφορούν μόνο τη μετρική accuracy και είναι τα μέγιστα που προέκυψαν ανά διαδικασία μάθησης. Επίσης, οι υπόλοιπες μέθοδοι που αξιοποιήθηκαν, δεν ξεπέρασαν την τιμή 0,89 στη μελέτη που χρησιμοποιήθηκε ως σημείο αναφοράς. Όμως, η προτεινόμενη αρχιτεκτονική Title – Text, δίνει σημαντικά μεγαλύτερη τιμή accuracy, όπως άλλωστε φαίνεται και παρακάτω.

Πίνακας 20 – Αποτελέσματα συνόλου δεδομένων ISOT με προσεγγίσεις άλλων μελετητών

Αρχιτεκτονική	**Accuracy**
SVM	0,86
kNN	0,83
Linear Regression	0,89

Ο λόγος που χρησιμοποιείται για σύγκριση αποκλειστικά η αρχιτεκτονική Title-Text είναι διότι η εξεταζόμενη μελέτη στην οποία γίνεται αναφορά, μελετά τα άρθρα συνδυαστικά με τους τίτλους. Αρχικά, θα παρουσιαστεί η περίπτωση όπου το μοντέλο δέχεται στην είσοδό του, τους κωδικοποιημένους τίτλους και τα κωδικοποιημένα άρθρα.

Πίνακας 21 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα άρθρα του συνόλου δεδομένων ISOT

Αρχιτεκτονική	**Accuracy**	**Recall**	**Precision**	**F1-score**
Title-Text	0,99 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)

Έπειτα, πραγματοποιήθηκε η ίδια διαδικασία για να μελετηθεί η περίπτωση που το μοντέλο δέχεται στην είσοδό του κωδικοποιήσεις τίτλων συνδυαστικά με τις αντίστοιχες κωδικοποιήσεις συμπυκνωμένου κειμένου.

Πίνακας 22 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα συμπυκνωμένα άρθρα του συνόλου δεδομένων ISOT

Αρχιτεκτονική	**Accuracy**	**Recall**	**Precision**	**F1-score**
Title-Text	0,99 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)	0,99 (+/- 0,00)

6.1.2 FakeNewsNet

Στη συνέχεια, θα παρουσιαστεί σύντομη σύγκριση των αποτελεσμάτων του συνόλου δεδομένων FakeNewsNet με προηγούμενες έρευνες. Στη μελέτη με τίτλο FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media (Shu K. , Mahudeswaran, Wang, Lee, & Liu, 2020) προτάθηκε μια σύνθετη προσέγγιση. Το σύνολο FakeNewsNet αποτελείται από δύο επιμέρους υποσύνολα που περιέχουν δείγματα

προερχόμενα από τους οργανισμούς GossipCop και PolitiFact. Στην προσέγγιση της συγκεκριμένης έρευνας, εξετάστηκαν ξεχωριστά τα υποσύνολα και παράχθηκαν διαφορετικά αποτελέσματα. Στην παρούσα εργασία όμως, τα δύο υποσύνολα ενοποιήθηκαν κατά την πειραματική διαδικασία.

Η προσέγγιση που προτάθηκε αξιοποιεί πληροφορία που παρέχεται από τα μεταδεδομένα των χρηστών του Twitter συνδυαστικά με το κείμενο. Για την κωδικοποίηση του κειμένου χρησιμοποιείται αρχιτεκτονική βασισμένη σε μηχανισμούς μακράς βραχύχρονης μνήμης (LSTM). Οι κωδικοποιήσεις προωθούνται ως είσοδος σε μηχανισμούς διανυσμάτων υποστήριξης, λογιστικής παλινδρόμησης, Naive-Bayes και μοντέλα βαθιάς μάθησης όπως συνελικτικά νευρωνικά μοντέλα (CNNs: Convolutional Neural Networks).

Πίνακας 23 – Αποτελέσματα συνόλου δεδομένων FakeNewsNet με προσεγγίσεις άλλων μελετητών

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
SVM (PolitiFact)	0,580	0,717	0,611	0,659
SVM (GossipCop)	0,470	0,451	0,462	0,456
Logistic Regression (PolitiFact)	0,642	0,543	0,757	0,633
Logistic Regression (GossipCop)	0,822	0,722	0,897	0,799
Naive Bayes (PolitiFact)	0,617	0,630	0,674	0,651
Naive Bayes (GossipCop)	0,704	0,765	0,735	0,798
CNN (PolitiFact)	0,629	0,456	0,807	0,583
CNN (GossipCop)	0,703	0,623	0,789	0,699

Με βάση τις παραπάνω τιμές των μετρικών και τα αποτελέσματα που παρουσιάζονται στον **Πίνακας 14** είναι φανερό πως η προτεινόμενη μέθοδος εφαρμοζόμενη στο σύνολο δεδομένων παρέχει καλύτερες επιδόσεις. Ο λόγος που αξιοποιούνται τα ίδια αποτελέσματα με το 5^{ο} κεφάλαιο είναι πως δεν υπήρξε κάποια ανάγκη για τροποποιήσεις ως προς τον τύπο των δειγμάτων ή την κατανομή τους σε βαθμίδες αξιοπιστίας.

6.1.3 FakeNewsChallenge

Τα υπόλοιπα 3 σύνολα δεδομένων, επειδή έχουν δείγματα που ταξινομούνται σε παραπάνω από δύο κατηγορίες, εμφανίζονται στην πλειοψηφία των μελετών με βάση τις αρχικές βαθμίδες εγκυρότητας που παρέχουν. Στη συνέχεια, θα γίνει σύντομη παρουσίαση των αποτελεσμάτων του FakeNewsChallenge (FNC) συγκριτικά με προγενέστερες έρευνες. Στη μελέτη με τίτλο Modeling the Fake News Challenge as a Cross-Level Stance Detection Task (Conforti, Pilehvar, & Collier, 2019) έγινε ταξινόμηση των δειγμάτων σε πολλαπλές κατηγορίες. Αγνοώντας τα δείγματα της κατηγορίας μη σχετιζόμενων άρθρων με τους αντίστοιχους τίτλους (unrelated), εφαρμόστηκαν δύο βασικές προσεγγίσεις. Η πρώτη, αφορά τη χρήση μοντέλου perceptron πολλαπλών στρωμάτων (MLP) και η δεύτερη αφορά στρώματα μακράς βραχύχρονης μνήμης δύο κατευθύνσεων (Bi-LSTM), συνδυαστικά με

μηχανισμούς self-attention. Επίσης, οι τίτλοι μελετώνται συνδυαστικά με τα άρθρα σαν μια κατηγορία δείγματος.

Παρακάτω, θα παρουσιαστούν τα αποτελέσματα των μεθόδων που προέκυψαν από την προαναφερθείσα μελέτη. Είναι φανερό, πως η αρχιτεκτονική Title-Text που προτάθηκε στην παρούσα εργασία παρέχει καλύτερα αποτελέσματα.

Πίνακας 24 - Αποτελέσματα συνόλου δεδομένων FakeNewsChallenge με προσεγγίσεις άλλων μελετητών

Αρχιτεκτονική	**Recall**	**Precision**	**F1-score**
MLP	0,361	0,388	0,367
Bi-LSTM	0,503	0,505	0,486

Όπως συνέβη και στην περίπτωση του συνόλου δεδομένων ISOT, η μελέτη στην οποία γίνεται αναφορά, εξετάζει ταυτόχρονα τους τίτλους συνδυαστικά με τα άρθρα. Ακόμα, το σύνολο δεδομένων έχει αναλυθεί ως προς τρεις βαθμίδες αξιοπιστίας εξαιρώντας την κατηγορία των unrelated δειγμάτων. Πρόκειται για δείγματα στα οποία δεν εμφανίζεται σύνδεση μεταξύ του τίτλου και του άρθρου. Οπότε, για να έχει νόημα η σύγκριση, θα πρέπει να χρησιμοποιηθεί η αρχιτεκτονική Title-Text καθώς και να διατηρηθούν όλες οι βαθμίδες αξιοπιστίας που προϋπήρχαν.

Ο λόγος που επιλέγεται να διατηρηθούν όλες οι κατηγορίες, σε αντίθεση με την μελέτη στην οποία γίνεται αναφορά, εντοπίζεται στο γεγονός πως με βάση το 1ο κεφάλαιο η απουσία σύνδεσης μεταξύ τίτλου και άρθρου (κατηγορία unrelated) είναι ενδεικτική ψευδούς δείγματος. Η συνθήκη αυτή, οδηγεί σε μειωμένη αποτελεσματικότητα της προτεινόμενης μεθόδου, συγκριτικά με την περίπτωση χρήσης τριών βαθμίδων αξιοπιστίας, ωστόσο τα αποτελέσματα είναι ιδιαιτέρως ικανοποιητικά.

Παρακάτω, παρουσιάζονται αναλυτικότερα τα αποτελέσματα της προτεινόμενης αρχιτεκτονικής που μελετά συνδυαστικά τους τίτλους και τα άρθρα. Αρχικά, θα παρουσιαστεί η περίπτωση, όπου το μοντέλο δέχεται στην είσοδό του παράλληλα τον κωδικοποιημένο τίτλο και το κωδικοποιημένο άρθρο. Η ταξινόμηση των δειγμάτων γίνεται ως προς τέσσερις κατηγορίες.

*Πίνακας 25 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα άρθρα του συνόλου δεδομένων FakeNewsChallenge **τεσσάρων** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	**Accuracy**	**Recall**	**Precision**	**F1-score**
Title-Text	0,76 (+/- 0,00)	0,76 (+/- 0,00)	0,75 (+/- 0,00)	0,69 (+/- 0,00)

Έπειτα, πραγματοποιήθηκε η ίδια διαδικασία για να μελετηθεί η περίπτωση που το μοντέλο δέχεται στην είσοδό του κωδικοποιήσεις τίτλων, συνδυαστικά με τις αντίστοιχες κωδικοποιήσεις συμπυκνωμένου κειμένου. Παρατηρείται πως τα αποτελέσματα και στις δύο περιπτώσεις είναι όμοια.

*Πίνακας 26 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στους τίτλους και τα συμπυκνωμένα άρθρα του συνόλου δεδομένων FakeNewsChallenge **τεσσάρων** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
Title-Text	0,76 (+/- 0,00)	0,76 (+/- 0,00)	0,75 (+/- 0,00)	0,69 (+/- 0,00)

6.1.4 LIAR

Η μελέτη που εισήγαγε το σύνολο LIAR (William Yang, 2017), προσφέρει κάποια αποτελέσματα ταξινόμησης που προέκυψαν με συμβατικές μεθόδους κωδικοποίησης. Οι προσεγγίσεις που χρησιμοποιήθηκαν, αφορούσαν κωδικοποίηση των δειγμάτων κειμένου με την τεχνική Word2Vec. Οι διαδικασίες μάθησης περιλάμβαναν μηχανισμούς διανυσμάτων υποστήριξης (SVMs), χρήση αλγορίθμου λογιστικής παλινδρόμησης (Logistic Regression), δικτύων βασισμένων σε στρώματα μακράς βραχύχρονης μνήμης δύο κατευθύνσεων (Bi-LSTMs) και δικτύων βασισμένων σε συνελικτικά στρώματα (CNNs).

Παρακάτω, παρουσιάζονται τα αποτελέσματα που προέκυψαν από την προαναφερθείσα μελέτη. Ωστόσο, εφαρμόζοντας την προτεινόμενη μέθοδο ως προς πολλαπλές κατηγορίες στο σύνολο δεδομένων LIAR, τα τελικά αποτελέσματα είναι αρκετά καλύτερα.

Πίνακας 27 – Αποτελέσματα συνόλου δεδομένων LIAR με προσεγγίσεις άλλων μελετητών

Αρχιτεκτονική	Accuracy
SVMs	0,255
Logistic Regression	0,247
Bi-LSTMs	0,233
CNNs	0,270

Αντίθετα με το 5^{ο} κεφάλαιο, στη συγκεκριμένη περίπτωση χρειάστηκε να εφαρμοστεί η προτεινόμενη μεθοδολογία ως προς τις έξι βαθμίδες αξιοπιστίας που υπήρχαν αρχικά στο σύνολο δεδομένων, με βάση τη μελέτη στην οποία γίνεται αναφορά. Αναλυτικότερα, τα αποτελέσματα κάθε αρχιτεκτονικής, που δέχεται μόνο ένα τύπο δείγματος ως είσοδο, παρουσιάζονται παρακάτω.

*Πίνακας 28- Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στα δείγματα του συνόλου δεδομένων LIAR **έξι** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,44 (+/- 0,00)	0,44 (+/- 0,00)	0,45 (+/- 0,00)	0,44 (+/- 0,00)
Bidirectional	0,46 (+/- 0,00)	0,46 (+/- 0,00)	0,48 (+/- 0,00)	0,46 (+/- 0,00)
FakeBERT	0,45 (+/- 0,00)	0,45 (+/- 0,00)	0,46 (+/- 0,00)	0,45 (+/- 0,00)
CNN-L2	0,42 (+/- 0,00)	0,42 (+/- 0,00)	0,42 (+/- 0,00)	0,40 (+/- 0,00)
LSTM	0,41 (+/- 0,00)	0,41 (+/- 0,00)	0,42 (+/- 0,00)	0,41 (+/- 0,00)
CNN-Light	0,44 (+/- 0,00)	0,44 (+/- 0,00)	0,45 (+/- 0,00)	0,44 (+/- 0,00)

Τα αποτελέσματα της προτεινόμενης μεθόδου είναι ανώτερα, συγκριτικά με τη μελέτη που αναφέρθηκε αρχικά. Ωστόσο, παρατηρείται πως η προτεινόμενη μεθοδολογία παρέχει πολύ μικρότερες τιμές μετρικών συγκριτικά με περιπτώσεις άλλων συνόλων δεδομένων (π.χ. ISOT). Το γεγονός αυτό δεν είναι παράξενο, αφού το σύνολο δεδομένων LIAR παρέχει έξι βαθμίδες αξιοπιστίας και ταυτόχρονα παρουσιάζει ανισορροπία. Ένας ομοιόμορφα τυχαίος ταξινομητής, σαν μέτρο σύγκρισης, αναμένεται να έχει ακρίβεια που δεν ξεπερνά το 20%, ενώ στην περίπτωση δύο βαθμίδων αξιοπιστίας το ποσοστό αυτό αυξάνεται στο 50%. Συνεπώς το μικρότερο ποσοστό στα παραπάνω αποτελέσματα είναι ενδεικτικό αποκλειστικά των ιδιαίτερων χαρακτηριστικών του συνόλου δεδομένων.

6.1.5 PHEME

Συνεχίζοντας με το σύνολο δεδομένων PHEME, θα γίνει σύγκριση της προτεινόμενης μεθόδου με τα αποτελέσματα της μελέτης με τίτλο A Novel Rumor Detection Method Based on Non-Consecutive Semantic Features and Comment Stance (Zhu, Gensheng, Sheng, & Xuejian, 2023). Η μελέτη αυτή, συνδυάζει τα χαρακτηριστικά του περιεχομένου των δημοσιεύσεων, των δημιουργών και της αλληλεπίδρασής των δημοσιεύσεων με τους χρήστες των μέσων κοινωνικής δικτύωσης.

Μεταξύ άλλων μεθόδων, χρησιμοποιήθηκαν προσεγγίσεις όπως ο μηχανισμός διανυσμάτων υποστήριξης, δίκτυα που συνδυάζουν συνελικτικά στρώματα με στρώματα μακράς βραχύχρονης μνήμης και δίκτυα που συνδυάζουν στρώματα μακράς βραχύχρονης μνήμης με μηχανισμούς προσοχής. Τα καλύτερα αποτελέσματα στη συγκεκριμένη μελέτη προέκυψαν από μια υβριδική μέθοδο με ονομασία SFCR, που συνδυάζει τα μεταδεδομένα του δημιουργού και τη χρήση μιας εκδοχής του BERT, με την ονομασία StanceBERTa. Παρακάτω, παρουσιάζονται τα αποτελέσματα της προαναφερθείσας μελέτης. Γίνεται αναφορά αποκλειστικά στην περίπτωση ταξινόμησης σε πολλαπλές κατηγορίες.

Πίνακας 29 - Αποτελέσματα συνόλου δεδομένων PHEME με προσεγγίσεις άλλων μελετητών

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
SVM	0,783	0,731	0,692	0,711
LSTM-CNN	0,804	0,811	0,801	0,806
LSTM-Attention	0,830	0,816	0,823	0,819
SFCR	0,851	0,854	0,842	0,848

Αντίθετα με το 5^{ο} κεφάλαιο, στη συγκεκριμένη περίπτωση χρειάστηκε να εφαρμοστεί η προτεινόμενη μεθοδολογία ως προς τις τρεις βαθμίδες αξιοπιστίας που υπήρχαν αρχικά στο σύνολο δεδομένων, με βάση τη μελέτη στην οποία γίνεται αναφορά. Κάθε μοντέλο, δέχεται μόνο ένα είδος δείγματος στην είσοδό του και εμφανίζει αρκετά ικανοποιητική απόδοση. Παρατηρείται, πως εμφανίζονται αρκετά πιο ικανοποιητικά αποτελέσματα με χρήση της προτεινόμενης μεθόδου.

*Πίνακας 30 - Αποτελέσματα εφαρμογής της προτεινόμενης μεθοδολογίας στα δείγματα του συνόλου δεδομένων PHEME **τριών** βαθμίδων αξιοπιστίας*

Αρχιτεκτονική	Accuracy	Recall	Precision	F1-score
CNN	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)
Bidirectional	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)
FakeBERT	0,95 (+/- 0,00)	0,95 (+/- 0,00)	0,96 (+/- 0,00)	0,95 (+/- 0,00)
CNN-L2	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)
LSTM	0,95 (+/- 0,00)	0,95 (+/- 0,00)	0,96 (+/- 0,00)	0,95 (+/- 0,00)
CNN-Light	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)	0,96 (+/- 0,00)

6.2 Συμπεράσματα

Με βάση τα αποτελέσματα των επιμέρους συγκρίσεων του παρόντος κεφαλαίου, είναι φανερό πως η μέθοδος ταξινόμησης που προτάθηκε είναι πολύ ικανοποιητική ως προς την απόδοσή της. Σε όλα τα σύνολα δεδομένων που αναλύθηκαν, παρατηρείται πως η κωδικοποίηση των δειγμάτων με χρήση του DistilBERT υπερτερεί έναντι προγενέστερων μεθόδων αναπαράστασης κειμένου. Επιπλέον, υπήρξαν περιπτώσεις που ακόμα και η αξιοποίηση μεθόδων κωδικοποίησης κειμένου με χρήση μηχανισμών προσοχής δεν κατέληξε σε αποτελέσματα ανώτερα της προτεινόμενης μεθοδολογίας.

Η διαπίστωση αυτή είναι πολύ ενθαρρυντική και αποδεικνύει δύο βασικές υποθέσεις. Αρχικά, αποδεικνύεται πως η μέθοδος είναι αποτελεσματική και μπορεί όντως να συμβάλλει στην πρόβλεψη της εγκυρότητας των δειγμάτων. Μάλιστα, η αξιοπιστία των μοντέλων ενισχύεται μέσα από τη διαδικασία της διασταυρωμένης επικύρωσης, που εξετάζει περαιτέρω την ικανότητά τους για γενίκευση. Επίσης, αποδεικνύεται πως η κωδικοποίηση κειμένου με βάση το πλαίσιο χρήσης των λέξεων είναι πολύ πιο αποτελεσματική από συμβατικές μεθόδους κωδικοποίησης. Το γεγονός αυτό είναι λογικό με βάση το θεωρητικό υπόβαθρο που παρουσιάστηκε, αφού γίνεται αποτύπωση πιο σύνθετων σημασιολογικών στοιχείων του κειμένου και τα εμφυτεύματα δεν είναι στατικά. Δηλαδή, πλέον οι αναπαραστάσεις αποδίδουν με μεγαλύτερη σαφήνεια τον τόνο του δημιουργού στα δείγματα προς μελέτη.

Ιδιαίτερη σημασία έχει η χρήση του DistilBERT, ανάμεσα σε παρόμοια μοντέλα. Συγκεκριμένα, ανάμεσα στις σύγχρονες τεχνικές αναπαράστασης (π.χ. BERT, GPT), το DistilBERT προσφέρει μειωμένο υπολογιστικό κόστος και χρόνο εκπαίδευσης λόγω της χρήσης απόσταξης γνώσης. Το συγκεκριμένο χαρακτηριστικό του DistilBERT μπορεί να είναι κομβικής σημασίας σε περιπτώσεις χρήσης όπου οι πόροι είναι περιορισμένοι. Αποφεύγεται η σπατάλη ενεργειακών πόρων συγκριτικά με μοντέλα εκατοντάδων εκατομμυρίων παραμέτρων που παράλληλα είναι υπολογιστικά ακριβά και πολύ πιο χρονοβόρα. Δηλαδή, πρόκειται για μια μέθοδο που συγκριτικά με παρόμοιες προσεγγίσεις μπορεί να κλιμακωθεί πιο εύκολα και είναι πιο οικονομική σε όλα τα επίπεδα αν και θεωρητικά προσφέρει ελαφρώς μειωμένη απόδοση ως προς την ακρίβεια των αποτελεσμάτων.

Εκτός από το DistilBERT, σημαντικό ρόλο διαδραματίζουν και οι αρχιτεκτονικές οι οποίες μετέπειτα αξιοποιούν τα εμφυτεύματα ως δεδομένα

εκπαίδευσης. Οι δοκιμές διαφορετικών αρχιτεκτονικών, έδωσαν τη δυνατότητα να εξεταστούν αρκετές τεχνικές κανονικοποιήσης και τρόποι οργάνωσης των επιμέρους στρωμάτων. Ωστόσο, ανάμεσα στις αρχιτεκτονικές που αναπτύχθηκαν υπήρξε μια που απέδωσε καλύτερα από τις υπόλοιπες. Για παράδειγμα, στο 5ο κεφάλαιο της παρούσας εργασίας, η σύγκριση των βέλτιστων αποτελεσμάτων ανά σύνολο δεδομένων υποδεικνύει πως η αρχιτεκτονική FakeBERT είναι κατά μέσο όρο πιο αποτελεσματική. Το γεγονός πως η FakeBERT επιστρέφει υψηλές τιμές μετρικών δεν συνεπάγεται πως είναι οικονομική ως προς τους υπολογιστικούς πόρους που απαιτεί, μιας και έχει τις περισσότερες εκπαιδεύσιμες παραμέτρους συγκριτικά με τις υπόλοιπες έξι αρχιτεκτονικές.

6.3 Μελλοντικές προεκτάσεις

Ένα πεδίο το οποίο παρουσιάζει ενδιαφέρον, είναι η ανάπτυξη μοντέλων που χρησιμοποιούν με ενοποιημένο τρόπο τα επιμέρους σύνολα, ως ένα σύνολο δεδομένων. Στην παρούσα εργασία, έχει γίνει αναφορά στη διαφορετική οργάνωση των συνόλων δεδομένων και στο γεγονός πως παρέχουν διαφορετικές μορφές κειμένου. Για παράδειγμα, προσφέρονται δείγματα τα οποία είναι δημοσιεύσεις στα μέσα κοινωνικής δικτύωσης, άρθρα από μέσα μαζικής ενημέρωσης, σύντομες δηλώσεις ή και τίτλοι. Έως τώρα, κάθε σύνολο δεδομένων έχει ερευνηθεί ξεχωριστά. Όπως είναι κατανοητό, κάθε μοντέλο μπορεί να εντοπίσει μόνο ένα από όλα τα είδη περιεχομένου, λόγω του συνόλου εκπαίδευσης που το τροφοδοτεί. Συνεπώς, κανένα από τα μοντέλα δεν θα ήταν χρήσιμο για ενσωμάτωση σε μια υπηρεσία με σκοπό την επικύρωση περιεχομένου, αφού δεν θα είχε επαρκή απόδοση. Θα χρειαζόταν να χρησιμοποιηθούν πολλαπλά μοντέλα με ξεχωριστά σύνολα εκπαίδευσης και με τη συμβολή ενός πρωτοκόλλου ομοφωνίας να προκύψει η ταξινόμηση του εκάστοτε δείγματος προς μελέτη. Η διαδικασία αυτή είναι χρονοβόρα και υπολογιστικά ακριβή. Όμως, αν γίνει επεξεργασία των επιμέρους συνόλων δεδομένων, ως ένα ολοκληρωμένο σύνολο, τότε θα απαιτείται μόνο ένα μοντέλο στο οποίο και θα βασίζεται η υπηρεσία επικύρωσης περιεχομένου. Η συγκεκριμένη πρακτική είναι πιο συμφέρουσα, ως προς τους απαιτούμενους πόρους και ίσως να παρουσιάζει και καλύτερα αποτελέσματα.

Τέλος, μια ακόμα προέκταση της παρούσας εργασίας, που αφορά κυρίως δείγματα στα μέσα κοινωνικής δικτύωσης, έχει να κάνει με συνδυασμό των προσεγγίσεων ταξινόμησης. Για να υπάρξει μια πιο λεπτομερής σκιαγράφηση των στοιχείων που συνθέτουν τα fake news στα μέσα κοινωνικής δικτύωσης, χρειάζεται να αναλυθεί το συγγραφικό ύφος και το περιεχόμενο (style-based approach), λαμβάνοντας όμως υπόψη το προφίλ του εκάστοτε δημιουργού (user-based approach) και την αλληλεπίδραση του κοινού με τη δημοσίευση (post-based approach). Έτσι, θα γίνει καλύτερα κατανοητό με ποιο τρόπο διαδίδονται τα fake news στα κοινωνικά δίκτυα, από τι είδος χρηστών και πώς αντιμετωπίζονται από την κοινότητα (π.χ. αριθμός κοινοποιήσεων και likes).

Βιβλιογραφία

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Ανάκτηση από http://download.tensorflow.org/paper/whitepaper2015.pdf

Ahmed, H., Traore, I., & Saad, S. (2017, October). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. doi:10.1007/978-3-319-69155-8_9

Ahmed, H., Traore, I., & Saad, S. (2018, January/February). Detecting opinion spams and fake news using text. *Journal of Security and Privacy, 1*(1).

Ahmed, S., Hinkelmann, K., & Corradini, F. (2020). Development of Fake News Model using Machine Learning through Natural Language Processing. *International Journal of Computer and Information Engineering, 14*(12). Ανάκτηση από https://doi.org/10.48550/arXiv.2201.07489

Alghamdi, J., Lin, Y., & Luo, S. (2022). A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection. *Information, 13*(12). Ανάκτηση από https://doi.org/10.3390/info13120576

Baptista, J., & Gradim, A. A. (2022). A Working Definition of Fake News. *Encyclopedia*, σσ. 632 - 645. Ανάκτηση από https://doi.org/10.3390/encyclopedia2010043

Camacho - Collados, J., & Pilehvar, M. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. Στο *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (σσ. 40-46). Association for Computational Linguistics. Ανάκτηση από https://aclanthology.org/W18-5406

Choraś, M., Pawlicki, M., Kozik, R., Demestichas, K., Kosmides, P., & Gupta, M. (2019). SocialTruth Project Approach to Online Disinformation (Fake News) Detection and Mitigation. Στο *ARES '19: Proceedings of the 14th International Conference on Availability, Reliability and Security* (σσ. 1-10). doi:10.1145/3339252.3341497

Conforti, C., Pilehvar, M., & Collier, N. (2019). Modeling the Fake News Challenge as a Cross-Level Stance Detection Task.

Conroy, N., Rubin, V., & Yimin, C. (2015). *Automatic Deception Detection: Methods for Finding Fake News.*

Demestichas, K., Remoundou, K., & Adamopoulou, E. (2020). Food for Thought: Fighting Fake News and Online Disinformation. *IT Professional, 22*(2), σσ. 28-34. doi:10.1109/MITP.2020.2978043

Deng, Z.-H. T.-W.-Q.-Y.-Q. (2004). A Comparative Study on Feature Weight in Text Categorization. Στο *Lecture Notes in Computer Science* (σσ. 588-597). doi:10.1007/978-3-540-24655-8_64

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Στο *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (σσ. 4171 - 4186). Association for Computational Linguistics. Ανάκτηση από https://doi.org/10.18653/v1/n19-1423

Erfani, S., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition, 58*, σσ. 121-134. Ανάκτηση από https://doi.org/10.1016/j.patcog.2016.03.028

Farokhian, M., Rafe, V., & Veisi, H. (2022). Fake news detection using parallel BERT deep neural networks. *ArXiv, abs/2204.04793*.

Fuhr, N. (1992, June). Probabilistic Models in Information Retrieval. *The Computer Journal, 35*(3), σσ. 243-255. Ανάκτηση από https://doi.org/10.1093/comjnl/35.3.243

Goswami, A., Kaliyar, R., & Narang, P. (2021, March). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications, 80*(8), σσ. 11765-11788. Ανάκτηση από https://doi.org/10.1007/s11042-020-10183-2

Harris, C., Millman, K., & van der Walt, S. J. (2020). Array programming with NumPy. *Nature 585*, σσ. 357-362. Ανάκτηση από https://doi.org/10.1038/s41586-020-2649-2

Hinton, G. (2012). A practical guide to training restricted Boltzmann machines. Στο *Neural Networks: Tricks of the Trade: Second Edition* (σσ. 599-612). Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *9*(8), σσ. 1735-1780.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Στο *International Conference on Machine Learning* (σσ. 448-456).

Kochkina, E., Liakata, M., & Arkaitz, Z. (2018, August). All-in-one: Multi-task Learning for Rumour Verification. *Proceedings of the 27th International Conference on Computational Linguistics*, 3402-3413. Association for Computational Linguistics. Ανάκτηση από https://aclanthology.org/C18-1288

McKinney, W. (2011). pandas: a Foundational Python Library for Data Analysis and Statistics. Ανάκτηση από https://www.dlr.de/sc/en/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations in Vector Space.* Ανάκτηση από https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Suttskever, I., Chen, K., Corrado, G., & Dean Jeff. (2013). Distributed Representations of Words and Phrases and their Compositionality. Στο *Advances in Neural Information Processing Systems 26.*

Mishra, S., Shukla, P., & Agarwal, R. (2022). Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets. *Wireless Communications and Mobile Computing*, σ. 18. Ανάκτηση από https://doi.org/10.1155/2022/1575365

P. Ksieniewicz, P. Z. (2020). Fake News Detection from Data Streams. *2020 International Joint Conference on Neural Networks (IJCNN)*, σσ. 1-8. doi:10.1109/IJCNN48605.2020.9207498

Pennington, J., Socher, R., & Manning, C. (2014). GloVE: Global Vectors for Word Representation. Doha: Association for Computational Linguistics. doi:10.3115/v1/D14-1162

Perez, F., & Granger, B. (2015). Project Jupyter: Computational narratives as the engine of collaborative data science. *Retrieved September, 11*(207), σ. 108. Ανάκτηση από http://archive.ipython.org/JupyterGrantNarrative-2015.pdf

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Στο *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (σσ. 2227-2237). New Orleans: Association for Computational Linguistics. doi:10.18653/v1/N18-1202

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training.* OpenAI.

Rocha, Y. d. (2023). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *J Public Health (Berl.)*(31), σσ. 1007 - 1016. Ανάκτηση από https://doi.org/10.1007/s10389-021-01658-z

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Ανάκτηση από https://doi.org/10.48550/arXiv.1910.01108

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context and. *Big Data, 8*, σσ. 171-188. doi:10.1089/big.2020.0062

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *8*(3), σσ. 171 - 188. Ανάκτηση από https://doi.org/10.1089/big.2020.0062

Shu, K., Wang, S., Silva, A., Tang, J., & Liu, H. (2017, August). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*. doi:10.1145/3137597.3137600

Simons, G., & Manoilo, A. (2021). The what, how and why of fake news: An overview. *World of Media Journal of Russian Media and Journalism Studies*.

Tandoc, E., Wei Lim, Z., & Ling, R. (2018). Defining “Fake News”. *Digital Journalism, 6:2, 137-153*. doi:10.1080/21670811.2017.1360143

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*. Ανάκτηση από https://doi.org/10.48550/arXiv.1706.03762

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60), σ. 3021. doi:10.21105/joss.03021

Watson, C. (2018). Information Literacy in a Fake/False News World: An Overview of the Characteristics of Fake News and its Historical Development. *International Journal of Legal Information, 46*(2), σσ. 93 - 96. Ανάκτηση από https://doi.org/10.1017/jli.2018.25

William Yang, W. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Στο *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (σσ. 422-426). Association for Computational Linguistics. doi:10.18653/v1/P17-2067

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . M. Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Στο *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (σσ. 38-45). Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6

Yoon, K. (2014). Convolutional Neural Networks for Sentence Classification. Στο *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (σσ. 1746-1751). Association for Computational Linguistics. doi:10.3115/v1/D14-1181

Zeiler, M. (2012). Adadelta: an adaptive learning rate method.

Zhu, Y., Gensheng, W., Sheng, L., & Xuejian, H. (2023). A Novel Rumor Detection Method Based on Non-Consecutive Semantic Features and Comment Stance. *IEEE Access*.

Παράρτημα Α

Α.1. ISOT

Στη συνέχεια, παρουσιάζονται τα επιμέρους τμήματα κώδικα στα οποία διαχωρίζεται η υλοποίηση που αναλύθηκε στο 4ο κεφάλαιο για το σύνολο δεδομένων ISOT.

- **Load modules:** Στο συγκεκριμένο τμήμα κώδικα, εισάγονται τα δεδομένα και οι απαραίτητες βιβλιοθήκες της γλώσσας προγραμματισμού Python.
- **Fine Tune DistilBERT:** Στο συγκεκριμένο τμήμα κώδικα, γίνεται η εισαγωγή και η διαδικασία προεκπαίδευσης του DistilBERT.
- **Create Embeddings:** Δημιουργούνται τα εμφυτεύματα όλων των κατηγοριών που παρέχονται από το σύνολο δεδομένων ISOT.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα τίτλων.
- **Texts:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα κειμένων, που παρέχονται από το σύνολο δεδομένων ISOT.
- **Max - Worth:** Εξετάζεται πως ανταποκρίνονται οι αρχιτεκτονικές στα εμφυτεύματα του max-worth περιεχομένου. Δηλαδή, εξετάζονται τα δείγματα που προκύπτουν μετά τη συμπύκνωση των άρθρων.
- **Text-Title-DistilBERT:** Εξετάζεται πως ανταποκρίνεται η αρχιτεκτονική Title-Text του 4ου κεφαλαίου στα εμφυτεύματα των τίτλων και των κειμένων των άρθρων.
- **Max Worth-Title-DistilBERT:** Εξετάζεται πως ανταποκρίνεται η αρχιτεκτονική Title-Text του 4ου κεφαλαίου στα εμφυτεύματα των τίτλων και των συμπυκνωμένων άρθρων.

Το αρχείο κώδικα το οποίο υλοποιεί την εφαρμογή της μεθόδου στους επιμέρους τύπους δειγμάτων του συνόλου δεδομένων ISOT είναι προσβάσιμο με χρήση του ακόλουθου συνδέσμου:

https://colab.research.google.com/drive/1uNe2zGEkfx1WuorQpYf_9O2mUHF7inWA?us`p=sharing

A.2. LIAR

Στη συνέχεια, παρουσιάζονται τα επιμέρους τμήματα κώδικα στα οποία διαχωρίζεται η υλοποίηση που αναλύθηκε στο 4ο κεφάλαιο για το σύνολο δεδομένων LIAR.

- **Load modules:** Στο συγκεκριμένο τμήμα κώδικα, εισάγονται τα δεδομένα και οι απαραίτητες βιβλιοθήκες της γλώσσας προγραμματισμού Python.
- **Fine Tune DistilBERT:** Στο συγκεκριμένο τμήμα κώδικα, γίνεται η εισαγωγή και η διαδικασία προεκπαίδευσης του DistilBERT. Το DistilBERT εκπαιδεύεται σε έξι βαθμίδες αξιοπιστίας.
- **Create Embeddings:** Γίνεται δημιουργία των εμφυτευμάτων που παρέχονται από το σύνολο δεδομένων LIAR.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα των δειγμάτων. Όμως, υπάρχουν έξι πιθανές βαθμίδες ταξινόμησης ως προς την αξιοπιστία. Δεν συμπεριλαμβάνεται η αρχιτεκτονική Title-Text, καθώς το σύνολο δεδομένων παρέχει μόνο ένα είδος δείγματος.
- **Dataset Binary:** Πραγματοποιείται ένωση των δειγμάτων των κατηγοριών mostly-true και true ως αληθή. Τα υπόλοιπα δείγματα του συνόλου δεδομένων θεωρούνται αναληθή. Έτσι, δημιουργούνται μόνο δύο βαθμίδες αξιοπιστίας. Στη συνέχεια, γίνεται εκ νέου προεκπαίδευση του DistilBERT. Το μοντέλο εκπαιδεύεται σε δύο βαθμίδες αξιοπιστίας.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα των δειγμάτων. Όμως, υπάρχουν δύο πιθανές βαθμίδες ταξινόμησης ως προς την αξιοπιστία.

Το αρχείο κώδικα το οποίο υλοποιεί την εφαρμογή της μεθόδου στους επιμέρους τύπους δειγμάτων του συνόλου δεδομένων LIAR είναι προσβάσιμο με χρήση του ακόλουθου συνδέσμου:

https://colab.research.google.com/drive/19stMTYBBXgGZGSwIK1Uw9XQwSIm6YZ8m?usp=sharing

A.3. PHEME

Στη συνέχεια, παρουσιάζονται τα επιμέρους τμήματα κώδικα στα οποία διαχωρίζεται η υλοποίηση που αναλύθηκε στο 4ο κεφάλαιο για το σύνολο δεδομένων PHEME.

- **Load modules:** Στο συγκεκριμένο τμήμα κώδικα, εισάγονται τα δεδομένα και οι απαραίτητες βιβλιοθήκες της γλώσσας προγραμματισμού Python.
- **Fine Tune DistilBERT:** Στο συγκεκριμένο τμήμα κώδικα, γίνεται η εισαγωγή και η διαδικασία προεκπαίδευσης του DistilBERT. Το DistilBERT εκπαιδεύεται σε τρεις βαθμίδες αξιοπιστίας.
- **Create Embeddings:** Γίνεται δημιουργία των εμφυτευμάτων που παρέχονται από το σύνολο δεδομένων PHEME.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα των δειγμάτων. Όμως, υπάρχουν τρεις πιθανές βαθμίδες ταξινόμησης ως προς την αξιοπιστία. Δεν συμπεριλαμβάνεται η αρχιτεκτονική Title-Text, καθώς το σύνολο δεδομένων παρέχει μόνο ένα είδος δείγματος.
- **Dataset Binary:** Αφαιρείται η κατηγορία δειγμάτων με τον χαρακτηρισμό “unverified”. Έτσι, δημιουργούνται μόνο δύο βαθμίδες αξιοπιστίας. Στη συνέχεια, γίνεται εκ νέου προεκπαίδευση του DistilBERT. Το μοντέλο εκπαιδεύεται σε δύο βαθμίδες αξιοπιστίας.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα των δειγμάτων. Όμως, υπάρχουν δύο πιθανές βαθμίδες ταξινόμησης ως προς την αξιοπιστία.

Το αρχείο κώδικα το οποίο υλοποιεί την εφαρμογή της μεθόδου στους επιμέρους τύπους δειγμάτων του συνόλου δεδομένων PHEME είναι προσβάσιμο με χρήση του ακόλουθου συνδέσμου:

https://colab.research.google.com/drive/1ZjGgn_RBMkNqCDzDLEGpunomGGgw9wQQ?usp=sharing

A.4. FakeNewsNet

Στη συνέχεια, παρουσιάζονται τα επιμέρους τμήματα κώδικα στα οποία διαχωρίζεται η υλοποίηση που αναλύθηκε στο 4ο κεφάλαιο για το σύνολο δεδομένων FakeNewsNet.

- **Load modules:** Στο συγκεκριμένο τμήμα κώδικα, εισάγονται τα δεδομένα και οι απαραίτητες βιβλιοθήκες της γλώσσας προγραμματισμού Python.
- **Fine Tune DistilBERT:** Εισάγεται το μοντέλο DistilBERT και πραγματοποιείται η διαδικασία προεκπαίδευσης του.
- **Create Embeddings:** Δημιουργούνται τα εμφυτεύματα μέσω των δειγμάτων, που παρέχονται από το σύνολο δεδομένων FakeNewsNet.
- **Statements – Titles:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα των δειγμάτων. Δεν συμπεριλαμβάνεται η αρχιτεκτονική Title-Text, καθώς το σύνολο δεδομένων παρέχει μόνο ένα είδος δείγματος.

Το αρχείο κώδικα το οποίο υλοποιεί την εφαρμογή της μεθόδου στους επιμέρους τύπους δειγμάτων του συνόλου δεδομένων FakeNewsNet είναι προσβάσιμο με χρήση του ακόλουθου συνδέσμου:

https://colab.research.google.com/drive/1msbxnM-Cex0LzuSB3VrqgB09aXuwy-nP?usp=sharing

A.5. FakeNewsChallenge

Στη συνέχεια, παρουσιάζονται τα επιμέρους τμήματα κώδικα στα οποία διαχωρίζεται η υλοποίηση που αναλύθηκε στο 4ο κεφάλαιο για το σύνολο δεδομένων FakeNewsChallenge.

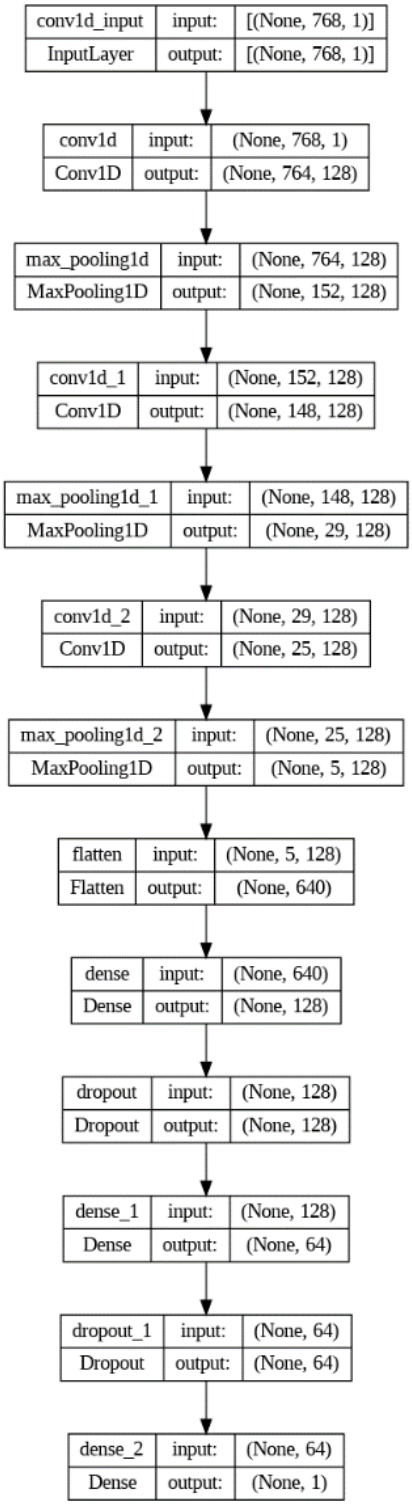
- **Load modules:** Στο συγκεκριμένο τμήμα κώδικα, εισάγονται τα δεδομένα και οι απαραίτητες βιβλιοθήκες της γλώσσας προγραμματισμού Python.
- **Fine Tune DistilBERT:** Σε αυτό το τμήμα κώδικα, γίνεται η εισαγωγή και η διαδικασία προεκπαίδευσης του DistilBERT. Το DistilBERT εκπαιδεύεται σε τέσσερις βαθμίδες αξιοπιστίας.
- **Create Embeddings:** Πραγματοποιείται δημιουργία των εμφυτευμάτων όλων των κατηγοριών που παρέχονται από το σύνολο δεδομένων FakeNewsChallenge.
- **Texts:** Εξετάζεται πως ανταποκρίνονται οι επιμέρους αρχιτεκτονικές του 4ου κεφαλαίου στα εμφυτεύματα τίτλων. Υπάρχουν τέσσερις βαθμίδες αξιοπιστίας ως προς την ταξινόμηση των δειγμάτων.
- **Max - Worth:** Εξετάζεται πως ανταποκρίνονται οι αρχιτεκτονικές στα εμφυτεύματα του max-worth περιεχομένου. Δηλαδή, εξετάζονται τα δείγματα που προκύπτουν μετά τη συμπύκνωση των άρθρων. Υπάρχουν τέσσερις βαθμίδες αξιοπιστίας ως προς την ταξινόμηση των δειγμάτων.
- **Text-Title-DistilBERT:** Συνδυάζονται τα εμφυτεύματα των τίτλων και των κειμένων των άρθρων. Τα δείγματα ταξινομούνται ως προς τέσσερις κατηγορίες.
- **Max Worth-Title-DistilBERT:** Συνδυάζονται τα εμφυτεύματα των τίτλων και των max-worth περιεχομένων. Τα δείγματα ταξινομούνται ως προς τέσσερις κατηγορίες.
- **Dataset Binary:** Τα δεδομένα των κατηγοριών agree και discusses συνδυάζονται ως αληθή και τα υπόλοιπα ως αναληθή. Άρα, τα δεδομένα κατηγοριοποιούνται πλέον σε δύο κατηγορίες. Έπειτα, στον συγκεκριμένο τομέα κώδικα υλοποιείται η διαδικασία που αναλύθηκε παραπάνω. Ωστόσο, πλέον οι κατηγορίες ταξινόμησης των δειγμάτων είναι δύο και όχι τέσσερις. Αρχικά, γίνεται η δημιουργία των εμφυτευμάτων, στη συνέχεια ακολουθεί η επεξεργασία των άρθρων, των συμπυκνωμένων άρθρων και εξετάζεται ο συνδυασμός τίτλων με τις προηγούμενες δύο κατηγορίες δειγμάτων.

Το αρχείο κώδικα το οποίο υλοποιεί την εφαρμογή της μεθόδου στους επιμέρους τύπους δειγμάτων του συνόλου δεδομένων FakeNewsChallenge είναι προσβάσιμο με χρήση του ακόλουθου συνδέσμου:

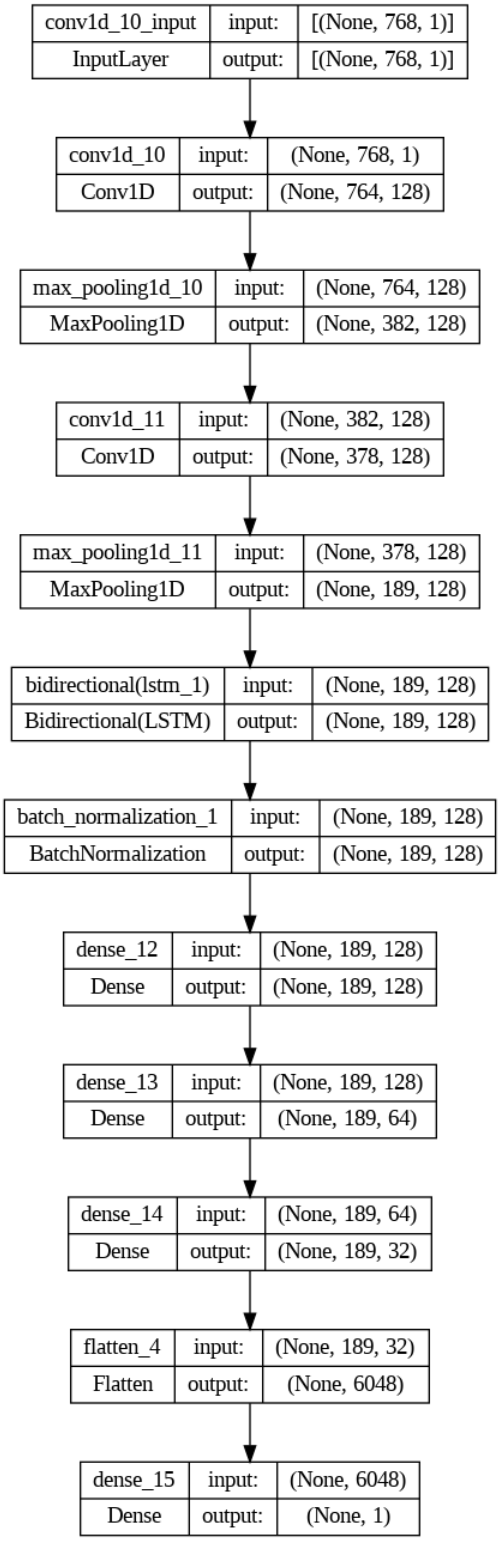
https://colab.research.google.com/drive/17yHQ62RqJxIWxG5BiUCTbVRqrB3C7EkU?usp=sharing

Παράρτημα Β

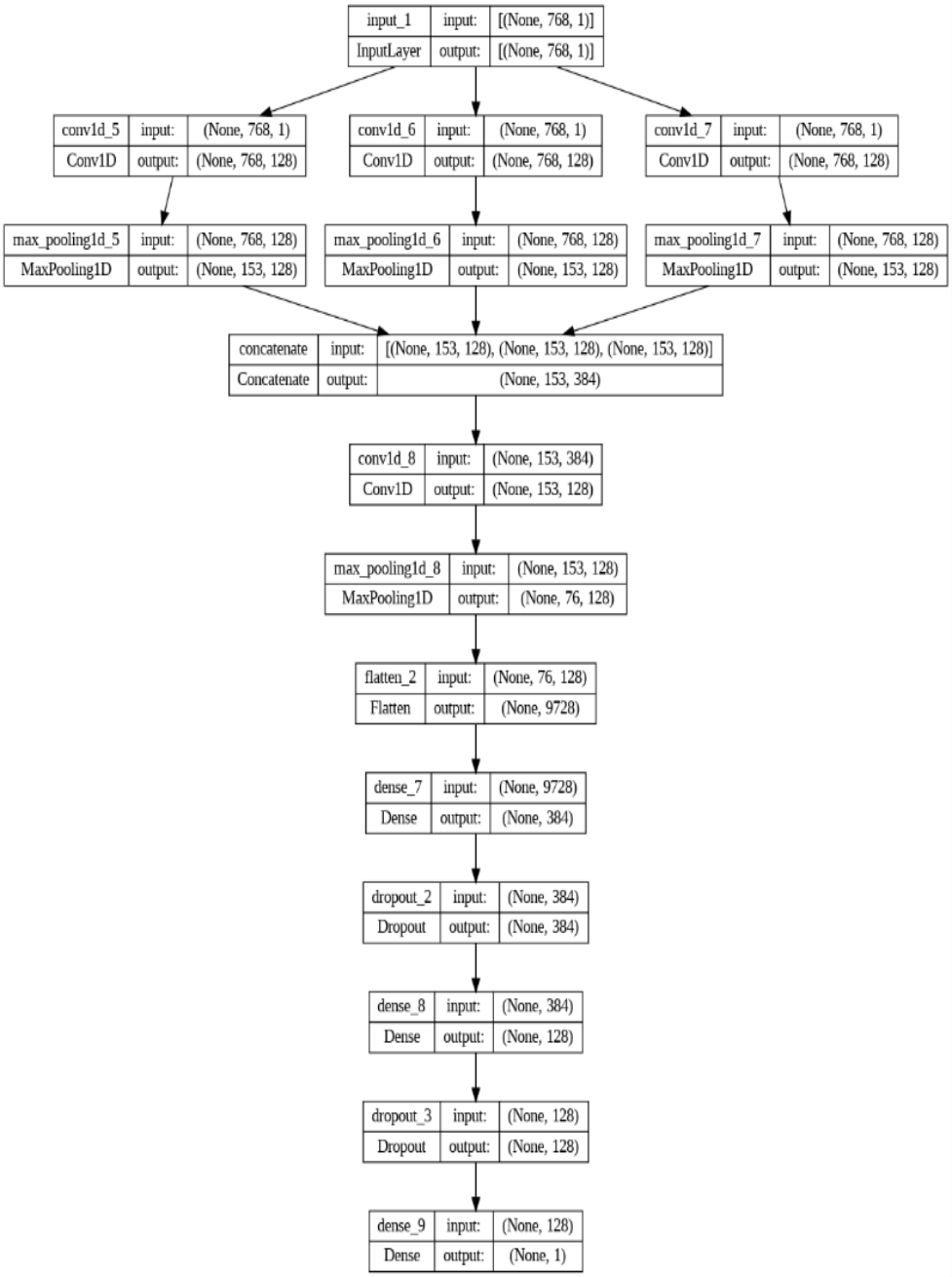
Β.1. CNN



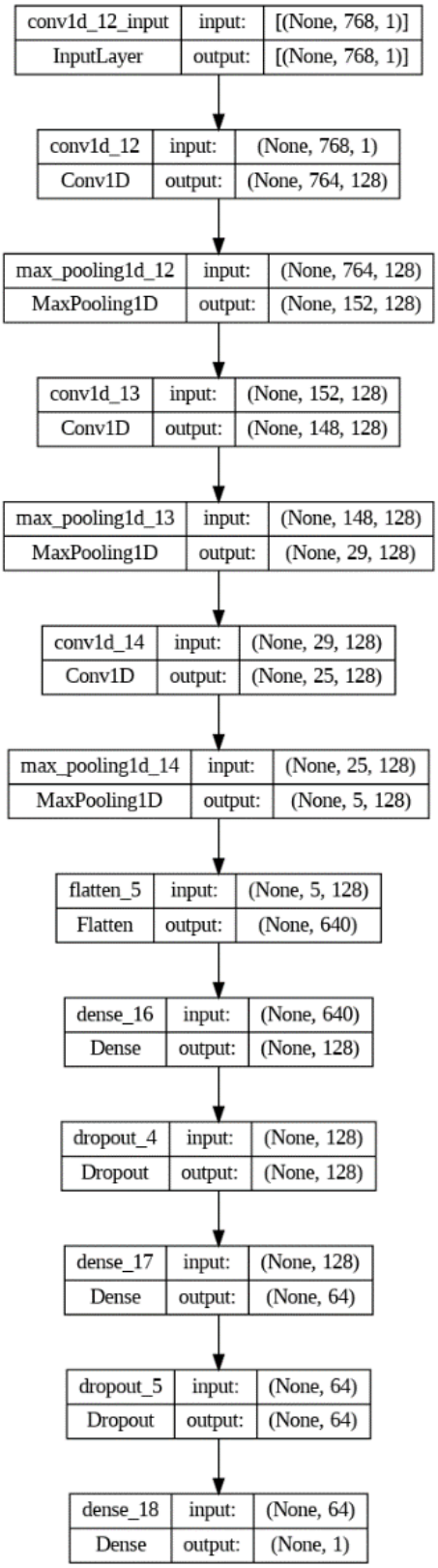
B.2. Bidirectional LSTM



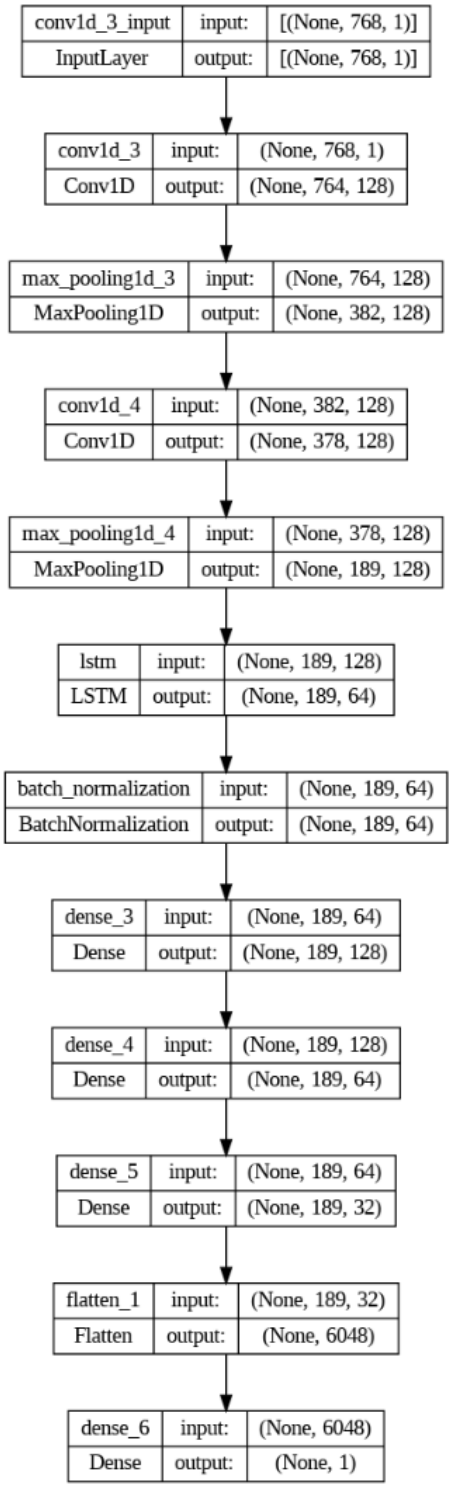
B.3. FakeBERT


input_1 | input: | [(None, 768, 1)]
InputLayer | output: | [(None, 768, 1)]
conv1d_5 | input: | (None, 768, 1)
Conv1D | output: | (None, 768, 128)
conv1d_6 | input: | (None, 768, 1)
Conv1D | output: | (None, 768, 128)
conv1d_7 | input: | (None, 768, 1)
Conv1D | output: | (None, 768, 128)
max_pooling1d_5 | input: | (None, 768, 128)
MaxPooling1D | output: | (None, 153, 128)
max_pooling1d_6 | input: | (None, 768, 128)
MaxPooling1D | output: | (None, 153, 128)
max_pooling1d_7 | input: | (None, 768, 128)
MaxPooling1D | output: | (None, 153, 128)
concatenate | input: | [(None, 153, 128), (None, 153, 128), (None, 153, 128)]
Concatenate | output: | (None, 153, 384)
conv1d_8 | input: | (None, 153, 384)
Conv1D | output: | (None, 153, 128)
max_pooling1d_8 | input: | (None, 153, 128)
MaxPooling1D | output: | (None, 76, 128)
flatten_2 | input: | (None, 76, 128)
Flatten | output: | (None, 9728)
dense_7 | input: | (None, 9728)
Dense | output: | (None, 384)
dropout_2 | input: | (None, 384)
Dropout | output: | (None, 384)
dense_8 | input: | (None, 384)
Dense | output: | (None, 128)
dropout_3 | input: | (None, 128)
Dropout | output: | (None, 128)
dense_9 | input: | (None, 128)
Dense | output: | (None, 1)

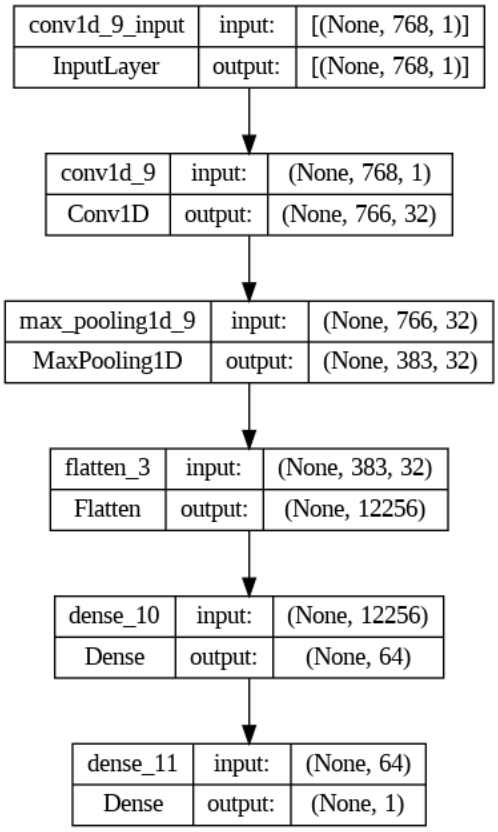
B.4. CNN-L2 Regularization



B.5. LSTM


conv1d_3_input | input: | [(None, 768, 1)]
InputLayer | output: | [(None, 768, 1)]
conv1d_3 | input: | (None, 768, 1)
Conv1D | output: | (None, 764, 128)
max_pooling1d_3 | input: | (None, 764, 128)
MaxPooling1D | output: | (None, 382, 128)
conv1d_4 | input: | (None, 382, 128)
Conv1D | output: | (None, 378, 128)
max_pooling1d_4 | input: | (None, 378, 128)
MaxPooling1D | output: | (None, 189, 128)
lstm | input: | (None, 189, 128)
LSTM | output: | (None, 189, 64)
batch_normalization | input: | (None, 189, 64)
BatchNormalization | output: | (None, 189, 64)
dense_3 | input: | (None, 189, 64)
Dense | output: | (None, 189, 128)
dense_4 | input: | (None, 189, 128)
Dense | output: | (None, 189, 64)
dense_5 | input: | (None, 189, 64)
Dense | output: | (None, 189, 32)
flatten_1 | input: | (None, 189, 32)
Flatten | output: | (None, 6048)
dense_6 | input: | (None, 6048)
Dense | output: | (None, 1)

B.6. CNN Light



B.7. Title – Text