ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΕΥΦΥΕΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΔΙΑΔΙΚΤΥΟΥ – WEB INTELLIGENCE

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανίχνευση ψευδών ειδήσεων στον πολιτικό λόγο χρησιμοποιώντας ανάλυση συναισθήματος και μοντελοποίηση θεμάτων

Του φοιτητή:
Γεώργιος Νάσκος
Αριθμός Μητρώου: 17/2023

Επιβλέπων Καθηγητής:
Παναγιώτης Αδαμίδης

Σεπτέμβριος 2025

Τίτλος Μ.Δ.Ε. Ανίχνευση Ψευδών Ειδήσεων στον Πολιτικό Λόγο Χρησιμοποιώντας Ανάλυση Συναισθήματος και Μοντελοποίηση Θεμάτων
Κωδικός Μ.Δ.Ε 24235
Ονοματεπώνυμο φοιτητή/τών Γεώργιος Νάσκος
Ονοματεπώνυμο εισηγητή Παναγιώτης Αδαμίδης
Ημερομηνία ανάληψης Μ.Δ.Ε 29-09-2024
Ημερομηνία περάτωσης Μ.Δ.Ε 01-09-2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως μεταπτυχιακή διπλωματική εργασία, στο ΠΜΣ «Ευφυείς Τεχνολογίες Διαδικτύου – Web Intelligence» του Τμήματος Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Αφιερωμένο στο δίκαιο και στην αλήθεια.»

Πρόλογος

Ο ψευδής λόγος προβληματίζει και δημιουργεί πάντα προβλήματα στην καθημερινότητα των ανθρώπων. Στην σημερινή εποχή τον συναντάμε πάρα πολύ στα μέσα κοινωνικής δικτύωσης μιας και είναι ο κύριος τρόπος ενημέρωσης για πολλούς ανθρώπους. Εκεί που μπορούμε όμως να συμφωνήσουμε όλοι είναι ότι τον συναντάμε σίγουρα στον πολιτικό λόγο, που εκεί τα προβλήματα που μπορεί να δημιουργήσει είναι πολλά γιατί επηρεάζει πόλεις, χώρες μπορεί και ολόκληρο τον κόσμο. Μέσα από τον ψευδή πολιτικό λόγο κατευνάζονται απόψεις, πράξεις και πιστεύω. Αυτό το καθιστά μείζον θέμα και όσο πιο νωρίς μπορέσουμε να τον ανιχνεύσουμε τόσο λιγότερο κακό θα δημιουργήσει.

Abstract

Nowadays, the flow and speed at which information spreads are extremely high. It is not uncommon for the information we read to be false, especially when it comes to political discourse. This can lead to misinformation, a distorted perception of reality, and poor decision-making in our lives.

In this study, we examine how an artificial intelligence system can detect falsehood in written political language, using topic and sentiment detection. Specifically, we use the information and titles of political articles provided by the researchers of the LAIR benchmark dataset, in order to extract the topics and sentiments contained in the article titles. Then, we construct a model that combines all this information to detect falsehood.We analyze the problems and challenges faced during the development of a falsehood detection model in written language.

The LAIR includes six different ordinal categories as targets, which makes the task even more difficult than a binary classification problem. We present related literature and examine how other researchers have approached the same problem, how they have handled the LAIR dataset, which information they have kept or simplified, and which features they have enhanced in order to improve their models' performance.

Finally, we propose two models. The first achieves higher accuracy than the existing models from the LAIR researchers, without using any additional information beyond topics and sentiment. However, it shows a significant deviation when it fails. The second model does not predict each label with high accuracy, but in most cases, it comes very close to the correct category. This model approaches the problem in a way that has not been previously explored in the LAIR literature, by taking advantage of the ordinal nature of the six categories and focusing on minimizing prediction error rather than simply maximizing accuracy.

Περίληψη

Στις μέρες μας, η ροή και η ταχύτητα διάδοσης της πληροφορίας είναι εξαιρετικά μεγάλες. Δεν είναι λίγες οι φορές που η πληροφορία που διαβάζουμε είναι ψευδής, ειδικά όταν πρόκειται για τον πολιτικό λόγο. Αυτό μπορεί να οδηγήσει σε παραπληροφόρηση, λανθασμένη αντίληψη της πραγματικότητας και λάθος αποφάσεις στη ζωή μας.

Στην παρούσα εργασία εξετάσουμε πως ένα σύστημα τεχνητής νοημοσύνης μπορεί να ανιχνεύσει το ψεύδος στον γραπτό πολιτικό λόγο, με την βοήθεια της ανίχνευσης θεμάτων και συναισθημάτων. Συγκεκριμένα, χρησιμοποιήσουμε τις πληροφορίες και τους τίτλους πολιτικών άρθρων, που μας παρέχουν οι ερευνητές του LAIR benchmark dataset, ώστε να εξάγουμε τα θέματα και τα συναισθήματα που περιέχουν οι τίτλοι των άρθρων, και έπειτα να κατασκευάσουμε ένα μοντέλο που συνδυάζει όλες αυτές τις πληροφορίες για να μπορέσει να ανιχνεύσει το ψεύδος. Αναλύουμε τα προβλήματα και τις δυσκολίες που αντιμετωπίζει κάποιος στην πορεία κατασκευής ενός μοντέλου ανίχνευσης ψεύδους στον γραπτό λόγο.

Το LAIR έχει για στόχους έξι διαφορετικές κλιμακωτές κατηγορίες, αυτό το καθιστά ακόμα πιο δύσκολο από μια δυαδική ταξινόμηση. Παρουσιάζουμε σχετική βιβλιογραφία και εξετάζουμε πως άλλοι ερευνητές έχουν προσεγγίσει το ίδιο πρόβλημα, πως έχουν διαχειριστεί το σύνολο δεδομένων του LAIR, ποιες πληροφορίες έχουν διατηρήσει ή απλοποιήσει, καθώς και ποιες έχουν εμπλουτίσει για τη βελτίωση της απόδοσης των μοντέλων τους.

Τέλος, προτείνουμε δύο μοντέλα. Το πρώτο παράγει υψηλότερη ακρίβεια από τα υπάρχοντα μοντέλα των ερευνητών του LAIR, χωρίς να χρησιμοποιεί επιπλέον πληροφορίες πέρα από τα θέματα και τα συναισθήματα. Ωστόσο, παρουσιάζει μεγάλη απόκλιση όταν αποτυγχάνει. Το δεύτερο μοντέλο δεν πετυχαίνει με ακρίβεια την κάθε πρόβλεψη, αλλά στις περισσότερες περιπτώσεις πλησιάζει πολύ την πραγματική κατηγορία. Το μοντέλο αυτό προσεγγίζει το πρόβλημα με τρόπο που δεν έχει εξεταστεί ξανά στη βιβλιογραφία του LAIR, αξιοποιώντας τη φύση των έξι κλιμακωτών κατηγοριών, και εστιάζει στη μείωση του σφάλματος πρόβλεψης, και όχι απλώς στην ακρίβεια.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Παναγιώτη Αδαμίδη, για την πολύτιμη καθοδήγηση, την τακτική επικοινωνία και τις εποικοδομητικές συμβουλές του καθ' όλη τη διάρκεια εκπόνησης της μεταπτυχιακής μου διπλωματικής εργασίας.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ψευδείς ειδήσεις

Η συνεχόμενη αύξηση της χρήσης των μέσων κοινωνικής δικτύωσης (social media) από τους πολίτες κάθε χώρας, αλλά και η ενημέρωσή τους από αυτά, έχει οδηγήσει και τους πολιτικούς να τα χρησιμοποιούν σε καθημερινή βάση για απλά ζητήματα αλλά και για σοβαρά. Όλοι μπορούμε να κατανοήσουμε ότι λόγω της απήχησης τους και του μεγάλου όγκου πληροφορίας, μπορούν και έχουν μεγάλη επίδραση στη γνώμη και στις απόψεις του κόσμου. Επίσης, λόγω του μεγάλου όγκου πληροφορίας, παρατηρούμε ότι υπάρχει μεγάλη παραπληροφόρηση και ψευδείς ειδήσεις. Αυτό δυσκολεύει τους πολίτες να διασταυρώσουν τις πληροφορίες για να μπορέσουν να μάθουν την αλήθεια. Καθώς οι ρυθμοί στις πόλεις, στα αστικά κέντρα, είναι υψηλοί, δεν υπάρχει χρόνος διαθέσιμος από τους πολίτες για να αφιερώσουν στην αναζήτηση της αλήθειας. Αυτά τα προβλήματα οδηγούν στην παραπληροφόρηση, τη μαζοποίηση και στη χαλιναγώγηση των πολιτών [1].

Ένα ακόμα ζήτημα είναι ότι στα άρθρα δεν είναι πάντα οι πληροφορίες εξολοκλήρου αληθείς ή ψευδείς, αλλά μπορεί να είναι μερικώς αληθείς ή μερικώς ψευδείς. Άρα θα βοηθούσε πάρα πολύ η ύπαρξη ενός συστήματος που να μπορεί να προβλέψει κατά πόσο μια πρόταση είναι αληθής ή ψευδής και αυτό να το κάνει σε κλιμακωτές κατηγορίες και όχι δυαδικές.

1.2 Σκοπός και στόχοι της εργασίας

Η παρούσα εργασία έχει ως κύριο στόχο τη μελέτη της δυνατότητας ανίχνευσης ψευδών ειδήσεων στον πολιτικό λόγο μέσω της αξιοποίησης τεχνικών επεξεργασίας φυσικής γλώσσας. Συγκεκριμένα, εξετάζονται μέθοδοι ανάλυσης συναισθήματος, θεματικής μοντελοποίησης και βαθιάς μάθησης με σκοπό την αυτόματη ταξινόμηση πολιτικών δηλώσεων ως αληθών ή ψευδών. Για την υλοποίηση της μελέτης χρησιμοποιείται το LIAR dataset, ένα δομημένο σύνολο δεδομένων που περιλαμβάνει σύντομες πολιτικές δηλώσεις, κατηγοριοποιημένες σε έξι επίπεδα αλήθειας, συνοδευόμενες από πλούσια μεταδεδομένα (όπως η ταυτότητα του ομιλητή, η κομματική του προέλευση, η θεσμική ιδιότητα του, η ημερομηνία της δήλωσης και η πηγή). Η εργασία επικεντρώνεται τόσο στην επεξεργασία των ίδιων των κειμένων των δηλώσεων, όσο και στην ενσωμάτωση των μεταδεδομένων

για την ενίσχυση της ακρίβειας των μοντέλων, όπως τη θεματική των δηλώσεων και την συναισθηματική τους ανάλυση. Σκοπός είναι η αποτίμηση της συνεισφοράς διαφορετικών τεχνικών και συνδυασμών χαρακτηριστικών στην επίδοση της ανίχνευσης ψευδούς πολιτικού λόγου.

1. Ποια είναι η επίπτωση του ψευδούς πολιτικού λόγου στην κοινωνία;
2. Ποιες είναι οι τεχνικές και μεθοδολογίες που χρησιμοποιούνται για την ανίχνευση ψεύδους στον γραπτό λόγο;
3. Γιατί χρειαζόμαστε συστήματα τεχνητής νοημοσύνης για την ανίχνευση του ψευδούς λόγου;
4. Ποιες πληροφορίες του γραπτού λόγου μπορούμε να χρησιμοποιήσουμε ή να εξάγουμε για να μπορέσουμε να πετύχουμε καλύτερα αποτελέσματα με τα συστήματα τεχνητής νοημοσύνης;
5. Πόσο αποδοτικό μπορεί να γίνει ένα τέτοιο σύστημα;

Η εργασία αυτή συμβάλλει στην έρευνα της ανίχνευσης ψευδών ειδήσεων στον πολιτικό λόγο μέσα από την παράλληλη και συγκριτική αξιοποίηση πολλαπλών τεχνικών εξαγωγής σημασιολογικής πληροφορίας. Πιο συγκεκριμένα, για τη θεματική ανάλυση των πολιτικών δηλώσεων χρησιμοποιούνται δύο διαφορετικές προσεγγίσεις: η κλασική μέθοδος θεματικής μοντελοποίησης LDA (Latent Dirichlet Allocation), καθώς και ένα σύγχρονο νευρωνικό μοντέλο τύπου LSTM (Long Short-Term Memory) το οποίο εκπαιδεύεται να αναγνωρίζει υποκείμενα θεματικά μοτίβα μέσα από την ακολουθιακή δομή των λέξεων. Παράλληλα, η ανάλυση συναισθήματος υλοποιείται με δύο επίσης εναλλακτικές προσεγγίσεις: τη λεξικογραφική μέθοδο VADER, η οποία βασίζεται σε προεπιλεγμένα συναισθηματικά λεξικά, και το προεκπαιδευμένο μετασχηματιστικό μοντέλο RoBERTa, το οποίο μπορεί να αποδώσει πιο πλούσια και συμφραζόμενη συναισθηματική πληροφορία. Όλες αυτές οι παραγόμενες πληροφορίες — θεματικά χαρακτηριστικά, συναισθηματικοί δείκτες, και μεταδεδομένα από το LIAR dataset — αξιοποιούνται σε ένα τελικό σύστημα, όπου εξετάζονται οι διάφοροι πιθανοί συνδυασμοί εισόδων με σκοπό τη βελτιστοποίηση της επίδοσης του μοντέλου ανίχνευσης ψευδών ειδήσεων.

Ένας επιπλέον τομέας καινοτομίας της εργασίας έγκειται στην προσέγγιση της στόχευσης των κατηγοριών αλήθειας του LIAR dataset. Αντί για απλή ταξινόμηση σε μία από τις έξι κατηγορίες (pants-fire, false, barely-true, half-true, mostly-true, true), υιοθετείται μια κλιμακωτή ή ιεραρχική προσέγγιση, κατά την οποία οι κατηγορίες αντιμετωπίζονται ως ένα συνεχές φάσμα αξιοπιστίας. Με τον τρόπο αυτό, δίνεται στο μοντέλο η δυνατότητα όχι μόνο να προβλέπει την κατηγορία με τη μέγιστη πιθανότητα, αλλά και να προσεγγίζει σωστά τον στόχο όταν η πρόβλεψη αποκλίνει ελαφρώς από τον στόχο. Αυτή η στρατηγική μειώνει την αυστηρότητα της αξιολόγησης και επιτρέπει την καλύτερη γενίκευση του μοντέλου σε πραγματικά δεδομένα, όπου τα όρια μεταξύ αλήθειας και ψεύδους συχνά δεν είναι απόλυτα διακριτά.

1.3 Μεθοδολογική Προσέγγιση

Τον ψευδή λόγο μπορούμε να τον συναντήσουμε σε διάφορες μορφές. Στην εργασία αυτήν εστιάζουμε στον γραπτό πολιτικό λόγο, πειραματιζόμαστε με τη χρήση της τεχνητής νοημοσύνης (ΤΝ) ώστε να προβλέψουμε κατά πόσο μια πρόταση πολιτικού περιεχομένου είναι ψευδής ή αληθής. Χρησιμοποιούμε ένα συγκεκριμένο σύνολο δεδομένων (dataset), το LIAR [2], που είναι γνωστό και σχεδιασμένο για benchmarking μοντέλα μηχανικής μάθησης (Machine Learning - ML). Λόγω ότι η πρόβλεψη του ψεύδους ή της αλήθειας είναι δύσκολο να γίνει έχοντας μόνο την πρόταση, επαυξάνουμε τα αρχικά δεδομένα του dataset με την πληροφορία των θεμάτων (topics) και της ανάλυσης συναισθήματος (sentiment analysis). Τέλος, έχει κατασκευαστεί ένα μοντέλο ιδιοκατασκευή (custom), που έχει ως είσοδο τα δεδομένα από το LIAR αλλά και τις προβλέψεις των θεμάτων και του sentiment analysis, και προσπαθεί να ταξινομήσει τις προτάσεις ανάμεσα σε 6 διαφορετικές κατηγορίες που παρέχει το LIAR. Εξάγουμε συμπεράσματα κατά πόσο αυτά τα επιπλέον μεταδεδομένα μας βοηθήσουν στην πρόβλεψη μας. Συγκρίνουμε τα αποτελέσματα μας με αυτά των ερευνητών του LAIR για να συμπεράνουμε άμα έχουμε βελτίωση της απόδοσης του μοντέλου.

Πιο αναλυτικά, για τα topics γίνεται χρήση του Latent Dirichlet Allocation (LDA) [3], αλλά και ενός custom μοντέλου, ειδικά μόνο για τα θέματα, με χρήση Long Short-Term Memory (LSTM) [4] και Deep Neural Networks (DNN) [5]. Για την ανίχνευση συναισθήματος γίνεται χρήση ενός lexical-based μοντέλου, το VADER (Valence Aware Dictionary and sEntiment Reasoner) [6], αλλά και του RoBERTa [7], που είναι ένα μοντέλο βαθιάς μάθησης (deep learning) [5] βασισμένο στο Bidirectional Encoder Representations from Transformers (BERT) [8], αλλά έχει γίνει μετεκπαίδευση σε περισσότερα δεδομένα και με μεγαλύτερη έμφαση στις υπερπαραμέτρους του μοντέλου και έτσι προσφέρει καλύτερη απόδοση.

Αφού έγινε ένα pre-processing των δεδομένων του LIAR [2], έγιναν πολλές δοκιμές για το τελικό μοντέλο που συνδυάζει όλα τα δεδομένα με διαφορετικές τεχνικές και μοντέλα, όπως LSTM [4], BERT [8] και Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNN) [9]. Τέλος, έχοντας τα αποτελέσματα από όλες τις δοκιμές και τους συνδυασμούς, συμπεραίνουμε αν και κατά πόσο βοηθούν οι πληροφορίες των θεμάτων και της ανίχνευσης συναισθήματος στο τελικό μοντέλο για την ταξινόμηση των προτάσεων στην κλίμακα αλήθειας-ψεύδους. Επίσης, γίνεται σύγκριση του καλύτερου μοντέλου αυτής της εργασίας με υπάρχοντα μοντέλα από άλλες δημοσιεύσεις βασισμένες στο LIAR dataset.

1.4 Δομή της εργασίας

Μετά από αυτήν την εισαγωγή, στο Κεφάλαιο 2 ακολουθεί μια βιβλιογραφική ανασκόπηση γενικά για τις ψευδείς ειδήσεις και μετά πιο συγκεκριμένα για ψευδείς ειδήσεις στον πολιτικό λόγο. Στη συνέχεια γίνεται περιγραφή και ανάλυση όλων των τεχνολογιών που χρησιμοποιήθηκαν στην εργασία αυτήν. Τέλος, γίνεται περιγραφή του LIAR dataset και ανάλυση των δεδομένων που παρέχει. Στο Κεφάλαιο 3 αναλύουμε με λεπτομέρειες τη μεθοδολογία που ακολουθήθηκε εργασία αυτή. Στο Κεφάλαιο 4 βρίσκονται τα αποτελέσματα και η αξιολόγηση των μοντέλων από τις εκτελέσεις που έγιναν πάνω στο LIAR dataset, έχοντας υπόψη τη μεθοδολογία που περιγράφτηκε στο Κεφάλαιο 3. Και στο τελευταίο κεφάλαιο

βρίσκονται τα συμπεράσματα της εργασίας και μελλοντικές ιδέες για συνέχεια της έρευνας στην ανίχνευση του ψευδού λόγου.

1.5 Επίλογος

Στην εισαγωγή αναφερθήκαμε στην θεματολογία και τους στόχους αυτής της εργασίας, που εστιάζει στην ανίχνευση ψευδών πολιτικών ειδήσεων, αξιοποιώντας διαφορετικά χαρακτηριστικά, όπως τα θέματα και την ανίχνευση συναισθήματος. Μέσα από την ανάλυση και την υλοποίηση αντίστοιχου μοντέλου αναδείχθηκαν οι δυνατότητες και οι προκλήσεις της αυτόματης ανίχνευσης ψευδών ειδήσεων. Τα αποτελέσματα δείχνουν πως ο συνδυασμός πολλών παραμέτρων μπορεί να ενισχύσει την ακρίβεια της ανίχνευσης, προσφέροντας ένα εργαλείο χρήσιμο για την καταπολέμηση της παραπληροφόρησης στον δημόσιο λόγο.

Κεφάλαιο 2

Βιβλιογραφική Ανασκόπηση

2.1 Εισαγωγή

Η διάδοση ψευδών ειδήσεων (fake news) αποτελεί ένα από τα πιο ανησυχητικά φαινόμενα της ψηφιακής εποχής, με σημαντικές επιπτώσεις στην ενημέρωση του κοινού, στη δημόσια συζήτηση και στη δημοκρατική διαδικασία. Η ανάγκη για αξιόπιστα και αποδοτικά εργαλεία ανίχνευσης ψευδούς περιεχομένου έχει οδηγήσει σε έντονη ερευνητική δραστηριότητα, με επίκεντρο την αξιοποίηση τεχνικών μηχανικής μάθησης και βαθιάς μάθησης. Οι πρόσφατες μελέτες επικεντρώνονται στην ανάλυση του γλωσσικού ύφους, του κοινωνικού πλαισίου και της πολυτροπικής πληροφορίας, με στόχο την ακριβή αναγνώριση παραπλανητικών ειδήσεων σε ευρεία κλίμακα. Οι παρακάτω παράγραφοι παρουσιάζουν βασικές ερευνητικές κατευθύνσεις και σύγχρονες προκλήσεις στο πεδίο της ανίχνευσης ψευδών ειδήσεων, όπως αναδεικνύονται από πρόσφατες επιστημονικές δημοσιεύσεις.

2.2 Ανίχνευση ψευδών ειδήσεων

Η ανίχνευση ψευδών ειδήσεων έχει αναδειχθεί σε κρίσιμο πεδίο έρευνας λόγω της αυξανόμενης εξάπλωσης παραπληροφόρησης μέσω ψηφιακών μέσων και κοινωνικών δικτύων. Οι σύγχρονες προσεγγίσεις βασίζονται σε τεχνικές ML και deep learning, αξιοποιώντας χαρακτηριστικά όπως το περιεχόμενο του κειμένου, το κοινωνικό πλαίσιο και εξωτερικές πηγές γνώσης. Σύμφωνα με τον Linmei et al. [10], οι μέθοδοι deep learning έχουν ξεπεράσει τις παραδοσιακές τεχνικές μηχανικής μάθησης, επιτυγχάνοντας υψηλότερη ακρίβεια στην ανίχνευση ψευδών ειδήσεων, ιδίως όταν συνδυάζονται με πληροφορίες από τα κοινωνικά δίκτυα και εξωτερικές βάσεις δεδομένων.

Ωστόσο, η αποτελεσματικότητα αυτών των μοντέλων επηρεάζεται από προκλήσεις, όπως η περιορισμένη διαθεσιμότητα αξιόπιστων και ποικιλόμορφων δεδομένων, η υπερπροσαρμογή (overfitting) σε συγκεκριμένα σύνολα δεδομένων και η ανάγκη προσαρμογής σε διαφορετικά πολιτισμικά και γλωσσικά πλαίσια. Η έρευνα του Elhadad et al. [11] επισημαίνει την ανάγκη για ανάπτυξη πολυτροπικών πλαισίων που συνδυάζουν κείμενο, εικόνες και κοινωνικό πλαίσιο, καθώς και για την ενίσχυση της ψηφιακής παιδείας των χρηστών.

Επιπλέον, η ταχεία εξέλιξη των τεχνολογιών τεχνητής νοημοσύνης, όπως τα μεγάλα γλωσσικά μοντέλα (Large Language Models - LLM), δημιουργεί νέες προκλήσεις στην ανίχνευση ψευδών ειδήσεων. Οι δημιουργοί παραπληροφόρησης μπορούν να εκμεταλλευτούν αυτές τις τεχνολογίες για την παραγωγή πιο πειστικών και δύσκολα ανιχνεύσιμων ψευδών περιεχομένων. Σύμφωνα με τον Wang et al. [12], είναι απαραίτητη η ανάπτυξη νέων μεθόδων ανίχνευσης που να μπορούν να αντιμετωπίσουν τις προκλήσεις που προκύπτουν από τη χρήση προηγμένων τεχνολογιών στην παραγωγή ψευδών ειδήσεων.

Από μια ακόμη έρευνα παρατηρούμε ότι τονίζουν την επίπτωση των κοινωνικών δικτύων στην καθημερινότητά μας, αλλά και τις ψευδείς ειδήσεις που διαχέονται μέσα από αυτά [13]. Η χρήση των κοινωνικών δικτύων για την κατανάλωση ειδήσεων αποτελεί ένα δίκοπο μαχαίρι. Από τη μία πλευρά, η ευκολία πρόσβασης, το χαμηλό κόστος και η ταχύτητα διάδοσης της πληροφορίας ωθούν τους ανθρώπους να ενημερώνονται μέσω αυτών των μέσων. Από την άλλη πλευρά, όμως, οι ίδιες αυτές δυνατότητες διευκολύνουν και την εξάπλωση των ψευδών ειδήσεων — περιεχόμενο δηλαδή χαμηλής ποιότητας, το οποίο περιλαμβάνει εσκεμμένα ψευδείς πληροφορίες. Η μαζική διάδοση τέτοιων ειδήσεων μπορεί να έχει σοβαρές επιπτώσεις τόσο στο άτομο όσο και στο κοινωνικό σύνολο. Γι' αυτόν τον λόγο, η ανίχνευση ψευδών ειδήσεων στα κοινωνικά δίκτυα έχει αναδειχθεί σε ένα νέο και ιδιαίτερα δραστήριο πεδίο έρευνας.

Η ανίχνευση ψευδών ειδήσεων στα κοινωνικά μέσα παρουσιάζει ιδιαιτερότητες που δυσκολεύουν την εφαρμογή παραδοσιακών αλγορίθμων, οι οποίοι είχαν σχεδιαστεί για συμβατικά μέσα ενημέρωσης [13]. Οι ψευδείς ειδήσεις είναι σκόπιμα διατυπωμένες ώστε να παραπλανούν τους αναγνώστες, κάτι που καθιστά δύσκολη τη διάκρισή τους μόνο μέσω της ανάλυσης περιεχομένου. Επομένως, απαιτείται και η αξιοποίηση επιπλέον δεδομένων, όπως οι αλληλεπιδράσεις των χρηστών στα κοινωνικά δίκτυα. Ωστόσο, αυτού του είδους τα δεδομένα παρουσιάζουν προκλήσεις από μόνα τους, καθώς είναι ογκώδη, ανολοκλήρωτα, ανοργάνωτα και συχνά θορυβώδη. Για την αντιμετώπιση αυτών των ζητημάτων, οι ερευνητές πραγματοποίησαν αυτή τη μελέτη, προσφέροντας μια ανασκόπηση του πεδίου που καλύπτει από ψυχολογικές και κοινωνικές θεωρήσεις μέχρι μεθόδους εξόρυξης δεδομένων, σύνολα δεδομένων αξιολόγησης και μελλοντικές ερευνητικές προοπτικές [13].

2.3 Ψευδείς ειδήσεις και πολιτικός λόγος

Η ανίχνευση ψευδών ειδήσεων στον πολιτικό λόγο αποτελεί κρίσιμο πεδίο έρευνας, καθώς οι ψευδείς πληροφορίες μπορούν να επηρεάσουν τη δημόσια γνώμη και να υπονομεύσουν τη δημοκρατική διαδικασία. Οι ερευνητές έχουν αναπτύξει διάφορες μεθόδους για την αντιμετώπιση αυτού του φαινομένου. Για παράδειγμα ο ερευνητής Raza [14] πρότεινε ένα πλαίσιο ανίχνευσης ψευδών ειδήσεων που βασίζεται σε αρχιτεκτονική Transformer, αξιοποιώντας πληροφορίες από άρθρα ειδήσεων και το κοινωνικό τους πλαίσιο για την ανίχνευση ψευδών ειδήσεων με υψηλή ακρίβεια.

Άλλη προσέγγιση παρουσιάζεται από Khan και την ομάδα του [15], οι οποίοι πραγματοποίησαν μια συγκριτική μελέτη διαφόρων μοντέλων μηχανικής μάθησης για την ανίχνευση ψευδών ειδήσεων. Η μελέτη τους ανέδειξε ότι τα προεκπαιδευμένα μοντέλα,

όπως το BERT [8], υπερέχουν στην ανίχνευση ψευδών ειδήσεων, ιδιαίτερα όταν τα διαθέσιμα δεδομένα είναι περιορισμένα.

Επιπλέον, η έρευνα των Sude, Sharon και Dvir-Gvirsman [16] διερεύνησε πως η κομματική ταυτότητα επηρεάζει την ικανότητα των ατόμων να εντοπίζουν ψευδείς ειδήσεις. Τα αποτελέσματα έδειξαν ότι άτομα με ισχυρή κομματική ταύτιση τείνουν να θεωρούν ακριβές τις πολιτικά συνεπείς ειδήσεις, ακόμη και όταν προέρχονται από άγνωστες πηγές, υπογραμμίζοντας την ανάγκη για ενίσχυση της κριτικής σκέψης και της παιδείας στα μέσα ενημέρωσης.

Τέλος, η έρευνα των Khan et al. [15] υπογραμμίζει τη σημασία της χρήσης προηγμένων μοντέλων μηχανικής μάθησης για την ανίχνευση ψευδών ειδήσεων, ειδικά σε πολιτικά ευαίσθητα περιβάλλοντα. Η αξιοποίηση τέτοιων τεχνικών μπορεί να συμβάλει σημαντικά στην προστασία της δημόσιας σφαίρας από την παραπληροφόρηση.

2.4 Ανάλυση συναισθήματος

Τα τελευταία χρόνια, με την ολοένα και μεγαλύτερη χρήση των social media, εμφανίζονται πάρα πολλές ψευδείς ειδήσεις. Για να μπορέσουν να πείσουν τους αναγνώστες, χρησιμοποιούν διάφορες τεχνικές, μια από αυτές είναι να δημιουργήσουν έντονα συναισθήματα στους αναγνώστες, όπως γέλιο, λύπη, εκνευρισμό και άλλα πολλά. Έτσι προκύπτει και η χρήση της ανίχνευσης συναισθήματος για να βοηθήσει στην εύρεση των ψευδών ειδήσεων [17]. Η ανίχνευση συναισθήματος μπορεί να χρησιμοποιηθεί από μόνη της ή και συνδυαστικά με άλλες τεχνικές και πληροφορίες. Τις περισσότερες φορές το συναντάμε συνδυαστικά, λόγω της δυσκολίας που υπάρχει στην εύρεση του ψεύδους, είναι δύσκολο να έχουμε καλά αποτελέσματα μόνο με αυτό [18].

Υπάρχουν δύο βασικές τεχνικές για την ανίχνευση συναισθήματος, οι λεξικοκεντρικές (vocabulary based) και deep learning ή LLM.

- **Vocabulary based** προσπαθούν να ταξινομήσουν ανάμεσα σε διάφορες κατηγορίες συναισθήματος χρησιμοποιόντας τις λέξεις που περιέχουν οι προτάσεις. Με κάποιον αλγόριθμο και τη χρήση της στατιστικής έχουν ανατεθεί βάρη στις λέξεις του λεξικού που τις χαρακτηρίζουν ως προς το συναίσθημα που εκφράζουν. Με αυτόν τον τρόπο παράγεται ένας τελικός αριθμός που δείχνει το γενικό συναίσθημα μιας πρότασης. Το πρόβλημα που συναντάμε, είναι ότι πολλές λέξεις από μόνες τους δεν μπορούν να εκφράσουν κάποιο συγκεκριμένο συναίσθημα αλλά αλλάζει ανάλογα το γενικό νόημα της πρότασης. Για παράδειγμα, μπορεί σε μια πρόταση με κυριολεκτική χροιά να την χαρακτηρίζαμε αρνητική, αλλά γνωρίζοντας ότι αυτός που το έγραψε κάνει σάτιρα, τότε να αναγνωρίζαμε ότι το συναίσθημα είναι θετικό. Επίσης, πολλές λέξεις έχουν ουδέτερο συναίσθημα, και πάλι αυτές μπορεί να αλλάξουν το συναίσθημα που εκφράζουν ανάλογα με το νόημα της πρότασης. Αυτά καθιστούν τα vocabulary based όχι και τόσο αποδοτικό συστήματα για την ταξινόμηση. Ένα τέτοιο σύστημα, στην περίπτωση μιας μη εποπτικής εκπαίδευσης (unsupervised training) μπορεί να φανεί πολύ χρήσιμο. Δηλαδή όταν δεν μπορούμε εμείς να ορίσουμε στόχους σε κάποιο

dataset που φτιάξαμε, γιατί είναι πολύ μεγάλο ή γιατί δεν υπάρχει ο χρόνος για να το κάνουμε, τότε αυτά τα συστήματα μπορούν να προβλέψουν τους στόχους σε αυτά τα dataset ώστε στη συνέχεια να εκπαιδεύσουμε ένα deep learning ή LLM. Σε αντίθετη περίπτωση, οι ερευνητές αναλαμβάνουν τον φόρτο της ανάγνωσης κάθε πρότασης και στην συνέχεια με την κρίση τους να ταξινομήσουν τα κείμενα ανάμεσα στις κατηγορίες συναισθήματος. Ένα από αυτά τα συστήματα είναι το: VADER [6] και το TextBlob [19].

- **Deep learning ή LLM** είναι διάφορα συστήματα με πάρα πολλές παραλλαγές και συνδυασμούς μεθοδολογιών όπως τα LSTM, CNN, Autoencoders, Transformers, Feed forward Neural Networks, που συνέχεια αναπτύσσονται και αναβαθμίζονται με ραγδαίο ρυθμό. Τα συστήματα αυτά είναι αυτοεκπαιδευόμενα και μπορούν να μάθουν από μόνα τους τη σημασία της κάθε λέξης μέσα στην πρόταση, ανάλογα με τη θέση της αλλά και ανάλογα με τις άλλες λέξεις που είναι πριν ή μετά από αυτήν. Έτσι, η κάθε λέξη μπορεί και έχει διαφορετικό νόημα ανάλογα με τις λέξεις που περιέχει η πρόταση και τη θέση που έχουν σε αυτήν. Στην πράξη έχουν δείξει μεγάλη επιτυχία, μπορούν να εκπαιδευτούν με επίβλεψη (supervised) αλλά και με ενίσχυση (reinforcement). Συνηθίζεται η αρχική εκπαίδευση να γίνεται σε όσο μεγαλύτερα dataset γίνεται με supervised τρόπο και στη συνέχεια να γίνεται ”fine tuning” με reinforcement. Τα προβλήματα που συναντάμε με αυτά τα συστήματα είναι ότι πρέπει να γίνουν πάρα πολλές δοκιμές στις υπερπαραμέτρους που έχουν και στο σχήμα τους για να μπορέσει να βρεθεί ένα αποδοτικό μοντέλο. Και όλα αυτά χωρίς να υπάρχει μια συγκεκριμένη συνταγή που μπορεί κάποιος να ακολουθήσει για να βρει το αποδοτικό μοντέλο. Όπως έχει ήδη αναφερθεί, λόγω ότι η πρόβλεψη και γενικά η ταξινόμηση του λόγου είναι δύσκολη διαδικασία, αυτό αναγκάζει τα μοντέλα να είναι μεγάλα και περίπλοκα, ειδικά αυτά που προσπαθούν να αποδώσουν καλά σε γενικά πλαίσια προτάσεων. Αυτό καθιστά τα μοντέλα αυτά να είναι αργά κατά την εκπαίδευσή τους και να χρειάζονται ακριβό υλικό (hardware) για να επιταχύνουμε την εκπαίδευση τους. Ένα πολύ γνωστό μοντέλο είναι το BERT [8] και το RoBERTa [7], τα οποία είναι προεκπαιδευμένα και γενικού σκοπού. Αυτά τα μοντέλα έχουν αποδώσει πολύ καλά αποτελέσματα σε πλήθος εφαρμογών, όπως προκύπτει από πολλές έρευνες και δημοσιεύσεις.

2.5 Θεματική μοντελοποίηση

Μια ακόμα πληροφορία που μπορούμε να χρησιμοποιήσουμε για την καλύτερη ταξινόμηση των fake news είναι η εύρεση του θέματος της πρότασης. Κάποια από τα θέματα μπορεί να έχουν πιο μεγάλη πιθανότητα να τα συναντάμε σε fake news ή μη και αυτό μπορεί να βοηθήσει το μοντέλο στο τελικό του συμπέρασμα. Όπως και με την ανίχνευση συναισθήματος υπάρχουν δυο βασικοί τρόπου για την εύρεση του θέματος.

- **Στατιστικοί μέθοδοι – clustering,** αυτοί οι μέθοδοι βασίζονται σε κάποια στατιστική συνάρτηση και αλγόριθμο ώστε να κατασκευαστούν ομάδες από λέξεις που πιθανός να έχουν κοινό νόημα που σχετίζεται με την συχνότητα της εμφάνισης τους μέσα στην πρόταση αλλά και σε διαφορετικές προτάσεις. Η κάθε ομάδα περιέχει λέξεις που θεωρεί η μέθοδος ότι είναι παρεμφερείς μεταξύ τους. Συνήθως έχουν και έναν συντελεστή βαρύτητας. Οι μέθοδοι αυτοί δεν μπορούν να μας δώσουν ένα όνομα για τις ομάδες που

φτιάχνουν ώστε να θεωρήσουμε ότι αυτό είναι το θέμα. Αυτό, θα πρέπει να το κάνουμε εμείς άμα θέλουμε. Είναι μέθοδοι για unsupervised learning, και αυτό μας βοηθάει όταν δεν έχουμε τα θέματα σαν στόχους (targets) στο dataset μας να μπορούμε να παράξουμε τα θέματα για τα κείμενα μας και στην συνέχεια να κάνουμε supervised training για την εύρεση των fake news [20]. Σε αυτήν την εργασία χρησιμοποιούμε το LDA [3] που είναι μια μέθοδος τέτοιου τύπου. Μπορεί να μας φτιάξει ομάδες από λέξεις, που θα θεωρήσουμε ότι είναι τα θέματα μας [21].

- **Deep learning,** όπως και με την ανίχνευση συναισθήματος, μπορεί να γίνει χρήση deep learning μοντέλων για την εύρεση θεμάτων. Αυτά τα μοντέλα εκπαιδεύονται με supervised learning, άρα πρέπει να γνωρίζουμε για το dataset μας τα θέματα για κάθε κείμενο. Στην εργασία αυτήν μελετάμε πως γίνεται να το πετύχουμε αυτό με την χρήση του LSTM [4] και των embeddings [22], [23].

2.6 Μέθοδοι ανίχνευσης ψευδών ειδήσεων

Όλες οι προηγούμενες μεθοδολογίες που περιγράψαμε μπορούν να χρησιμοποιηθούν για την πρόβλεψη του ψευδού λόγου. Ενώ μπορούμε να χρησιμοποιήσουμε αυτές τις μεθοδολογίες από μόνες τους για την πρόβλεψη του ψευδού λόγου, στην πράξη συμπεραίνουμε ότι με αυτόν τον τρόπο δεν μπορούμε να πετύχουμε καλά αποτελέσματα. Για αυτό τα περισσότερα συστήματα και μοντέλα είναι συνδυασμός πολλών μεθόδων και τεχνικών. Όπως σε αυτήν την δημοσίευση που με την χρήση ενός autoencoder κάνουν επαύξηση (augment) των προτάσεων του LAIR ώστε να βοηθήσουν στη συνέχεια το CNN μοντέλο τους να πετύχει καλύτερο score. Πολλοί ερευνητές χρησιμοποιούν έτοιμα pre-trained μοντέλα όπως το BERT [8] και το RoBERTa [7] και τα κάνουν fine-tune στο dataset τους. Έχουν δείξει καλά αποτελέσματα, όπως σε αυτήν την έρευνα [24] που συγκρίνει πολλά τέτοια μοντέλα μεταξύ τους πάνω στο LAIR dataset και καταλήγει ότι το RoBERTa είναι το καλύτερο από αυτά που δοκίμασαν.

2.7 Έρευνες με το LAIR dataset

Η έρευνα των Bibek Upadhayay και Vahid Behzadan [25] αναδεικνύει τη σημασία της συναισθηματικής ανάλυσης στην ανίχνευση ψευδών ειδήσεων, εστιάζοντας σε σύντομες δηλώσεις που συναντώνται συχνά στα μέσα κοινωνικής δικτύωσης. Η εργασία αυτή εισάγει το σύνολο δεδομένων Sentimental LIAR [25], το οποίο αποτελεί μια επεκταμένη έκδοση του αρχικού συνόλου δεδομένων LIAR [2], προσθέτοντας συναισθηματικής ανάλυσης για κάθε δήλωση. Η προσέγγιση αυτή βασίζεται στην υπόθεση ότι οι ψευδείς δηλώσεις συχνά περιέχουν έντονα συναισθηματικά στοιχεία, τα οποία μπορούν να χρησιμοποιηθούν για την ανίχνευσή τους.

Το Sentimental LIAR [25] προσαρμόζει το αρχικό σύνολο δεδομένων LIAR, το οποίο περιλαμβάνει 12.8 χιλιάδες δηλώσεις από το PolitiFact.com, ταξινομημένες σε συγκεκριμένες κατηγορίες ψεύδους. Στη νέα αυτή έκδοση, οι πολυκατηγορηματικοί στόχοι του LIAR (όπως ”half-true”, ”false”, ”barely-true”, ”pants-fire”) μετατρέπονται σε δυαδική κωδικοποίηση, όπου οι δηλώσεις που χαρακτηρίζονται ως ”true” παραμένουν αληθείς, ενώ όλες οι

άλλες κατηγορίες αναγνωρίζονται ως ψευδείς. Επιπλέον, οι ονομασίες των ομιλητών αντικαθίστανται με αριθμητικά αναγνωριστικά για την αποφυγή προκαταλήψεων που σχετίζονται με την κειμενική αναπαράσταση των ονομάτων. Η επεκταμένη αυτή έκδοση περιλαμβάνει επίσης χαρακτηριστικά συναισθηματικής ανάλυσης που προκύπτουν από την ανάλυση συναισθημάτων και συναισθημάτων των δηλώσεων, χρησιμοποιώντας εργαλεία όπως το Google NLP API και το IBM NLP API. Αυτά τα χαρακτηριστικά περιλαμβάνουν γενικά συναισθηματικά σκορ (θετικά ή αρνητικά) καθώς και επιμέρους συναισθηματικά σκορ για έξι συναισθηματικές καταστάσεις: θυμός, λύπη, αηδία, φόβος και χαρά.

Η μελέτη προτείνει μια νέα αρχιτεκτονική βαθιάς μάθησης βασισμένη στο μοντέλο BERT-Base για την ταξινόμηση των δηλώσεων ως αληθείς ή ψευδείς. Τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη αρχιτεκτονική, όταν εκπαιδεύεται στο Sentimental LIAR, μπορεί να επιτύχει ακρίβεια 70%, βελτιώνοντας κατά περίπου 30% τα προηγούμενα αποτελέσματα για το dataset LIAR. Αυτή η βελτίωση υπογραμμίζει τη σημασία της συναισθηματικής και συναισθηματικής ανάλυσης στην ανίχνευση ψευδών ειδήσεων, ιδιαίτερα σε σύντομες δηλώσεις που συχνά περιέχουν έντονα συναισθηματικά στοιχεία [25].

Η μελέτη των Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha και Nikos Komninos [26] διερευνά την επίδραση της ενσωμάτωσης γνώσης σε προεκπαιδευμένα γλωσσικά μοντέλα για την ανίχνευση ψευδών ειδήσεων. Οι συγγραφείς αξιολογούν διάφορες μεθόδους ενσωμάτωσης γνώσης, κυρίως χρησιμοποιώντας το Wikidata ως βάση γνώσης, σε δύο δημοφιλή σύνολα δεδομένων για ψευδείς ειδήσεις: το LIAR, ένα πολιτικά προσανατολισμένο σύνολο δεδομένων, και το COVID-19 [27], ένα σύνολο δεδομένων με μηνύματα που δημοσιεύτηκαν στα μέσα κοινωνικής δικτύωσης σχετικά με την πανδημία COVID-19.

Τα πειράματα δείχνουν ότι τα μοντέλα με ενισχυμένη γνώση μπορούν να βελτιώσουν σημαντικά την ανίχνευση ψευδών ειδήσεων στο LIAR, όταν η βάση γνώσης είναι σχετική και ενημερωμένη. Ωστόσο, τα μικτά αποτελέσματα στο COVID-19 υπογραμμίζουν τη σημασία των χαρακτηριστικών του στυλ και την ανάγκη για εξειδικευμένες και επίκαιρες βάσεις γνώσης [26].

Η εργασία αυτή συμβάλλει σημαντικά στην κατανόηση του τρόπου με τον οποίο η ενσωμάτωση γνώσης μπορεί να ενισχύσει τα γλωσσικά μοντέλα για την ανίχνευση ψευδών ειδήσεων, προσφέροντας νέες προσεγγίσεις και εργαλεία για την ανάπτυξη πιο αποτελεσματικών συστημάτων ανίχνευσης. Η προσέγγιση αυτή ανοίγει νέους δρόμους για την ενσωμάτωση δομημένης γνώσης στην ανάλυση κειμένων, προσφέροντας μια πιο ολοκληρωμένη κατανόηση των παραμέτρων που επηρεάζουν την αλήθεια ή την ψευδαισθησία των δηλώσεων [26].

Η μελέτη των Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej και Asma Khaled Aldubaikil [28], παρουσιάζει ένα καινοτόμο μοντέλο βαθιάς μάθησης βασισμένο σε σύνολο για την ταξινόμηση ειδήσεων ως ψευδείς ή αληθείς, χρησιμοποιώντας το σύνολο δεδομένων LIAR. Η εργασία αυτή επιδιώκει να αντιμετωπίσει την αυξανόμενη διάδοση ψευδών ειδήσεων μέσω των κοινωνικών μέσων, αξιοποιώντας προηγμένες τεχνικές μηχανικής μάθησης για την ανίχνευση και ταξινόμησή τους [28].

Στην έρευνα τους [28] δεν παρέχονται λεπτομέρειες σχετικά με τη μέθοδο κωδικοποίησης των ετικετών (labels) του συνόλου δεδομένων LIAR. Το LIAR περιλαμβάνει έξι κατηγορίες αλήθειας: “pants-fire”, “false”, “barely-true”, “half-true”, “mostly-true” και “true”. Ωστόσο, στην εν λόγω μελέτη, οι συγγραφείς αναφέρονται σε ταξινόμηση των ειδήσεων ως “fake” ή “real”, υποδεικνύοντας ότι πιθανώς εφάρμοσαν δυαδική κωδικοποίηση των ετικετών. Αν και δεν διευκρινίζεται η ακριβής αντιστοίχιση, μια κοινή προσέγγιση είναι η ομαδοποίηση των κατηγοριών “pants-fire”, “false” και “barely-true” ως “fake”, και των “half-true”, “mostly-true” και “true” ως “real”. Ωστόσο, χωρίς ρητή αναφορά στην εργασία, δεν μπορούμε να επιβεβαιώσουμε την ακριβή μέθοδο που χρησιμοποιήθηκε. Στην προσέγγισή τους, οι συγγραφείς χρησιμοποίησαν δύο διαφορετικά μοντέλα βαθιάς μάθησης λόγω της φύσης των χαρακτηριστικών του συνόλου δεδομένων. Για το κειμενικό χαρακτηριστικό ”statement”, εφαρμόστηκε ένα μοντέλο Bi-LSTM-GRU-dense, ενώ για τα υπόλοιπα χαρακτηριστικά, όπως το πολιτικό κόμμα, το θέμα και την ημερομηνία, χρησιμοποιήθηκε ένα πυκνό νευρωνικό δίκτυο. Η συνδυαστική αυτή προσέγγιση επέτρεψε την αξιοποίηση τόσο των κειμενικών όσο και των μεταδεδομένων χαρακτηριστικών για την ενίσχυση της ακρίβειας του μοντέλου [28].

Τα πειραματικά αποτελέσματα της μελέτης ήταν εντυπωσιακά. Χρησιμοποιώντας μόνο το χαρακτηριστικό ”statement”, το προτεινόμενο μοντέλο πέτυχε ακρίβεια 89.8%, ανάκληση 91.6%, ακρίβεια 91.3% και F-score 91.4%. Αυτά τα αποτελέσματα είναι αξιοσημείωτα σε σύγκριση με προηγούμενες μελέτες που χρησιμοποιούν το LIAR, υποδεικνύοντας την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας. Η εργασία αυτή συμβάλλει σημαντικά στην κατανόηση του τρόπου με τον οποίο τα συνδυαστικά μοντέλα βαθιάς μάθησης μπορούν να ενισχύσουν την ακρίβεια στην ανίχνευση ψευδών ειδήσεων [28].

Η εργασία με τίτλο “Web-Informed-Augmented Fake News Detection Model Using Stacked Layers of Convolutional Neural Network and Deep Autoencoder” [29] παρουσιάζει ένα καινοτόμο μοντέλο για την ανίχνευση ψευδών ειδήσεων, το οποίο συνδυάζει CNN με βαθύ autoencoder. Το μοντέλο αυτό εφαρμόζεται στο σύνολο δεδομένων LIAR, το οποίο περιλαμβάνει έξι κατηγορίες αλήθειας, που στην έρευνα τους προσπαθούν να προβλέψουν και τις έξι κατηγορίες και όχι να μειώσουν τις κατηγορίες σε δύο.

Η αρχιτεκτονική του μοντέλου αποτελείται από 11 στρώματα. Αρχικά, τα κείμενα μετατρέπονται σε ακολουθίες ακεραίων μέσω ενός λεξικού και στη συνέχεια ενσωματώνονται χρησιμοποιώντας το προεκπαιδευμένο μοντέλο GloVe [30], με κάθε λέξη να αναπαρίσταται από ένα διάνυσμα 100 διαστάσεων. Ακολουθούν συνελικτικά και max pooling στρώματα για την εξαγωγή χαρακτηριστικών, ενώ ένα επίπεδο flatten μετατρέπει τα πολυδιάστατα χαρακτηριστικά σε μονοδιάστατο διάνυσμα. Έχουν κάνει και επαύξηση των δεδομένων του dataset από διάφορες πηγές που τις θεωρούν έμπιστες, έτσι ώστε το ”statement”, που είναι η βασική πληροφορία του LAIR, από μια πρόταση που είναι να γίνει μια παράγραφος. Αυτό το στάδιο είναι πολύ συμαντικό για την επίτευξη των τελικών αποτελεσμάτων [29]. Την επαύξηση των δεδομένω την έκαναν επιλέγοντας κείμενο που είναι σχετικό με κάποιον τίτλο από το dataset. Χρησιμοποιόντας την Ευκλείδεια απόσταση ανάμεσα στα διανήσματα των προτάσεων τις παραγράφου, και τους τίτλου του dataset. Αυτές τις προτάσεις τις άντλησαν από την μηχανή αναζήτησης της Google.

Το μοντέλο περιλαμβάνει επίσης ένα autoencoder, το οποίο αποτελείται από ένα επίπεδο

κωδικοποίησης και ένα επίπεδο αποκωδικοποίησης. Το autoencoder εκπαιδεύεται με μη εποπτευόμενο τρόπο για να μάθει μια συμπιεσμένη αναπαράσταση των χαρακτηριστικών, βοηθώντας στην αποφυγή υπερπροσαρμογής και στη βελτίωση της γενίκευσης του μοντέλου. Η τελική έξοδος παράγεται μέσω μιας συνάρτησης SoftMax, η οποία παρέχει τις πιθανότητες για κάθε μία από τις έξι κατηγορίες αλήθειας [29]. Η μελέτη αυτή δείχνει ότι ο συνδυασμός CNN και autoencoder μπορεί να είναι αποτελεσματικός στην πολυκατηγορική ταξινόμηση ψευδών ειδήσεων, προσφέροντας μια προσέγγιση που αξιοποιεί τόσο τα χαρακτηριστικά του κειμένου όσο και την ικανότητα του autoencoder να μάθει συμπιεσμένες αναπαραστάσεις [29].

2.8 Σύγκριση ερευνών στο LAIR dataset

Στον Πίνακα 2.1 φαίνονται τα καλύτερα αποτελέσματα από κάποιες έρευνες που ασχολήθηκαν με το LAIR dataset. Παρόλο που το dataset δίνει 6 labels για στόχους, παρατηρούμε ότι αρκετές έρευνες εστιάζουν σε κάποια από αυτά και κάνουν το πρόβλημα από multiclass σε binary classification. Αυτό κάνει σίγουρα το πρόβλημα πιο εύκολο αλλά ταυτόχρονα ακυρώνει την δυνατότητα σύγκρισης των αποτελεσμάτων με άλλες έρευνες που έχουν θέσει διαφορετικούς στόχους. Για να μπορέσει να γίνει αντικειμενική σύγκριση των αποτελεσμάτων ανάμεσα σε έρευνες πρέπει οι μεθοδολογίες να έχουν εκτελεστεί με τους ίδιους στόχους αλλά και με τα ίδια ακριβώς training, validation και testing sets. Στην έρευνα του Aslam [28] γίνεται σύγκριση της μεθοδολογίας του με άλλες προηγούμενες αλλά οι προηγούμενες έρευνες είχαν θέσει και τους 6 στόχους το dataset αλλά ο Aslam μόνο τα true και false. Που αυτό καθιστά το μοντέλο του να έχει διαφορετικούς στόχος αλλά και διαφορετικό sets. Όταν κάποιος έχει μετατρέψει το πρόβλημα σε binary classification τότε το βασικό του accuracy είναι το 0.5 που είναι το αποτέλεσμα από τυχαίες προβλέψεις. Ενώ κάποιες που έχει κρατήσει και τους 6 στόχους έχει σαν τυχαίο accuracy το 0.16. Παρατηρούμε από την τεράστια διαφορά αυτών των δύο αριθμών ότι δεν μπορούμε να κάνουμε έτσι σύγκριση αποτελεσμάτων. Στην δημοσίευση του LAIR έγιναν και εκτελέσεις με CNN και LSTM πάνω στο dataset με όλους τους στόχους. Το καλύτερο accuracy που πέτυχαν ήταν με το Hybrid CNN και με την χρήση όλως των δεδομένων του dataset. Ο Bibek Upadhayay [25] έκανε χρήση του BERT και από τα διαγράμματα που παρέχει στην δημοσίευση παρατηρούμε ότι παθαίνει overfit από τις πρώτες και όλας εποχές. Και εμείς δοκιμάσαμε το BERT και είχαμε ακριβώς το ίδιο πρόβλημα και κάνοντας χρήση dropout σε όλα τα layers του BERT και του υπόλοιπου μοντέλου. Την μεγάλη διαφορά στα αποτελέσματα με όλους του στόχους την έκανε ο Ali [29] με accuracy 0.8959 και F1 score 0.6909. Αυτό το κατάφερε έχοντας κάνει augmentation του dataset με παραγράφου αντί τους τίτλους που παρέχει το LAIR. Από αυτά μπορούμε να συμπεράνουμε ότι το LAIR dataset είναι δύσκολο στην πρόβλεψη των στόχων του χωρίς να κάνουμε κάποιου είδους augmentation. Επίσης ότι την σημαντικότερη επίπτωση στα αποτελέμσατα την έχει το κείμενο που θα δώσουμε στο μοντέλο μας και όχι τόσο τα μεταδεδομένα που του παρέχουμε. Τα περισσότερα μοντέλα πετυχαίνουν καλύτερα αποτελέσματα όταν το κείμενο που τους δίνουμε είναι περίπου μια παράγραφος και με αυτόν τον τρόπο τα διευκολύνουμε στο να μην πάθουν νωρίς overfit, όπως έγινε με το BERT. Σε αυτήν την εργασία δεν έγινε χρήση μόνο των νέων παραγράφων αλλά και άλλες μεταπληροφορίες. Στην παρούσα εργασία συνδυάζουμε τη θεματική ανάλυση και την ανάλυση συναισθήματος αξιοποιώντας δύο

Πίνακας 2.1: Αποτελέσματα ερευνών με χρήση του dataset LAIR. Παρατηρούμε ότι δεν χρησιμοποιήθηκαν όλοι οι στόχοι του dataset από όλες τις έρευνες.

Μελέτη	**Παρατηρήσεις**	**Μοντέλα**	**Accuracy**	**F1-score**
Aslam [28]	Κράτησαν μόνο τις εγγραφές με label false και true. Το καλύτερο score ήταν μόνο με τα statement	Bi-LSTM	0.898	0.914
LAIR [2]	Έχουν όλους τους στόχους και τα καλύτερα score ήταν με όλα τα δεδομένα του dataset	CNN + LSTM	0.274	—
Bibek Upadhayay [25]	Μετέτρεψαν το dataset σε binary classification με το να κάνουν τους πρώτους τρεις στόχους false και τους άλλους τρεις true	BERT + CNN	0.700	0.637
Ali [29]	Κράτησαν όλους τους στόχους. Έκαναν augment του LAIR dataset με παραγράφους αντί απλά τους τίτλους από άρθρα.	CNN Autoencoder	0.896	0.691

διαφορετικές μεθόδους για κάθε περίπτωση: LDA και LSTM για την εξαγωγή θεματικών χαρακτηριστικών, και VADER και RoBERTa για τη συναισθηματική αποτίμηση. Ο στόχος δεν είναι μόνο η επίτευξη καλύτερων αποτελεσμάτων στην πρόβλεψη της αλήθειας των δηλώσεων, αλλά και η αξιολόγηση της συνεισφοράς κάθε μεθόδου ξεχωριστά στη συνολική απόδοση του τελικού μοντέλου. Παράλληλα, εξετάζουμε δύο διαφορετικές προσεγγίσεις ως προς την πρόβλεψη των έξι κατηγοριών του LIAR dataset: αφενός ως διακριτές, ανεξάρτητες κατηγορίεςόπως ακολουθείται παραδοσιακά στη σχετική βιβλιογραφία και αφετέρου ως κλιμακωτές, ιεραρχικούς στόχους, ώστε να ληφθεί υπόψη η σημασιολογική εγγύτητα μεταξύ των επιπέδων αλήθειας. Με τον τρόπο αυτό, η εργασία επιχειρεί να διερευνήσει σε βάθος ποιες πτυχές του γλωσσικού περιεχομένου συμβάλλουν ουσιαστικά στην ανίχνευση ψευδούς λόγου, αλλά και ποια προσέγγιση στόχευσης αποδίδει πιο ρεαλιστικά και σταθερά αποτελέσματα.

2.9 Επίλογος

Από πολλές έρευνες που έχουν γίνει πάνω στο θέμα [10], [11], [12], [13], παρατηρούμε ότι λόγω της φύσης του προβλήματος που έχουμε να αντιμετωπίσουμε είναι δύσκολο η πρόβλεψη του ψευδή λόγο μόνο από το κείμενο. Για αυτό πολλές έρευνες μειώνουν τις κλάσεις από 6 που έχει το LIAR σε 2. Αυτό από μόνο του βοηθάει το μοντέλο να πετύχει καλύτερα αποτελέσματα αλλά χάνει την λεπτομερείς πρόβλεψη και απομακρύνεται από την πραγματικότητα. Παρατηρώντας το LAIR οι περισσότερες προτάσεις δεν είναι τελείως ψευδής ή αληθείς αλλά και τα δύο ταυτόχρονα. Επίσης παρατηρούμε ότι χρειαζόμαστε διάφορες άλλα μεταπληροφορίες, όπως context – topic, ανίχνευση συναισθήματος, πληροφορίες για τον ομιλητή και παρελθοντικές του ομιλίες. Όσο μεγαλύτερες προτάσεις ή παραγράφους έχουμε που θέλουμε να τις ταξινομήσουμε τόσο πιο εύκολο θα είναι για ένα μοντέλο να αντλήσει δεδομένα και πληροφορίες από αυτά για να παρέχει καλύτερα αποτελέσματα.

Κεφάλαιο 3

Μεθοδολογία

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο μελετάμε πιο αναλυτικά το LAIR dataset, ποια είναι τα δεδομένα που μας προσφέρει πόσα δεδομένα έχει και άλλες πληροφορίες για την καλύτερη κατανόηση του. Στην συνέχεια γίνεται η περιγραφή των βημάτων που έγιναν για την εύρεση θεμάτων με το LDA και το LSTM. Για την ανίχνευση συναισθήματος με vocabulary based και RoBERTa και τέλος το τελικό μοντέλο που είναι ο συνδυασμός των παραπάνω ώστε να γίνει η πρόβλεψη των ψευδών ειδήσεων του LAIR.

3.2 LAIR dataset

Το LAIR [2] είναι ένα benchmark dataset, δηλαδή έχει σχεδιαστεί με τέτοιον τρόπο που ο σκοπός του είναι να μπορούν άλλοι ερευνητές να το χρησιμοποιήσουν για να κρίνουν πόσο καλό είναι το μοντέλο τους. Το dataset είναι ήδη χωρισμένο σε train, validation, test sets και έτσι μπορεί να γίνει και σύγκριση ανάμεσα σε διαφορετικά μοντέλα από διαφορετικούς ερευνητές, γιατί μπορούν να συγκρίνουν τις ίδιες μετρικές πάνω στο ίδιο ακριβώς test set.

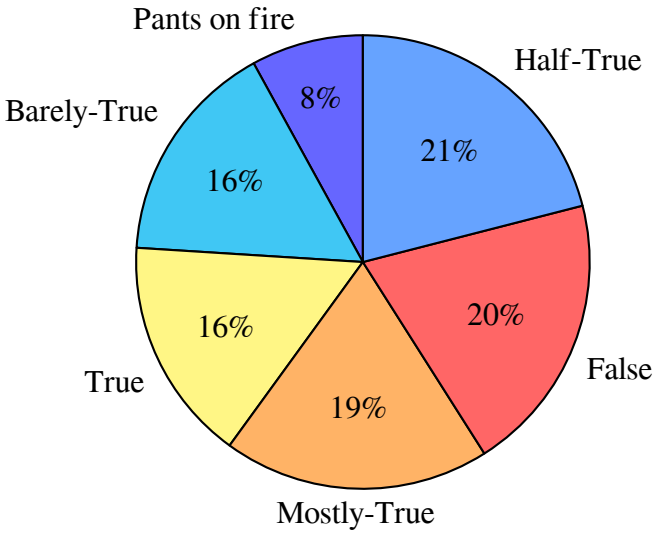
Για να κατασκευαστεί το dataset αυτό οι ερευνητές του μάζεψαν τίτλους από online άρθρα, σχετικά με την πολιτική την οικονομία, την επικαιρότητα και άλλους τομείς. Στην συνέχεια έκαναν την δική τους έρευνα προσπαθώντας να διασταυρώνουν πληροφορίες από άλλες έμπιστες πηγές σχετικά με το πόσο αληθής είναι ο τίτλος. Τελικά έφτιαξαν 6 κατηγορίες:

1. **Pants on fire**, από μια Αμερικάνικη έκφραση που λένε συνήθως τα παιδιά σε αυτόν που είναι ψεύτης αλλά δεν το παραδέχεται: “Lair, lair, pants on fire”.
2. **False**
3. **Barely-True**
4. **Half-True**
5. **Mostly-True**
6. **True**

Αυτές οι κατηγορίες είναι και τα targets του dataset. Το dataset αυτό μπορούμε να το δούμε σαν ένα classification πρόβλημα αλλά και σαν regression γιατί οι στόχοι είναι κλιμακωτές κατηγορίες. Λόγω ότι είναι δύσκολο dataset για να εκπαιδευτεί κάποιο μοντέλο πάνω του, κάποιοι ερευνητές περιορίζουν τους στόχους σε δυαδικό πρόβλημα μετατρέποντας τις πρώτες 3ς κατηγορίες σε False και τις άλλες 3ς σε true. Άλλοι εστιάζουν μόνο στα pants on fire και φτιάχνουν μοντέλα που προσπαθούν να προβλέψουν άμα είναι pants on fire ή όχι. Υπάρχουν και κάποιοι που κατασκευάζουν μοντέλα που προβλέπουν μόνο τα true και όλα τα άλλα τα διαχειρίζονται σαν false.

Το dataset περιέχει:

- Στο train set: 10240 εγγραφές.
- Στο validation set: 1284 εγγραφές.
- Στο test set: 1267 εγγραφές.

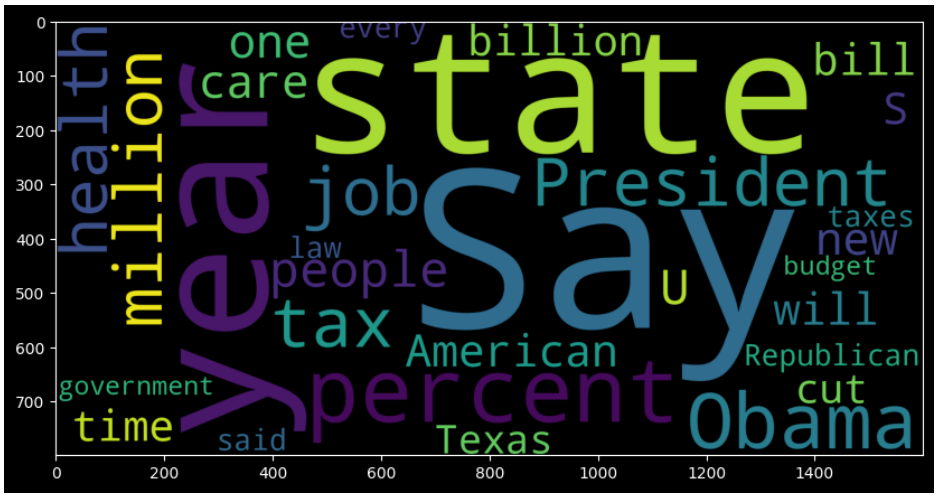


Σχήμα 3.1: Παρατηρούμε οτι δεν είναι ισάξια κατανεμημένες οι κλάσεις στο training set, με αυτήν του pants on fire να έχει σημαντικά μικρότερο ποσοστό εμφάνισης.

Γενικά οι εγγραφές του δεν είναι πολλές για να μπορέσει να μάθει ένα μοντέλο καλά το dataset. Οι προτάσεις που περιέχει είναι το κύριο χαρακτηριστικό του και είναι τίτλοι από άρθρα. Αυτό σημαίνει ότι είναι μικρές σε πλήθος λέξεων γιατί οι τίτλοι συνήθως είναι μικροί σε μέγεθος. Έτσι το μοντέλο δεν έχει αρκετή πληροφορία για να μπορέσει να μάθει. Οι τίτλοι από άρθρα συνήθως έχουν σκοπό να κεντρίσουν το ενδιαφέρον του αναγνώστη και να τον κάνει να πατήσει να δει το ολόκληρο άρθρο. Έτσι καταλήγουν να περιέχουν πολλές λέξεις που δεν έχουν κάποιο ειδικό νόημα αλλά βρίσκονται εκεί για αυτόν τον σκοπό.

Το dataset περιέχει 14 στήλες:

1. **ID:** Είναι απλά ένα μοναδικό διακριτικό για κάθε εγγραφή.
2. **Label (text):** Είναι το target, δηλαδή οι 6 κατηγορίες Pants on fire μέχρι True.
3. **Statement (text):** Είναι ο τίτλος του άρθρου. Που πολύ συχνά είναι κάτι που ειπώθηκε από κάποιον πολιτικό.
4. **Subjects (text):** Μια ή και παραπάνω λέξεις που περιγράφει το θέμα του statement.
5. **Speaker (text):** Το όνομα αυτού που είπε το statement. Άμα δεν το έχει πει κάποιος είναι κενό.
6. **Speaker's job title (text):** Ο επαγγελματικός τίτλος του speaker. Άμα δεν υπάρχει ομιλητής τότε είναι κενό.
7. **State info (text):** Η πολιτεία της Αμερικής που ειπώθηκε το statement αυτό. Άμα δεν το γνωρίζουμε ή δεν υπάρχει κάποια συγκεκριμένη πολιτεία τότε είναι κενό.
8. **Party affiliation (text):** Με μια λέξη περιγράφει τον κομματικό προσανατολισμό του speaker. Άμα δεν το γνωρίζουμε ή δεν έχουμε speaker τότε είναι κενό.
9. **Barely true counts (integer):** Το πλήθος τον εγγραφών που ο συγκεκριμένος speaker είπε κάτι που ήταν barely true. Έχει την τιμή μηδέν άμα δεν έχουμε speaker.
10. **False counts (integer):** Το πλήθος τον εγγραφών που ο συγκεκριμένος speaker είπε κάτι που ήταν false. Έχει την τιμή μηδέν άμα δεν έχουμε speaker.
11. **Half true counts (integer):** Το πλήθος τον εγγραφών που ο συγκεκριμένος speaker είπε κάτι που ήταν half true. Έχει την τιμή μηδέν άμα δεν έχουμε speaker.
12. **Mostly true counts (integer):** Το πλήθος τον εγγραφών που ο συγκεκριμένος speaker είπε κάτι που ήταν mostly true. Έχει την τιμή μηδέν άμα δεν έχουμε speaker.
13. **Pants on fire counts (integer):** Το πλήθος τον εγγραφών που ο συγκεκριμένος speaker είπε κάτι που ήταν pants on fire. Έχει την τιμή μηδέν άμα δεν έχουμε speaker.
14. **Context (text):** Με λίγες λέξεις περιγράφει την θεματική του statement. π.χ. "a tv ad", "Radio interview", "an article".



Σχήμα 3.2: Στο word cloud παρατηρούμε ότι υπάρχουν πολλές φορές οι λέξεις ”Say, state, year, precent, President, health, American”. Με αυτό το διάγραμμα μπορούμε να έχουμε μια γενική εποπτεία στο dataset.

Το LAIR [2] dataset περιέχει λέξεις που έχουν να κάνουν με τον πολιτικό λόγο και αυτό μπορεί να φανεί καλύτερα στο Σχήμα 3.2 που είναι ένα Word Cloud. Δηλαδή ένα διάγραμμα με τις 30 πιο συχνές λέξεις που περιέχονται στο dataset και το μέγεθος της λέξης δείχνει την συχνότητα της λέξης, όσο πιο μεγάλη είναι στο διάγραμμα τόσο πιο μεγάλη συχνότητα έχει.

Κατά την αναζήτηση τρόπων για την καλύτερη κατανόηση του dataset, και θέλοντας να δούμε άμα η εύρεση των θεμάτων θα βοηθούσε στο τελικό μοντέλο για την ταξινόμηση στις κατηγορίες του ψεύδους, έγινε και μια στατιστική ανάλυση για να φανεί άμα υπάρχει στατιστική συσχέτιση ανάμεσα στο θέμα της πρότασης και το target, δηλαδή σε ποια κατηγορία ψεύδους ανήκει. Άμα υπάρχει σημαντική στατιστική συσχέτιση τότε είναι πιθανό η χρήση των θεμάτων για τον τελική πρόβλεψη να βοηθήσουν στην ταξινόμηση του ψεύδους. Έτσι έγινε η χρήση του “x τετράγωνο” (chi2) και τα αποτελέσματα ήταν:

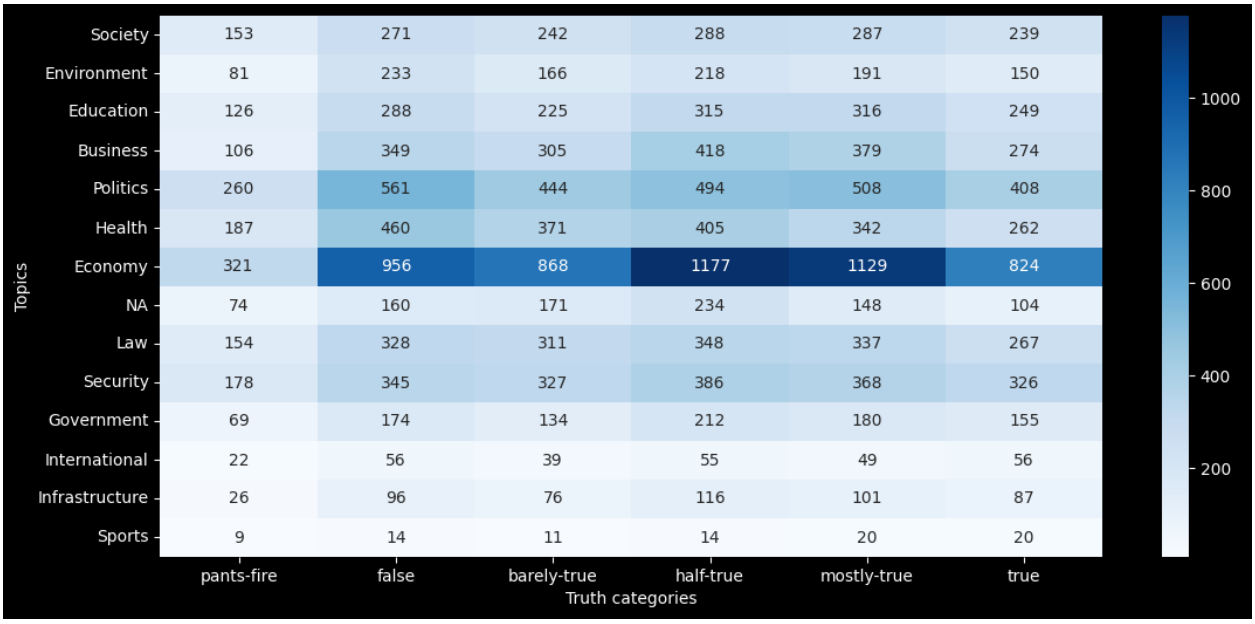
- **Chi-Square:** 1386.28
- **p-value:** 8.254610384282318e-47

Τα αποτελέσματα της ανάλυσης με τη χρήση του τεστ χ^2 αποκαλύπτουν μια ισχυρή στατιστική συσχέτιση ανάμεσα στις έξι κατηγορίες ψεύδους, στις οποίες μπορεί να ανήκει μια πρόταση, και στα αντίστοιχα θέματα (topics) που τις χαρακτηρίζουν. Η ιδιαίτερα υψηλή τιμή του χ^2 (1386.28), σε συνδυασμό με την εξαιρετικά μικρή τιμή p (8.25×10^{-47}), υποδεικνύουν ότι η κατανομή των ψευδών δηλώσεων δεν είναι τυχαία σε σχέση με τα θέματα που εξετάζονται, αλλά υπάρχει μια σαφής και σημαντική σύνδεση μεταξύ τους. Αυτό σημαίνει ότι η μηδενική υπόθεση H_0, η οποία υποθέτει ότι δεν υπάρχει συσχέτιση μεταξύ των κατηγοριών ψεύδους και των θεμάτων, απορρίπτεται με μεγάλη βεβαιότητα, καθώς $p \leq 0.05$. Συνεπως, κάθε κατηγορία ψεύδους τείνει να σχετίζεται με συγκεκριμένα θέματα, γεγονός που μπορεί να αποτυπώνει διαφορετικές στρατηγικές ή μοτίβα παρουσίασης ψευδών πληροφοριών ανάλογα με το περιεχόμενο και το πλαίσιο. Η στατιστική αυτή συσχέτιση προσφέρει ουσιαστικές

Πίνακας 3.1: Περιγράφεται αναλυτικά η αντιστοιχία ανάμεσα στα γενικά θέματα και στα θέματα του dataset. Αυτό έγινε για απλούστευση των θεμάτων με τελικό σκοπό την διευκόλυνση των μοντέλων στην ταξινόμηση των θεμάτων.

Γενικό υπερ-θέμα	**Dataset θέμα**
Politics	foreign-policy candidates-biography ethics voting-record elections campaign-advertising bipartisanship campaign-finance kagan-nomination debates obama-birth-certificate bush-administration polls
Economy	economy poverty federal-budget bankruptcy income taxes stimulus deficit debt wealth welfare pensions retirement state-finances trade tourism lottery agriculture state-budget gas-prices city-budget
Health	health-care public-health medicare medicaid drugs marijuana ebola autism food-safety hunger human-rights disability food animals Alcohol
Environment	energy climate-change environment water oil-spill cap-and-trade natural-disasters fires weather
Society	diversity women pop-culture gays-and-lesbians sexuality civil-rights islam lgbt homeless religion census marriage gambling abortion population families children fake-news
Security	homeland-security terrorism military public-safety veterans crime guns defense patriotism nuclear social-security
Education	education history science technology college school research pundits
Government	government-regulation government-efficiency public-service city-government county-government transparency urban county-budget states
Law	legal-issues immigration death-penalty criminal-justice law lawsuits rights privacy redistricting congress congressional-rules supreme-court financial-regulation market-regulation
Business	small-business labor unions workers corporations business startups consumer-safety job-accomplishments jobs
Infrastructure	infrastructure transportation housing roads bridges florida recreation
International	afghanistan iraq israel china space
Sports	baseball sports

ενδείξεις για την κατανόηση και την ταξινόμηση των ψευδών ειδήσεων, ενισχύοντας την αξιοπιστία των μοντέλων που επιχειρούν να διαχωρίσουν και να αναλύσουν τέτοιου είδους δεδομένα. Επιπλέον, η γνώση αυτών των σχέσεων μπορεί να συμβάλει στην ανάπτυξη πιο εξειδικευμένων εργαλείων εντοπισμού ψεύδους, εστιάζοντας όχι μόνο στη γλωσσική ανάλυση αλλά και στο θεματικό πλαίσιο, κάτι που είναι κρίσιμο σε εφαρμογές όπως η ανίχνευση παραπληροφόρησης και η αξιολόγηση της αξιοπιστίας περιεχομένου. Συνολικά, η στατιστική αυτή συσχέτιση προσδίδει βάθος στην ανάλυση και ανοίγει νέους δρόμους για την περαιτέρω μελέτη των σχέσεων μεταξύ θεμάτων και τύπων ψεύδους.



Σχήμα 3.3: Στο heatmap παρατηρούμε στον άξονα Υ 13 γενικά θέματα και ένα NA που είναι το άγνωστο θέμα. Στον άξονα Χ τις κατηγορίες ψεύδους. Το ποιο hot θέμα είναι το Economy και έχει τα περισσότερα half-true. Από τα διάσπαρτα hot points που έχει μπορούμε να κατανοήσουμε την πολυπλοκότητα του dataset.

Για την καλύτερη μελέτη και στην συνέχεια διευκόλυνση του LSTM για την πρόβλεψη των θεμάτων, έγινε μια ομαδοποίηση των θεμάτων που παρέχει το LAIR με την στήλη Subjects. Η ομαδοποίηση αυτή έχει γίνει με χειροκίνητο τρόπο και με την εννοιολογική συνάφεια των θεμάτων. Σε κάποια θέματα δεν ήταν ευδιάκριτος ο εννοιολογικός τους χαρακτήρας, γιατί είναι γενικά θέματα που θα μπορούσαν να σταθούν σε διαφορετικές θεματικές. Αυτό οδηγεί σε κάποιους συμβιβασμούς και υποκειμενικές αποφάσεις για την αντιστοίχιση αυτών των θεμάτων. Για παράδειγμα στον Πίνακα 3.1, για το υπερ-θέμα Health, είναι πιο ξεκάθαρος ο συσχετισμός του με τα θέματα health-care, medicate, drugs ενώ το Alcohol είναι πιο γενικό. Θα μπορούσε, εκτός από το Health, να συσχετιστεί με το Society π.χ ”Στατιστικές έρευνες έδειξαν την αύξηση κατανάλωσης αλκοόλ από νεαρούς.” ή και με το Economy π.χ ”Ο φόρος στα αλκοολούχα ποτά θα αυξηθεί τον επόμενο μήνα.”. Κατα την ομαδοποίηση υπήρχαν περιπτώσεις που ένα θέμα του dataset θα μπορούσε να συσχετιστεί με παραπάνω από ένα από τα 13 υπερ-θέματα, γιατί το νόημα του είναι τόσο γενικό που το επιτρέπει. Αποφασίσαμε κάθε θέμα να εμφανίζεται μόνο σε υπερ-θέμα, έτσι κάθε θέμα είναι συσχετισμένο μόνο με ένα από τα 13 υπερ-θέματα. Στην πραγματικότητα αυτά που παρέχει το dataset δεν είναι τόσο θέματα ή περιεχόμενο (context) αλλά περισσότερο ”tags” που συναντάμε σε web sites.

Στο Σχήμα 3.3 απεικονίζεται ένα heatmap που στον άξονα Υ έχουμε μια ομαδοποίηση των θεμάτων του dataset σε 13 πιο γενικά υπερ-θέματα Πίνακας 3.1. Ο άξονας Χ περιλαμβάνει τις 6 κατηγορίες ψεύδους που παρέχει το dataset. Από τα αριστερά προς τα δεξιά είναι από το πιο ψευδής μέχρι το πιο αληθής. Παρατηρούμε ότι το πιο hot θέμα είναι το Economy που περιέχει τα περισσότερα ψέματα αλλά και αλήθειες. Άλλα επίκεντρα θέματα είναι τα Politics, Health, Law, Security και Business. Το θέμα NA είναι όσα από τα θέματα του dataset δεν μπορούσαν να

αντιστοιχηθούν σε κανένα από τα 13 υπερ-θέματα, είτε γιατι ήταν πολύ γενικά είτε γιατί δεν είχαν αρκετά ευδιάκριτο νόημα. Από την κλίμακα του ψεύδους παρατηρούμε ότι το half-true είναι το πιο hot. Αυτό σημαίνει ότι οι περισσότερες προτάσεις περιέχουν ψέμα και αλήθεια μαζί. Δεύτερο είναι το mostly-true και τρίτο το false. Το γενικό συμπέρασμα από το heatmap είναι οτι το dataset είναι αρκετά μπερδεμένο παρατηρώντας τις κατηγορίες ψεύδους και τα θέματα. Αυτό είναι ένα στοιχειο που επιβεβαιώνει ακόμα μια φορά την πολυπλοκότητα του dataset.

3.3 Pre-processing

Η διαδικασία του pre-processing στα datasets που προορίζονται για την εκπαίδευση γλωσσικών μοντέλων αποτελεί ένα από τα πιο κρίσιμα στάδια πριν την έναρξη της εκμάθησης. Τα δεδομένα κειμένου, σε αντίθεση με άλλα είδη δεδομένων, περιλαμβάνουν ένα ευρύ φάσμα από παραλλαγές, θόρυβο και ασυνέπειες που μπορούν να επηρεάσουν σημαντικά την ποιότητα του μοντέλου. Η απομάκρυνση ανεπιθύμητων χαρακτήρων, η κανονικοποίηση (όπως π.χ. η μετατροπή όλων των λέξεων σε πεζά), η διόρθωση ορθογραφικών λαθών και η διαχείριση των διαφορετικών μορφών γλωσσικής έκφρασης είναι ενέργειες που συμβάλλουν στην ομοιογένεια των δεδομένων. Εάν δεν υπάρξει η κατάλληλη προεργασία, το μοντέλο κινδυνεύει να μάθει λανθασμένα μοτίβα ή να υπερεστιάσει από θορυβώδεις πληροφορίες, γεγονός που υποβαθμίζει την ικανότητά του να γενικεύει σωστά σε νέα, άγνωστα δεδομένα.

Επιπλέον, το pre-processing διασφαλίζει ότι το training dataset αντανακλά με συνέπεια τη γλωσσική ποικιλία και πολυπλοκότητα που το μοντέλο καλείται να χειριστεί. Η σωστή επιλογή και φίλτραρισμα των πηγών, αλλά και η ισορροπημένη εκπροσώπηση διαφορετικών θεμάτων και ύφους, αποτελούν θεμέλιο για την παραγωγή ενός υπεύθυνου και αποτελεσματικού γλωσσικού μοντέλου. Ακόμα και λεπτές διαφοροποιήσεις στα στάδια του pre-processing μπορεί να έχουν σημαντική επίπτωση στην τελική απόδοση του μοντέλου, τόσο από πλευράς ακρίβειας όσο και δεοντολογίας. Έτσι, δεν είναι υπερβολή να ειπωθεί ότι ένα καλό pre-processing μπορεί να καθορίσει την επιτυχία ή αποτυχία ολόκληρου του μοντέλου.

Στο LAIR έγινε pre-processing για την στήλη του statement. Αυτό γίνεται ώστε να αφαιρέσουμε περιττές λέξεις που δεν προσφέρουν κάποιο νόημα για τα μοντέλα μας αλλά και να κανονικοποιήσουμε τις προτάσεις σε συγκεκριμένες λέξεις. Επίσης μπορούμε να αφαιρέσουμε λέξεις με πολύ μεγάλη ή πολύ μικρή συχνότητα εμφάνισης, θεωρώντας ότι είναι θόρυβος. Ένα ακόμα θετικό με το epre-processing είναι ότι μικραίνει το μέγεθος των δεδομένων το dataset και έτσι είναι πιο εύκολα διαχειρίσιμο από εμάς και από την GPU.
Σε αυτήν την εργασία έγιναν πειραματισμοί με τις παρακάτω τεχνικές pre-processing:

- **Lower case:** Είναι μετατροπή όλων των χαρακτήρων ενός κειμένου σε πεζά (lower case) αποσκοπεί στην ενοποίηση λέξεων που εμφανίζονται με διαφορετική μορφή κεφαλαίων, μειώνοντας την πολυπλοκότητα του λεξιλογίου και αποτρέποντας την ξεχωριστή αντιμετώπιση όρων που είναι ουσιαστικά ίδιοι, όπως ”Computer” και ”computer”. Βοηθάει στην κανονικοποίηση του dataset.
- **Remove stop words** Η απομάκρυνση των λεγόμενων stop words, δηλαδή λέξεων που δεν φέρουν ιδιαίτερο σημασιολογικό βάρος (όπως ”to”, ”a”, ”at”), βοηθά στη μείωση

της πληθικότητας του κειμένου και ενισχύει την αποτελεσματικότητα των μοντέλων, τα οποία έτσι επικεντρώνονται στους πιο ουσιώδεις όρους. Αυτές οι λέξεις είναι που βοηθάνε στην σύνταξη και την δομή της πρότασης αλλά δεν προσθέτουν κάποιο ειδικό νόημα και έτσι δεν βοηθάνε τα μοντέλα μας. Υπάρχουν μοντέλα όπως αυτά που μπορούν να κάνουν attentions - Transformers που μπορούν να εκμεταλλευτούν τα stop words.

- **Lemmatize:** Η λεμματοποίηση είναι η διαδικασία μετατροπής μιας λέξης στη βασική, γραμματικά σωστή μορφή της (λήμμα), λαμβάνοντας υπόψη το συντακτικό της ρόλο. Σε αντίθεση με πιο απλές μεθόδους, διατηρεί τη σημασιολογική συνοχή των λέξεων και είναι ιδιαίτερα χρήσιμη για την ανάλυση περιεχομένου με μεγαλύτερη ακρίβεια. Αυτή η διαδικασία είναι πιο απλή στην Αγγλική γλώσσα γιατί δεν υπάρχουν πολλές περιπτώσεις που αλλάζουν οι καταλήξεις των λέξεων.

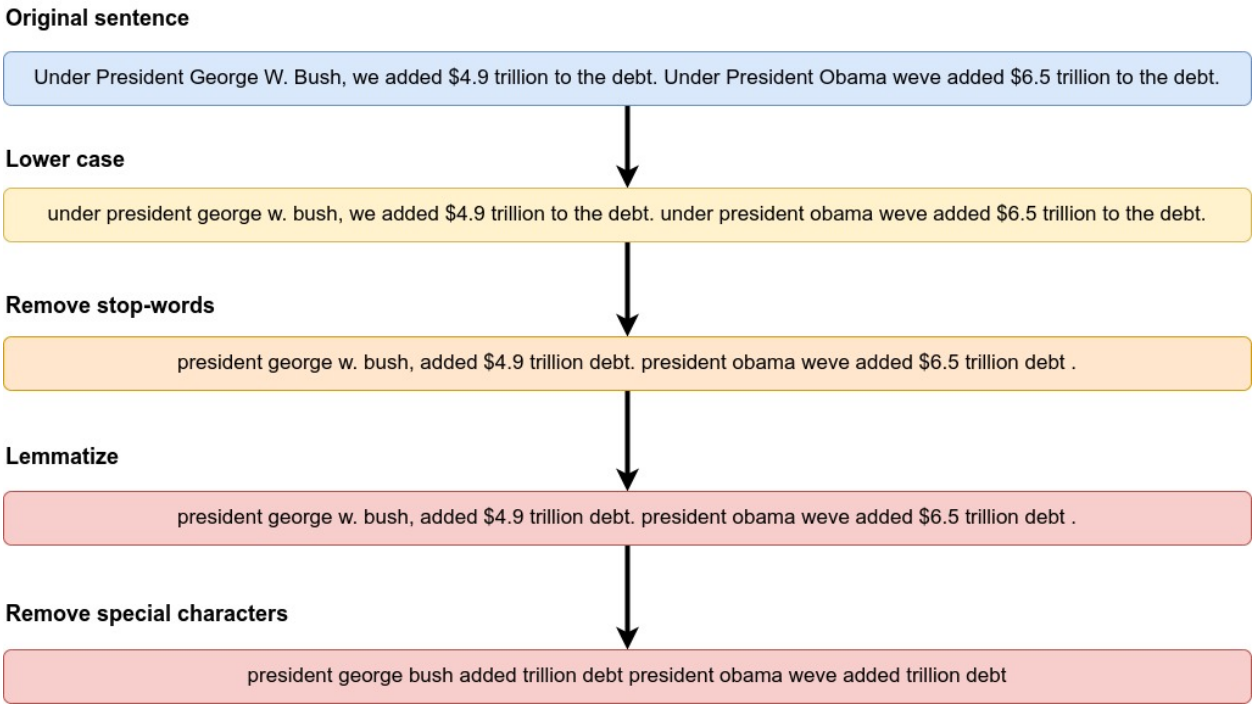
- **Stemming:** Η τεχνική του stemming αποσκοπεί στην αποκοπή καταλήξεων από λέξεις, ώστε να περιορίζονται στη ρίζα τους, ακόμα κι αν αυτή δεν αποτελεί έγκυρη λέξη. Αν και πρόκειται για πιο ”βίαιη” προσέγγιση σε σχέση με τη λεμματοποίηση, μπορεί να συμβάλλει σημαντικά στη μείωση της διασποράς όρων σε μεγάλα σύνολα δεδομένων και στην μείωση του μεγέθους του dataset, ώστε να χωράει πιο εύκολα στην RAM και στην GPU.

- **Filter special characters:** Η απομάκρυνση ειδικών χαρακτήρων όπως παρενθέσεις, εισαγωγικά ή σημεία στίξης αποτελεί μια βασική μορφή καθαρισμού των προτάσεων, είναι απαραίτητη για την αποφυγή θορύβου και την εστίαση σε γλωσσικές μονάδες που φέρουν πραγματικό νοηματικό φορτίο. Επίσης μπορεί να γίνει και αφαίρεση αριθμών ή και γενικά όλων των χαρακτήρων που δεν είναι στο αλφάβητο. Επίσης σε αυτό το στάδιο αφαιρούνται και πολλαπλά κενά, άμα υπάρχουν.

- **Remove noise:** Ο όρος ”θόρυβος” στο κείμενο αναφέρεται σε λέξεις που εμφανίζονται πολύ σπάνια ή πολύ συχνά στο dataset. Άμα τις αφαιρέσουμε τότε θα βοηθήσουμε το μοντέλο στην πρόληψη overfitting και του bias σε συγκεκριμένες λέξεις που αυτό οδηγεί σε καλύτερη γενίκευση.

Επίσης όλες οι προτάσεις μετατράπηκαν σε σταθερό μήκος (πλήθος λέξεων). Επιλέχθηκε το μήκος 100. Από μια ανάλυση που έγινε στην κατανομή των προτάσεων, που είναι pre-processed, φάνηκε ότι οι περισσότερες προτάσεις περιέχουν περίπου 10 λέξεις. Παρόλα αυτά υπήρχαν και κάποιες προτάσεις που είχαν πολύ μεγαλύτερο αριθμό λέξεων από 10 για αυτόν τον λόγο επιλέχθηκε το σταθερό μήκος των 100 λέξεων, που κάλυπτε σχεδόν όλο το εύρος. Ο μέσος όρος λέξων είναι 10, είναι λογικό γιατί είναι τίτλοι από άρθρα και ως συνήθως οι τίτλοι έχουν μικρό μέγεθος, που τελικά γίνεται ακόμα μικρότερα μετά από το pre-processing. Παρατηρούμε στο Σχήμα 3.4 πως η αρχική πρόταση μπορεί να απλοποιηθεί, να κανονικοποιηθεί και στο τέλος να έχει μόνο τις πιο σημαντικές λέξεις ώστε να καταφέρει το μοντέλο να εστιάσει σε αυτές και να κατανοήσει καλύτερα το νόημα της πρότασης.

Στο pre-processing έχει πολύ σημαντικό ρόλο η σειρά με την οποία θα εφαρμοστούν οι τεχνικές γιατί αλλάζοντας την σειρά θα αλλάξει και το τελικό αποτέλεσμα. Στην εργασία έγιναν πολλά πειράματα με διαφορετικούς συνδυασμούς των τεχνικών και καταλήξαμε σε διαφορετικό



Σχήμα 3.4: Στο word cloud παρατηρούμε ότι υπάρχουν πολλές φορές οι λέξεις ”Say, state, year, precent, President, health, American”. Με αυτό το διάγραμμα μπορούμε να έχουμε μια γενική εποπτεία στο dataset.

pre-processing ανάλογα με την μεθοδολογία που θα ακολουθούσε στην συνέχεια, όπως LDA, Sentiment Analysis ή Classification. Στον Πίνακα 3.2 φαίνονται πιο αναλυτικά ποιες τεχνικές pre-processing χρησιμοποιήθηκαν για ποια μεθοδολογία.

Εν κατακλείδι η διαδικασία του pre-processing αποτελεί θεμέλιο λίθο στην επεξεργασία φυσικής γλώσσας και τη δημιουργία αποτελεσματικών γλωσσικών μοντέλων. Μέσω του προσεκτικού καθαρισμού και της ομογενοποίησης των δεδομένων, το pre-processing εξασφαλίζει ότι τα δεδομένα εισόδου είναι όσο το δυνατόν πιο καθαρά, συνεπή και αντιπροσωπευτικά της γλώσσας που πρόκειται να αναλυθεί. Τεχνικές όπως η μείωση των χαρακτήρων σε πεζά, η αφαίρεση κοινών και μη πληροφοριακών λέξεων (stop words), η λεμματοποίηση και η στέλεχωση, καθώς και το φιλτράρισμα ειδικών χαρακτήρων και ο περιορισμός θορύβου, συμβάλλουν στο να μειωθεί η πολυπλοκότητα των δεδομένων, ενώ παράλληλα διατηρούν το ουσιαστικό νοήμα. Η σημασία αυτής της φάσης δεν περιορίζεται μόνο στη βελτίωση της ποιότητας των δεδομένων, αλλά επηρεάζει άμεσα και την απόδοση των μοντέλων, καθώς τα καθαρότερα και πιο συνεπή δεδομένα οδηγούν σε πιο ακριβείς και αξιόπιστες προβλέψεις. Επιπλέον, η σωστή διαχείριση του pre-processing επιτρέπει τη διαχείριση διαφορετικών γλωσσικών ιδιαιτεροτήτων και τη μείωση των σφαλμάτων που προκύπτουν από θόρυβο ή ασυνέπειες στα δεδομένα. Συνολικά, το pre-processing λειτουργεί ως γέφυρα ανάμεσα στα ακατέργαστα δεδομένα και την αποτελεσματική ανάλυση, καθιστώντας το αναπόσπαστο και κρίσιμο στάδιο σε κάθε γλωσσικό έργο και εφαρμογή.

Πίνακας 3.2: Παρατηρούμε ότι κάποιες βασικές τεχνικές χρησιμοποιήθηκαν σε όλες τις μεθοδολογίες και κάποιες που είναι πιο ειδικές φαίνονται χρήσιμες μόνο σε κάποιες μεθοδολογίες.

Pre-processing	**Μεθοδολογίες**				
Τεχνικές	**LDA θέματα**	**LSTM θέματα**	**Lexical Sentiment Analysis**	**RoBERTa Sentiment Analysis**	**Τελικό μοντέλο**
Lower case	✓	✓	✓	✓	✓
Remove stop words	✓	✓	✓	✓	✓
Lemmatize	✓	✓	✓	✓	✓
Stemming	✓	✓	—	—	—
Filter special characters	✓	✓	✓	✓	✓
Remove noise	✓	✓	—	—	—

3.4 Latent Dirichlet allocation

Όπως παρουσιάστηκε προηγουμένως από άλλες δημοσιεύσεις στο LAIR η εξαγωγή θεμάτων μπορεί να βοηθήσει στην καλύτερη πρόβλεψη και ταξινόμηση των ψευδών κατηγοριών. Το LIAR [2] έχει ήδη μια στήλη που θα μπορούσαμε να την δούμε σαν θέματα για κάθε πρόταση. Παρόλα αυτά στην εργασία αυτή έχουμε δοκιμάσει να παράξουμε τα δικά μας θέματα. Αυτό το κάναμε γιατί σε ένα πιο ρεαλιστικό σενάριο, και όχι στο benchmark dataset LIAR, δεν θα είχαμε το θέμα για κάθε πρόταση και θα έπρεπε να το βρούμε εμείς. Πρώτα έγιναν δοκιμές με το LDA για να δούμε πως θα μπορέσει να ομαδοποιήσει τις λέξεις από τις προτάσεις με τέτοιο τρόπο ώστε να έχουν ένα κοινό νόημα για να παραχθεί κάποιο θέμα.
Το LDA [3] έχει κάποιες βασικές παραμέτρους που μας βοηθάνε να το κατευθύνουμε ώστε να μας παράξει τα θέματα έτσι όπως τα θέλουμε:

- **Αριθμός θεμάτων:** Ο αριθμός θεμάτων είναι πολύ βασικός γιατί όσο το ορίσουμε τόσα ακριβώς θέματα θα προσπαθήσει να φτιάξει. Πρακτικά κάθε θέμα είναι μια ομάδα (cluster) από λέξεις που η κάθε λέξη έχει έναν συντελεστή βαρύτητας.

- **Άλφα:** Είναι ένας συντελεστής που ελέγχει κατά πόσο ένα κείμενο μπορεί να περιέχει πολλά θέματα ή λιγότερα. Όσο πιο μεγάλος είναι ο αριθμός αυτός τόσο πιο γενικά γίνονται τα θέματα και μπορούν να ταιριάξουν περισσότερα σε κάποια πρόταση.

- **Ήτα:** Ελέγχει τις λέξεις που περιέχονται στα θέματα. Όσο μεγαλύτερος αριθμός τόσο ένα θέμα θα περιέχει περισσότερες λέξεις που αυτό μπορεί να οδηγήσει και στο γεγονός ότι κάποιες από αυτές τις λέξεις θα τις συναντήσουμε και σε πολλά άλλα θέματα. Ενώ άμα είναι μικρός αριθμός θα προσπαθήσει κάθε λέξη να περιοριστεί σε λιγότερα θέματα.

Δεν υπάρχει κάποιος τρόπος να γνωρίζουμε από πριν τις καλύτερες τιμές για το dataset μας. Αυτό που μπορούμε να κάνουμε είναι να δοκιμάσουμε πολλές διαφορετικές τιμές

και συνδυασμούς αυτών των παραμέτρων με cross validation και παρατηρώντας κάποιες μετρικές, να αποφασίσουμε ποιον συνδυασμό θα κρατήσουμε. Μια βασική μετρική είναι το coherence score που μετράει κατά πόσο οι λέξεις σε κάθε θέμα έχουν σημαντική σημασιολογική ομοιότητα. Όσο μεγαλύτερος ο αριθμός του coherence score τόσο το καλύτερο.

Αρχικά πριν την εκτέλεση του LDA έγινε pre-processing στις προτάσεις του dataset όπως περιγράφεται στον Πίνακα 3.2. Ακολούθησαν πολλές δοκιμές με διαφορετικές τιμές για τις παραμέτρους του LDA, που είναι, ο αριθμός των θεμάτων που θα κατασκευαστούν, το άλφα και το ήτα. Για τον υπολογισμό της αποτελεσματικότητας των θεμάτων που παρήγαγε η κάθε δοκιμή χρησιμοποιήθηκε το coherence score CV [31]. Η μέτρηση του coherence score, και ειδικά η χρήση του CV coherence [31], αποτελεί μια από τις πιο αξιόπιστες μεθόδους για την αξιολόγηση της ποιότητας των θεμάτων που εξάγονται μέσω του αλγορίθμου LDA [3]. Το CV coherence score [31] βασίζεται σε ένα συνδυαστικό μοντέλο που ενσωματώνει στατιστικά μέτρα, όπως η συνάφεια μεταξύ των όρων ενός θέματος, η συχνότητα συν-εμφάνισης τους σε παράθυρα κειμένου και η χρήση μεθόδων επαύξησης πληροφορίας. Πρακτικά, το CV coherence [31] αποσκοπεί στο να προσεγγίσει τον τρόπο με τον οποίο ένας άνθρωπος θα έκρινε τη συνοχή ενός συνόλου λέξεων που περιγράφουν ένα θέμα. Όσο υψηλότερη είναι η τιμή του score, τόσο μεγαλύτερη είναι η θεματική συνοχή, υποδηλώνοντας πως οι λέξεις που σχετίζονται με το κάθε θέμα έχουν λογική και εννοιολογική συνοχή μεταξύ τους. Η αξιολόγηση με το CV coherence [31] καθίσταται κρίσιμη κατά τη διαδικασία επιλογής του ιδανικού αριθμού θεμάτων για ένα μοντέλο LDA, καθώς συχνά παρατηρείται πως η αύξηση του αριθμού των θεμάτων οδηγεί σε overfitting και μείωση της συνοχής. Έτσι, το CV coherence χρησιμοποιείται συστηματικά για τον προσδιορισμό της βέλτιστης παραμετροποίησης ενός μοντέλου θεματικής μοντελοποίησης, παρέχοντας έναν ισχυρό ποσοτικό δείκτη για την ποιότητα της εξαγόμενης πληροφορίας. Για αυτόν τον λόγο έγιναν πολλές δοκιμές και ο υπολογισμός του CV κάθε φορά με την τεχνική του cross-validation που θα μας δώσει μια πιο σωστή και σφαιρική απόδοση των θεμάτων του LDA.

3.5 Εξαγωγή θεμάτων με LSTM

Το LDA [3] μπορεί να δώσει αποτελέσματα και χωρίς να υπάρχει κάποιος στόχος. Τα αποτελέσματα του όμως είναι λογικό να μην είναι τόσο καλά όσο άμα εκπαιδεύαμε ένα supervised μοντέλο. Λόγο ότι το dataset παρέχει μια στήλη που μοιάζει με θέματα σκεφτήκαμε να εκπαιδεύσουμε και ένα μοντέλο supervised που θα έχει σαν στόχο αυτά τα θέματα. Τα θέματα αυτά που παρέχει το dataset είναι παραπάνω από 100 και για αυτό έγινε χρήση των γενικών θεμάτων Πίνακας 3.1. Το LSTM είναι ένα είδος νευρωνικού δικτύου που χρησιμοποιείται κυρίως για την επεξεργασία ακολουθιακών δεδομένων, όπως κείμενα, σειρές χρόνου ή ήχος. Σε αντίθεση με τα παραδοσιακά Recurrent Neural Networks (RNN), το LSTM έχει τη δυνατότητα να ”θυμάται” πληροφορίες για μεγάλο χρονικό διάστημα, χάρη στη δομή του που περιλαμβάνει κυψέλες μνήμης και μηχανισμούς πύλης (gates) που ελέγχουν τη ροή των πληροφοριών. Αυτό το καθιστά ιδιαίτερα χρήσιμο σε εφαρμογές όπως η μηχανική μετάφραση, η αναγνώριση φωνής, η ανάλυση συναισθήματος και η πρόβλεψη τιμών σε χρηματοοικονομικά δεδομένα. Με λίγα λόγια, το LSTM επιτρέπει στα συστήματα τεχνητής νοημοσύνης να αντιλαμβάνονται τη σημασία της χρονικής ακολουθίας και των συσχετίσεων

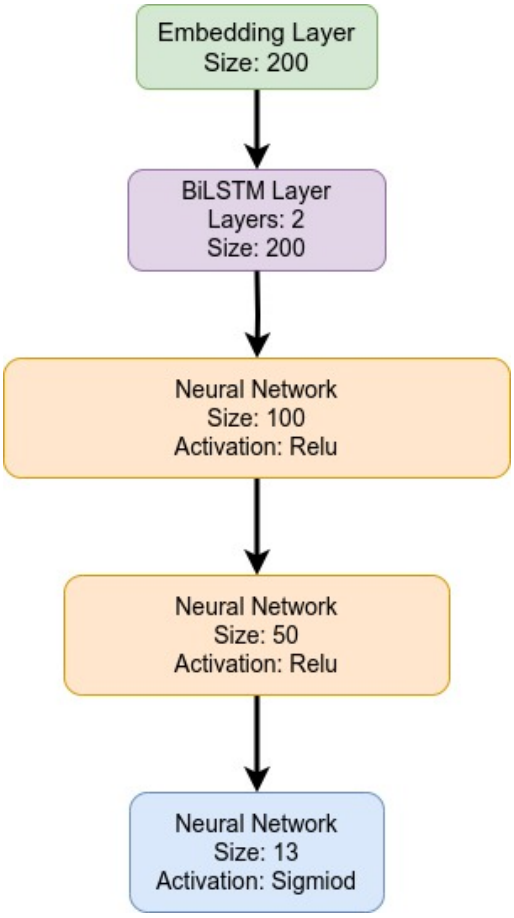
μέσα σε δεδομένα που εξελίσσονται με τον χρόνο. Το LSTM έχει αποδειχθεί ότι μπορεί να βοηθήσει πολύ στην μηχανική μάθηση προτάσεων γιατί μπορεί να απομνημονεύσει πληροφορίες από διαφορετικές λέξεις μέσα στην ίδια πρόταση. Επίσης άμα χρησιμοποιηθεί και σαν bidirectional έχει την δυνατότητα να λάβει υπόψη του και λέξεις με την ανάποδη σειρά από αυτή που έχει μάθει. Στις προτάσεις έγινε pre-processing όπως περιγράφεται από τον Πίνακα 3.2. Στο Σχήμα 3.5 απεικονίζεται το μοντέλο LSTM για την πρόβλεψη των θεμάτων. Υπάρχει στην αρχή ένα στρώμα (layer) embeddings που αντλούν πληροφορίες από το sentence column του dataset για την καλύτερη εκμάθηση του μοντέλου. Τα embeddings μπορούν να βοηθήσουν επίσης στην καλύτερη κατανόηση της πρότασης από το μοντέλο και συγκεκριμένα στην περίπτωση αυτή βοηθάνε το επόμενο layer που είναι το LSTM. Σε αντίθετη περίπτωση θα έπρεπε να λαμβάνει το LSTM τις προτάσεις απλά κωδικοποιημένες (encoded), που αυτό δεν βοηθάει στην εκπαίδευση. Οι προτάσεις είναι encoded με την τεχνική του δείκτη (index). Δηλαδή ή κάθε λέξη του train set λαμβάνει ένα μοναδικό index και η κάθε πρόταση γίνεται μια ακολουθία από indexes σύμφωνα με τις λέξεις που περιέχει. Στον Πίνακα 3.3 βρίσκονται τα υπόλοιπα χαρακτηριστικά του μοντέλου ή αλλιώς υπερπαράμετροι. Όπως είναι λογικό μια πρόταση μπορεί να περιγραφεί με παραπάνω από ένα θέμα, έτσι και το μοντέλο μπορεί στις προβλέψεις του να δώσει παραπάνω από ένα θέμα για μια συγκεκριμένη πρόταση, αυτό ονομάζεται multilabel classification.

Λόγω ότι οι κλάσεις δεν είναι ισορροπημένες σε πλήθος έχει γίνει η εκπαίδευση με κανονικοποιημένα βάρη (normalized weights) για κάθε μια κλάση. Το LAIR dataset παρέχει ήδη validation και testing set. Έτσι για την εκπαίδευση έγινε χρήση του validation για την καλύτερη επίβλεψη της πορείας του μοντέλου. Έχουν χρησιμοποιηθεί σημεία αναφοράς (checkpoints) που παρέχει το pytorch lighting, ώστε να καταγράφονται snapshots του μοντέλου ανάλογα την συνθήκη που έχουμε ορίσει εμείς. Εμείς ορίσαμε να αποθηκεύονται 4 snapshot κατά το μικρότερο validation error, έτσι άσχετα με το αν το μοντέλο προς το τέλος των εποχών έχει πάθει overfit, εμείς να έχουμε αποθηκεύσει το snapshot του σε πιο παλιά εποχή που το validation error ήταν καλύτερο. Τέλος επιλέχθηκε το καλύτερο μοντέλο λαμβάνοντας υπόψη την απόδοση του στο test set και έχοντας για μετρική το accuracy.

Συμπερασματικά, γίνεται φανερό πως η χρήση του μοντέλου LSTM μπορεί να αποτελέσει ένα ιδιαίτερα ισχυρό εργαλείο στην πρόβλεψη των θεμάτων μέσα σε προτάσεις. Το LSTM, χάρη στην αρχιτεκτονική του επιτρέπει την αποθήκευση και διατήρηση μακροχρόνιων εξαρτήσεων στη ροή της πληροφορίας, καταφέρνει να συλλάβει τη σημασιολογική συνοχή και τη χρονική αλληλουχία λέξεων και φράσεων. Αυτό το χαρακτηριστικό είναι κρίσιμο για την κατανόηση του ευρύτερου νοήματος ενός κειμένου και συνεπως για την αποτελεσματική αναγνώριση των θεμάτων. Μέσω της εκπαίδευσης, το LSTM μπορεί να "μάθει" μοτίβα που συνδέουν συγκεκριμένες λέξεις ή φράσεις με αντίστοιχα νοηματικά πλαίσια, επιτρέποντας έτσι την ακριβέστερη κατηγοριοποίηση και πρόβλεψη του εκάστοτε θέματος.

3.6 Ανίχνευση συναισθήματος με χρήση λεξικού

Η ανάλυση συναισθήματος, ειδικά όταν εφαρμόζεται μέσω λεξικοκεντρικών (lexical-based) μεθόδων, έχει αναδειχθεί ως ένα από τα χρήσιμα εργαλεία στην προσπάθεια πρόβλεψης και εντοπισμού ψευδών ειδήσεων. Στην προσέγγιση αυτή, γίνεται χρήση προτυποποιημένων



Σχήμα 3.5: Το σχήμα περιγράφει αναλυτικά το μέγεθος και το σχήμα του μοντέλου για την πρόβλεψη των θεμάτων με την χρήση embeddings, LSTM και neural networds. Στο στρώμα εξόδου (output layer) έχει activation function sigmoid λόγο του loss function που είναι το binary cross entropy.

λεξικών που περιλαμβάνουν λέξεις με προκαθορισμένο συναισθηματικό φορτίο (θετικό, αρνητικό ή ουδέτερο), και βάσει της συχνότητας και της έντασης τους στο κείμενο, αποδίδεται ένα γενικό συναισθηματικό προφίλ. Η παρατήρηση ότι τα ψευδώς ειδησεογραφικά κείμενα συχνά χαρακτηρίζονται από υπερβολικό συναισθηματισμό, φορτισμένη γλώσσα και επιτηδευμένο ύφος, έχει οδηγήσει στην αξιοποίηση τέτοιων τεχνικών για την ανίχνευση τους. Αυτή η μέθοδος έχει το πλεονέκτημα της απλότητας και της διαφάνειας, καθώς επιτρέπει την ερμηνεία των αποτελεσμάτων έχοντας τις ίδιες τις λέξεις του κειμένου. Η μέθοδος αυτή μας βοηθάει ιδιαίτερα όταν δεν έχουμε τους targets που να περιγράφουν το συναίσθημα, γιατί είναι unsupervised μέθοδος, που αυτή είναι και η περίπτωση για το LAIR dataset.

Όπως και με όλες τις προηγούμενες μεθοδολογίες έγινε pre-processing στα δεδομένα με τις τεχνικές που φαίνονται στον Πίνακα 3.2. Για την ανίχνευση συναισθήματος με lexical τρόπο χρησιμοποιήθηκε το VADER [6]. Το VADER είναι ένα λεξικογραφικό εργαλείο ανάλυσης συναισθήματος, σχεδιασμένο ειδικά για τη συναισθηματική αποτίμηση κειμένων μικρής έκτασης, όπως αναρτήσεις σε κοινωνικά δίκτυα, τίτλοι ειδήσεων ή σύντομες δηλώσεις.

Πίνακας 3.3: Οι υπερπαράμετροι που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου LSTM για την πρόβλεψη των θεμάτων.

Υπερπαράμετρος	**Περιγραφή**
Loss function	Binary cross entropy (multilabel)
Optimizer	Adam [32]
Learning rate	Fixed: 0.001
Max epochs	30
Batch size	100
Weighted targets	✓

Βασίζεται σε ένα εκτενές συναισθηματικό λεξικό που περιλαμβάνει λέξεις και φράσεις με προαποφασισμένες βαθμολογίες συναισθηματικού φορτίου (θετικό, αρνητικό ή ουδέτερο). Το VADER λαμβάνει επίσης υπόψη του ενισχυτικές ή αντιστρεπτικές λέξεις (όπως “πολύ”, “όχι”, “λιγάκι”) καθώς και τη χρήση κεφαλαίων, σημείων στίξης και emoticons, ώστε να αποδώσει ένα συνολικό ”σκορ” συναισθήματος για κάθε πρόταση ή κείμενο. Το τελικό αποτέλεσμα είναι ένα σύνολο τεσσάρων μετρικών: θετικό, αρνητικό, ουδέτερο και ένα σύνθετο σκορ (compound score), το οποίο χρησιμοποιείται συχνά ως γενικός δείκτης συναισθηματικού προσανατολισμού. Η απλότητα και η ταχύτητά του καθιστούν το VADER ιδιαίτερα χρήσιμο για εφαρμογές όπου απαιτείται γρήγορη και ερμηνεύσιμη συναισθηματική ανάλυση. Η πρόβλεψη του συναισθήματος έγινε και στα 3 set train, validation, testing. Δεν χρειάστηκε να δοθεί κάτι συγκεκριμένο για υπερπαραμέτρους. Το VADER [6] μπορεί να ταξινομήσει τις προτάσεις σε τρεις κατηγορίες:

1. Αρνητική (Negative)
2. Ουδέτερη (Neutral)
3. Θετική (Positive)

Υπάρχουν και άλλα μοντέλα που μπορούν να ταξινομήσουν και σε περισσότερες πιο αναλυτικές κατηγορίες το συναίσθημα. Εμάς μας αρκούν αυτές γιατί είναι οι πιο συνηθισμένες και επίσης το RoBERTa που χρησιμοποιούμε στην συνέχεια ταξινομεί στις ίδιες κλάσεις και έτσι μπορούμε να έχουμε μια σύγκριση των δυο μοντέλων στα τελικά αποτελέσματα.

3.7 Ανίχνευση συναισθήματος RoBERTa

Η εφαρμογή του RoBERTa, ενός βελτιστοποιημένου γλωσσικού μοντέλου βασισμένου στην αρχιτεκτονική BERT [8], έχει φανεί ιδιαίτερα αποτελεσματική στην ανάλυση συναισθήματος, ειδικά σε περιπτώσεις όπου απαιτείται κατανόηση των φράσεων και των νοηματικών της γλώσσας. Το BERT είναι ένα μοντέλο φυσικής γλώσσας που βασίζεται στην αρχιτεκτονική των transformers και έχει σχεδιαστεί για να κατανοεί το πλαίσιο των λέξεων μέσα σε ένα κείμενο, τόσο από τα αριστερά όσο και από τα δεξιά. Αυτό του επιτρέπει να συλλαμβάνει πιο

βαθιά σημασιολογικά χαρακτηριστικά και να κατανοεί καλύτερα τη σημασία μιας πρότασης ή παραγράφου. Έχει χρησιμοποιηθεί ευρέως σε εφαρμογές όπως η αναζήτηση, η απάντηση σε ερωτήσεις, η κατηγοριοποίηση κειμένου και αναγνώριση οντοτήτων. Η δύναμή του έγκειται στην ικανότητά του να μαθαίνει τις σχέσεις μεταξύ των λέξεων σύμφωνα με το πλαίσιο τους, γεγονός που το καθιστά ιδιαίτερα αποδοτικό σε σύνθετα γλωσσικά καθήκοντα. Στο πλαίσιο της ανίχνευσης ψεύδους σε προτάσεις πολιτικού λόγου, το RoBERTa έχει χρησιμοποιηθεί με σκοπό την εύρεση του συναισθήματος, αξιοποιώντας την ικανότητα του να αναλύει τη συναισθηματική φόρτιση και τις γλωσσικές επιλογές των ομιλητών. Ιδιαίτερη έμφαση δίνεται στην αναγνώριση υπερβολικά θετικών ή αρνητικών εκφράσεων, χαρακτηριστικό που πολλές φορές συνδέεται με προσπάθειες πειθούς ή χειραγώγησης του αναγνώστη. Η προσέγγιση αυτή βασίζεται στην παραδοχή ότι το συναισθηματικό προφίλ ενός κειμένου μπορεί να λειτουργήσει ως έμμεσος δείκτης ειλικρίνειας ή ψεύδους, και η απόδοση του RoBERTa σε τέτοια καθήκοντα δείχνει πως τα σύγχρονα γλωσσικά μοντέλα μπορούν να συμβάλουν ουσιαστικά στην ανάλυση του πολιτικού λόγου και στην καταπολέμηση της παραπληροφόρησης.

Και το RoBERTa μπορεί να ταξινομήσει στις τρεις ίδιες κατηγορίες όπως και το VADER [6]. Στο επόμενο κεφάλαιο των αποτελεσμάτων, περιγράφουμε πιο αναλυτικά τις ταξινομήσεις του. Δεν χρειάστηκε να γίνει κάποια συγκεκριμένη αλλαγή στις υπερπαραμέτρους του. Η πρόβλεψη του συναισθήματος έγινε και στα 3 set train, validation, testing. Όπως και με την περίπτωση του VADER [6] δεν έχομε τα πραγματικά συναισθήματα για την κάθε πρόταση του dataset και για αυτόν τον λόγο δεν μπορούμε να παράξουμε κάποια μετρική για αυτές τις προβλέψεις.

3.8 Ανίχνευση ψεύδους τελικό μοντέλο

Το τελικό μοντέλο που κατασκευάστηκε είναι απόρροια πολλών δοκιμών που έγιναν με διαφορετικούς τρόπους pre-processing, διαφορετικά δεδομένα από το dataset. Όπως την επιπρόσθετη πληροφορία από την εύρεση των θεμάτων και το συναίσθημα analysis καθώς και των τελικών μετρικών αποτελεσμάτων πάνω στο test set. Έχουν δοκιμαστεί πολλές μεθοδολογίες όπως ο συνδυασμός LSTM με custom embeddings, LSTM συνδυαστικά με BERT, CNN συνδυαστικά με BERT, CNN συνδυαστικά με custom embeddings. Στην εργασία δεν περιγράφουμε όλα τα αποτελέσματα από όλους αυτούς τους συνδυασμούς αλλά μόνο τον καλύτερο συνδυασμό που ήταν το CNN με custom embeddings.

Αρχικά είχαν γίνει δοκιμές μόνο με τις προτάσεις του dataset, που είναι η στήλη statement. Αλλά λόγο της φύσης του προβλήματος είναι αδύνατον το μοντέλο να μάθει να προβλέπει το ψεύδος μόνο και μόνο από τις προτάσεις. Χρειάζεται να βασιστεί σε πολύ μεγάλο βαθμό σε άλλες μεταπληροφορίες που όσες περισσότερες είναι τόσο το καλύτερο. Για αυτόν τον λόγο χρειάστηκε να γίνει pre-processing και στις υπόλοιπες στήλες του dataset όπως περιγράφεται στον Πίνακα 3.4. Για τις στήλες που έχουν σαν τύπο κείμενο (text) έχει γίνει αντικατάσταση των NA τιμών με το text ”NA” και σε αυτές που είναι αριθμός έχει γίνει αντικατάσταση των NA με μηδέν. Οι text στήλες στην συνέχεια έγιναν encode με index based τρόπο όπως έγιναν και οι προτάσεις κατά το pre-processing. Οι text στήλες περιέχουν ποιοτικές τιμές, ένας τρόπος για να γίνουν encode είναι με το one-hot encoding. Αλλά με αυτόν τον τρόπο πολύ

Πίνακας 3.4: Αναλυτικά η αντιστοιχία ανάμεσα στα γενικά θέματα και στα θέματα του dataset. Αυτό έγινε για απλούστευση των θεμάτων με τελικό σκοπό την διευκόλυνση των μοντέλων στην ταξινόμηση των θεμάτων.

Στήλη του dataset	**Τύπος**	**Αντικατάσταση NA**	**Fill zero**	**Index based encoding**	**Standard scaling**
Speaker	Text	✓	—	✓	—
Job title	Text	✓	—	✓	—
State	Text	✓	—	✓	—
Affiliation	Text	✓	—	✓	—
General Context	Text	✓	—	✓	—
Barely true counts	Numeric	—	✓	—	✓
False counts	Numeric	—	✓	—	✓
Half true counts	Numeric	—	✓	—	✓
Mostly true counts	Numeric	—	✓	—	✓
Pants on fire counts	Numeric	—	✓	—	✓

γρήγορα θα γεμίζαμε με στήλες από 0 και 1 γιατί έχουν πολλές διαφορετικές τιμές. Αυτό που επιλέχθηκε είναι να μείνουν ως έχει και να περάσουν από ένα layer embedding για να μάθει το μοντέλο το ειδικό νόημα της κάθε λέξης. Για τις αριθμητικές στήλες οι τιμές τους κανονικοποιήθηκαν με την τεχνική του standard scaling. Ο Τύπος 3.1 περιγράφει πως γίνεται αυτός ο υπολογισμός για μια τιμή x μιας στήλης. Με την κανονικοποίηση αυτήν βοηθάμε το μοντέλο να εκπαιδευτεί πιο εύκολα πάνω σε αυτές τις τιμές. Αυτό το pre-processing έγινε ώστε να μπορέσουν να χρησιμοποιήθουν όλα τα μεταδεδομένα που παρέχει το dataset για να μπορέσουμε να αυξήσουμε την επιτυχία του μοντέλου. Οι τιμές των θεμάτων και του συναισθήματος είναι κανονικοποιημένες με one hot encoding, γιατί τα συναισθήματα είναι μόνο τρια άρα θα έχουμε τρεις στήλες παραπάνω και τα θέματα είναι 13 άρα θα έχουμε 13 στήλες παραπάνω. Στο σύνολο 16 στήλες για συναισθήματα μαζί με θέματα, που σαν αριθμός δεν είναι και τόσο μεγάλος για επιπρόσθετες στήλες στο dataset. Όλες μαζί οι στήλες του dataset και οι επιπρόσθετες είναι 26.

$$z = \frac{x - \mu}{s} \tag{3.1}$$

- z: Η τελική τιμή του standard scale.
- μ: Η μέση τιμή όλης της στήλης.
- s: Το standard deviation της στήλης.
- x: Η τιμή της στήλης που θέλουμε να κανονικοποιηθεί.

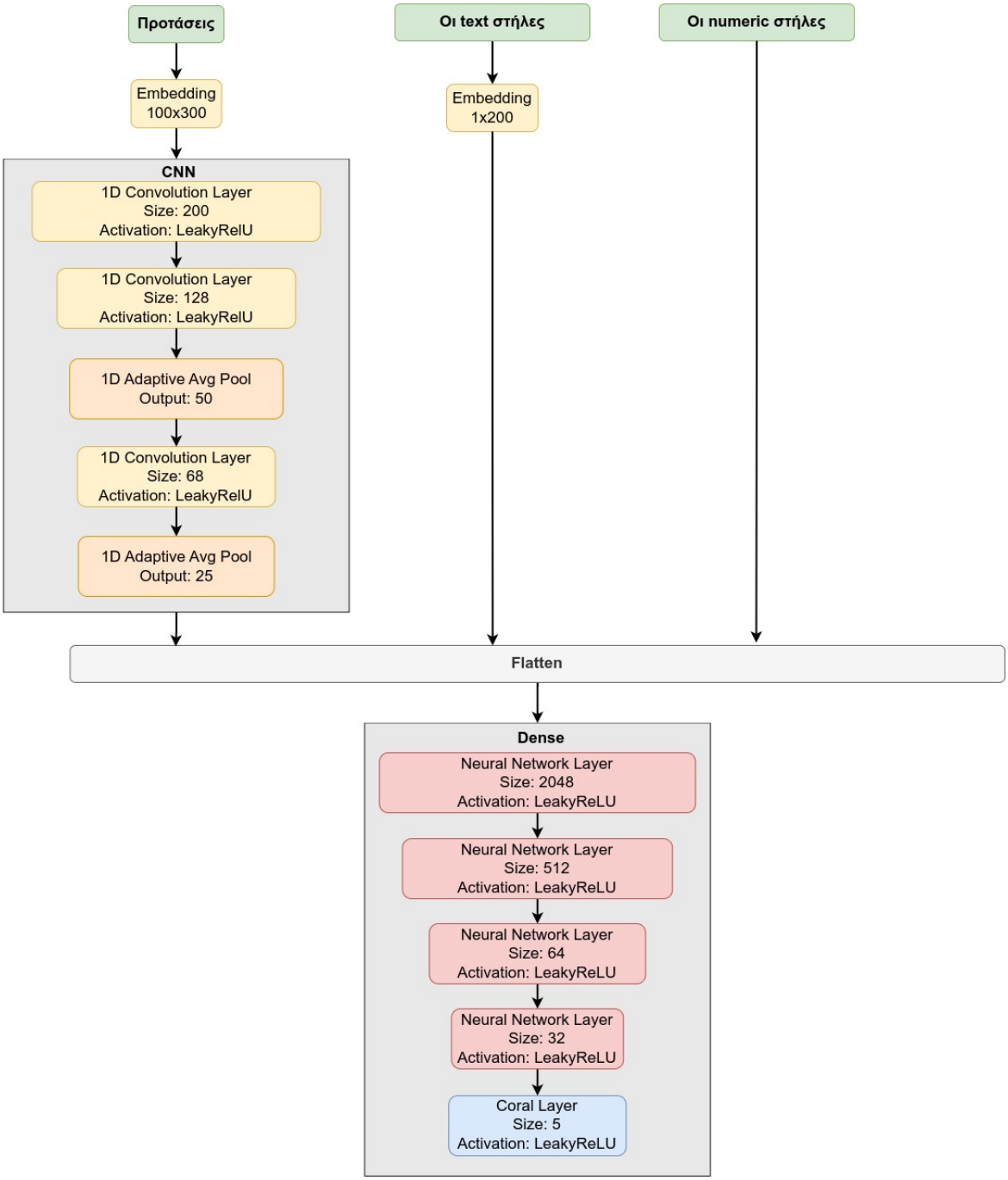
Κατά τους πειραματισμούς με το μοντέλο δοκιμάστηκαν διάφοροι αλγόριθμοι

βελτιστοποίησης (optimization). Τα καλύτερα αποτελέσματα τα είχαμε με τον Adam [32] που αυτός χρησιμοποιείται για τα τελικά αποτελέσματα. Το Σχήμα 3.6 απεικονίζει το σχήμα του τελικού μοντέλου, όπου παρατηρούμε ότι οι προτάσεις περνάνε από embeddings και μετά από ένα CNN μοντέλο, τα text input περνάνε και αυτά από embeddings. Οι έξοδοι (output) από το CNN και τα embeddings μαζί με τα input των κανονικοποιημένων αριθμητικών στηλών γίνονται ένα μονοδιάστατο διάνυσμα που είναι η είσοδος στο τελικό dense δίκτυο. Επίσης δοκιμάστηκε και το BERT αντί των embeddings για τις προτάσεις αλλά χωρίς κάποιο καλό αποτέλεσμα. Με το BERT το μοντέλο είχε πολύ πιο αργή εκπαίδευση και πάθαινε overfit σε λίγες εποχές, ακόμα και με dropout και weight decay. Οι υπερπαράμετροι του μοντέλου φαίνονται καλύτερα στον Πίνακα 3.5. Εδώ έχουμε ένα δυναμικό learning rate (reduce on plateau) και συγκεκριμένα αυτό που μειώνεται όταν η τιμή του validation error μένει σταθερή για πάνω από 2 εποχές. Αυτό βοηθάει όσο περνάνε οι εποχές και το validation error δεν αλλάζει και πολύ να μειώνεται το learning rate ώστε να μπορέσει το μοντέλο να κάνει μικροδιορθώσεις στα βάρη του και να πάει σε ένα καλύτερο αποτέλεσμα.

Πολλές έρευνες έχουν διαφοροποιήσει τους στόχους του LAIR και προσπαθούν να κάνουν το πρόβλημα binary classification, άλλοι εστιάζουν μόνο σε μια κλάση και άλλοι προσπαθούν να προβλέψουν και τις 6 κλάσεις μαζί (multilabel classification). Εμείς χρησιμοποιήσουμε το LAIR όπως πραγματικά είναι σχεδιασμένο, δηλαδή το μοντέλο να μπορεί να προβλέψει μια από τις 6 κλάσεις multilabel classification. Αυτό που παρατηρήσαμε είναι ότι οι κλάσεις δεν είναι ποιοτικές τιμές αλλά κλιμακωτές κατηγορίες, όπως για παράδειγμα ένα star rating σύστημα, ξεκινώντας από την κατηγορία pants on fire μέχρι την κατηγορία true. Έτσι αποφασίσαμε το μοντέλο να μην εκπαιδευτεί με τον κλασικό τρόπο του binary cross entropy αλλά με έναν τρόπο που να λαμβάνει υπόψη του την απόσταση που έχει η πρόβλεψη από τον στόχο. Για παράδειγμα άμα ο στόχος είναι η κλάση true αλλά το μοντέλο προβλέψει mostly-true τότε με το binary cross entropy το σφάλμα θα είναι μεγάλο γιατί δεν πέτυχε την κατηγορία. Αμα παρατηρήσουμε πιο προσεκτικά αντιλαμβανόμαστε ότι η πρόβλεψη δεν ήταν και τόσο άστοχη γιατί απέχει μόνο μια κατηγορία μακριά από τον στόχο. Άρα ένας τρόπος για να κάνουμε το μοντέλο να λαμβάνει υπόψη του την απόσταση της πρόβλεψης από τον στόχο είναι να διαχειριστούμε τους στόχους σαν regression πρόβλημα. Στην αναζήτηση άμα υπάρχει κάτι πιο εξειδικευμένο που να μπορεί να μας βοηθήσει σε αυτό, βρήκαμε το CORAL [33] που είναι σχεδιασμένο για ordinal classification. Το CORAL έχει επίσης και ένα output layer ώστε να

Πίνακας 3.5: Οι υπερπαράμετροι που χρησιμοποιήθηκαν για την εκπαίδευση του τελικού μοντέλου, για την πρόβλεψη των ψευδών ειδήσεων. Παρατηρούμε ότι ο Optimizer είναι το CORAL [33] που ειδικεύεται στην εκμάθηση στόχων που είναι ordered regression.

Υπερπαράμετρος	**Περιγραφή**
Loss function	Binary cross entropy (multilabel)
Optimizer	CORAL
Learning rate	Reduce learning rate on plateau
Max epochs	50
Batch size	25
Weighted targets	✓



Σχήμα 3.6: Στο σχήμα απεικονίζεται το τελικό μοντέλο με χρήση embeddings και CNN. Παρατηρούμε ότι οι προτάσεις περνάνε από ενα CNN δίκτυο ώστε να αντλήσει περισσότερες πληροφορίες το μοντέλο από αυτές. Οι υπόλοιπες είσοδοι μαζί με την έξοδο από το CNN μαζεύονται σε ένα μονοδιάστατο διάνυσμα που είναι και η είσοδος στο dense δίκτυο. Στην έξοδο έχει ένα CORAL layer.

παρέχει τις προβλέψεις με τον αναμενόμενο τρόπο που χρειάζεται η συνάρτηση σφάλματος του CORAL. Στον Πίνακα 3.6 γίνεται σύγκριση των προβλέψεων του CORAL σε σχέση με τις

Πίνακας 3.6: Συγκρίνουμε τις διαφορές ανάμεσα στο κλασικό πολλαπλών κλάσεων (multiclass) output και στο CORAL καθώς και πως αυτά τα output θα έπρεπε να ήταν για να πετύχουν τον στόχο.

	Προβλέψεις	
Target εύρος 0 μέχρι 5	**Κλασικό multiclass**	**CORAL**
2	[0, 0, 1, 0, 0, 0]	[0, 1, 0, 1, 0]
5	[0, 0, 0, 0, 0, 1]	[1, 1, 1, 1, 1]
0	[1, 0, 0, 0, 0, 0]	[0, 0, 0, 0, 0]
3	[0, 0, 0, 1, 0, 0]	[1, 0, 0, 1, 1]

προβλέψεις ενός κλασικού multiclass classification. Το CORAL αυτό που διαφοροποιεί είναι το πως αντιλαμβανόμαστε την κλάση που έχει προβλέψει το μοντέλο, έπειτα χρησιμοποιεί το binary cross entropy για την εφαρμογή του loss. Το CORAL έχει για output $N - 1$ στόχους. Για να προβλέψει τον στόχο 3 αρκεί στις εξόδους να παραχθούν 3 output που να έχουν τιμή μεγαλύτερη ή ίση με 0.5, άσχετα με την θέση που έχουν στην έξοδο τους. Αυτό μπορούμε να το εντοπίσουμε καλύτερα στον Πίνακα 3.6. Αυτό βοηθάει το μοντέλο να κατανοήσει από μόνο του και να ορίσει ποια έξοδος θα αντιστοιχεί σε ποια κατηγορία, αρκεί το πλήθος των εξόδων που έχουν τιμές πάνω από 0.5 να ισούται με τον στόχο.

3.9 Επίλογος

Σε αυτό το κεφάλαιο περιγράψαμε αναλυτικά το LAIR dataset και το pre-processing που έχει γίνει για την καλύτερη διαχείριση των δεδομένων του. Περιγράψαμε αναλυτικά τις μεθοδολογίες που έχουν ακολουθηθεί στην ανίχνευση των θεμάτων με την χρήση του LDA αλλά και του custom LSTM μοντέλου. Στην συνέχεια αναλύσαμε τις μεθοδολογίες που μας βοήθησαν να εξάγουμε το συναίσθημα από τις προτάσεις του dataset. Αυτό έγινε με την χρήση του VADER που είναι μια μέθοδος λεξικού αλλά και με το RoBERTa που είναι ένα προηγμένο μοντέλο βαθιάς μάθησης και transformers. Υπενθυμίζεται ότι, όλα τα επιπρόσθετα δεδομένα αναμένουμε να βοηθήσουν στην καλύτερη πρόβλεψη του μοντέλου μας σε σχέση με την περίπτωση χρήσης μόνο των δεδομένων που παρέχει το LAIR. Τέλος, αναφερθήκαμε στον συνδυασμό όλων των παραπάνω μεθοδολογίων και πως αυτά μπορούν να βοηθήσουν το τελικό custom μοντέλο για την πρόβλεψη των 6 κλάσεων ψεύδους. Στο επόμενο κεφάλαιο μελετάμε τα τελικά αποτελέσματα από όλες τις δοκιμές που έγιναν πάνω στο τελικό μοντέλο αλλά και των επιμέρους μεθοδολογιών για την ανίχνευση των θεμάτων και του συναισθήματος.

Κεφάλαιο 4

Αποτελέσματα

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο εξετάζουμε τα αποτελέσματα από όλες τις μεθοδολογίες που χρησιμοποιήθηκαν σε αυτήν την εργασία. Αρχικά, από τις μεθοδολογίες για την εύρεση των θεμάτων και στην συνέχεια με τις μεθοδολογίες για την ανίχνευση του συναισθήματος. Τέλος, αναλύουμε τα αποτελέσματα του τελικού μοντέλου με όλους του πιθανούς συνδυασμούς δεδομένων. Χωρίς την χρήση των επιπλέων πληροφοριών από τις μεθοδολογίες, με την πληροφορία των θεμάτων, με την πληροφορία των συναισθημάτων, συνδυασμός πληροφοριών θεμάτων και συναισθημάτων. Έτσι μπορούμε να εξάγουμε καλύτερα συμπεράσματα για το πόσο βοήθησε η κάθε επιπλέον πληροφορία.

4.2 Αποτελέσματα εξαγωγής θεμάτων

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, πραγματοποιήσαμε στατιστική ανάλυση με σκοπό να διερευνήσουμε την ύπαρξη στατιστικής συσχέτισης μεταξύ των θεμάτων και των κατηγοριών ψεύδους. Από την ανάλυση αυτή προέκυψε ότι υπάρχει ισχυρή στατιστική συσχέτιση, όπως αποδεικνύεται και από την τιμή του x^2 (1386,28). Το heatmap που παρουσιάζεται στο Σχήμα 3.3 δείχνει ότι η πιο συχνή συσχέτιση παρατηρείται μεταξύ της κατηγορίας ”half-true” και του θέματος ”Economy”. Γενικότερα, οι περισσότερες εγγραφές κατατάσσονται στην κατηγορία ”half-true”, γεγονός που υποδηλώνει ότι οι περισσότερες προτάσεις περιέχουν ταυτόχρονα αλήθεια και ψεύδος, σε περίπου ίσα μέρη. Η παρατήρηση αυτή υπογραμμίζει ότι οι περισσότεροι τίτλοι άρθρων συνδυάζουν στοιχεία αλήθειας και ψεύδους, δημιουργώντας επιπλέον πρόκληση για τα μοντέλα μας, καθώς δυσχεραίνεται η ακριβής ταξινόμησή τους ως αληθών ή ψευδών. Συγκεκριμένα, τα μοντέλα ενδέχεται να παρουσιάσουν μεροληψία υπέρ της ενδιάμεσης αυτής κατηγορίας, με συνέπεια να μειώνεται η ακρίβεια στην αναγνώριση των ακραίων κατηγοριών. Από τα θέματα, το ”Economy” εμφανίζει τον μεγαλύτερο αριθμό εγγραφών με σημαντική διαφορά σε σχέση με τα υπόλοιπα. Ακολουθεί το ”Politics”, όπου οι εγγραφές κατανεμήθηκαν σχετικά ισόποσα μεταξύ των κατηγοριών, και τρίτο είναι το ”Health”, το οποίο σχετίζεται περισσότερο με τις τέσσερις κεντρικές κατηγορίες και λιγότερο με τις ακραίες (”pants on fire”, ”true”).

Πίνακας 4.1: Οι παράμετροι που χρησιμοποιήθηκαν για την πρόβλεψη των θεμάτων για το LDA, μαζί με τα διαστήματα τιμών που ερευνήθηκαν αλλά και τις τελικές τιμές που επιλεχθηκαν κατά το CS score.

Παράμετρος	**Διαστήματα τιμών**	**Τελική Τιμή**
Πλήθος θεμάτων	[2, 15] με βήμα 1	10
Άλφα	[0.01, 1.1] με βήμα 0.01 και [0.001, 0.1] με βήμα 0.005	0.001
Ήτα	[0.01, 1.1] με βήμα 0.01 και [0.001, 0.1] με βήμα 0.005	0.021

Συνολικά, παρατηρείται ότι οι συσχετίσεις είναι πιο έντονες στις κεντρικές κατηγορίες παρά στις ακραίες, γεγονός που, σε συνδυασμό με την ακαταστασία των συσχετίσεων, αποτελεί ένα επιπλέον εμπόδιο που τα μοντέλα θα πρέπει να ξεπεράσουν προκειμένου να επιτύχουν ικανοποιητική γενίκευση και να αποφύγουν τη μεροληψία. Παρόλα αυτά περιμένουμε να έχουμε κάποια βελτίωση στα αποτελέσματα μας όταν χρησιμοποιηθεί η πληροφορία των θεμάτων στο τελικό μοντέλο.

4.2.1 Αποτελέσματα εξαγωγής θεμάτων με LDA

Για το LDA έγιναν πολλές δοκιμές με διαφορετικές τιμές για το πλήθος των θεμάτων και τις τιμές Άλφα και Ήτα. Η αξιολόγηση για τις καλύτερες τιμές έγινε με την μέθοδο του cross validation σε 10 folds και το score με το CV coherence. Έγιναν πολλές δοκιμές αλλάζοντας κάθε φορά μόνο μία παράμετρο και κρατώντας όλες τις άλλες σταθερές στις αρχικές τους τιμές. Οι αρχικές τιμές και οι τελικές επιλέχθηκαν εμπειρικά. Στον Πίνακα 4.1 υπάρχουν, τα διαστήματα τιμών για κάθε παράμετρο μαζί με την τελική τιμή που επιλέχθηκε ανάλογα με το καλύτερο CV score. Η σειρά με την οποία έγιναν οι δοκιμές δεν ήταν τυχαία, πρώτα ξεκινήσαμε με το πλήθος των topics, γιατί σαν παράμετρος επηρεάζει το Άλφα. Αυτό ισχύει γιατί το Άλφα ελέγχει κατά πόσο τα θέματα που θα παραχθούν θα είναι γενικά και θα ταιριάζουν σε πολλές προτάσεις ή θα είναι ειδικά και θα ταιριάζουν σε λίγες. Συνεπως, αν επιλέγαμε πρώτα το Άλφα και στην συνέχεια καταλήγαμε σε άλλο πλήθος θεμάτων, τότε το Άλφα που επιλέξαμε μπορεί να μην παρήγαγε το καλύτερο score και θα έπρεπε να ερευνηθεί ξανά. Αφού βρέθηκε το καλύτερο score έχοντας υπόψη το πλήθος θεμάτων και το Άλφα ξεκίνησαν οι δοκιμές για την καλύτερη τιμή του Ήτα. Οι δοκιμές για το Άλφα και το Ήταν έγιναν σε δύο κύκλους. Ο πρώτος κύκλος με μεγαλύτερη αρχική και τελική τιμή αλλά και βήμα, για να προσεγγισθεί πιο γρήγορα η βέλτιστη τιμή. Ο δεύτερος κύκλος έγινε με μικρότερη αρχική και τελική αλλά και βήμα, ώστε να βρεθεί η βέλτιστη τιμή με μεγαλύτερη ακρίβεια. Κατά την εκτέλεση των δοκιμών, γινόταν έλεγχος του CV score με γραφική παράσταση, ώστε να τερματιστεί ο αλγόριθμος με χειροκίνητο τρόπο, όταν το score ξεκινούσε και μειωνόταν, με αυτόν τον τρόπο έγινε εξοικονόμηση χρόνου. Οι μικρές τιμές στο Άλφα και στο Ήτα δείχνουν ότι τα καλύτερα αποτελέσματα τα πετύχαμε όταν οι λέξεις αλλά και τα θέματα ήταν πιο συγκεκριμένα και όχι γενικά.

Η τελική τιμή του CV score ήταν 0.3620. Το score αυτό δείχνει ότι το μοντέλο καταφέρνει να ταυτίσει προτάσεις με θέματα αλλά δεν το πετυχαίνει και με τον καλύτερο τρόπο. Τα θέματα που κατασκευάστηκαν από το LDA φαίνονται στον Πίνακα 4.2, καταγράφουμε τις

λέξεις που αποτελούν κάθε θέμα αλλά και τα βάρη που έχουν ανατεθεί σε κάθε μια λέξη. Μπορούμε να παρατηρήσουμε ότι κάποιες από τις λέξεις υπάρχουν σε παραπάνω από ένα θέμα αλλά με διαφορετική βαρύτητα. Επίσης το LDA δεν μπορεί να παράξει λέξεις που θα

Πίνακας 4.2: Τα 10 θέματα που έχουν παραχθεί από το LDA. Στην δεύτερη στήλη καταγράφουμε τις λέξεις που αποτελούν κάθε θέμα και το βάρος για αυτήν. Η περιγραφή είναι δοσμένη από εμάς για να είναι πιο εύκολο να κατανοήσουμε το γενικό νόημα των λέξεων για κάθε θέμα.

#	**Λέξεις**	**Περιγραφή**
1	$0.032 \times say + 0.022 \times illeg + 0.022 \times immigr + 0.019 \times trump + 0.015 \times donald + 0.011 \times money + 0.010 \times border + 0.010 \times would + 0.007 \times state + 0.007 \times fund$	Immigration
2	$0.042 \times year + 0.024 \times budget + 0.022 \times say + 0.018 \times school + 0.017 \times billion + 0.016 \times debt + 0.016 \times citi + 0.015 \times state + 0.011 \times last + 0.011 \times spend$	Public budget
3	$0.044 \times say + 0.025 \times romney + 0.018 \times mitt + 0.014 \times court + 0.009 \times one + 0.008 \times suprem + 0.007 \times marriag + 0.007 \times gay + 0.006 \times said + 0.006 \times state$	Law
4	$0.019 \times oil + 0.016 \times year + 0.011 \times campaign + 0.010 \times ga + 0.010 \times day + 0.010 \times one + 0.009 \times said + 0.008 \times republican + 0.007 \times state + 0.007 \times never$	Gas
5	$0.024 \times peopl + 0.021 \times american + 0.018 \times health + 0.015 \times insur + 0.015 \times million + 0.012 \times go + 0.011 \times govern + 0.011 \times get + 0.011 \times say + 0.011 \times year$	Health
6	$0.035 \times state + 0.027 \times clinton + 0.023 \times say + 0.021 \times unit + 0.021 \times hillari + 0.013 \times peopl + 0.010 \times gun + 0.008 \times countri + 0.008 \times percent + 0.007 \times check$	Clinton
7	$0.051 \times percent + 0.028 \times state + 0.026 \times rate + 0.014 \times year + 0.013 \times say + 0.013 \times unemploy + 0.013 \times averag + 0.012 \times countri + 0.011 \times nation + 0.011 \times scott$	Jobs
8	$0.029 \times say + 0.029 \times health + 0.029 \times care + 0.024 \times vote + 0.016 \times law + 0.015 \times republican + 0.015 \times bill + 0.014 \times democrat + 0.010 \times plan + 0.010 \times new$	Politics
9	$0.076 \times tax + 0.051 \times job + 0.017 \times creat + 0.016 \times rais + 0.015 \times sinc + 0.015 \times say + 0.014 \times year + 0.014 \times percent + 0.013 \times state + 0.013 \times cut$	Tax
10	$0.074 \times obama + 0.054 \times presid + 0.049 \times say + 0.032 \times barack + 0.014 \times would + 0.013 \times senat + 0.012 \times support + 0.012 \times said + 0.011 \times bush + 0.008 \times plan$	Barak Obama

περιέγραφαν το κάθε θέμα ώστε να μας δώσει ένα γενικό νόημα, για αυτό προσπαθήσαμε να αποδώσουμε εμείς το γενικό νόημα των λέξεων για κάθε θέμα, αυτό καταγράφεται στην τρίτη στήλη ”Περιγραφή” του Πίνακα 4.2.

Παρατηρούμε στον Πίνακα 4.2 ότι τα βάρη έχουν μικρές τιμές. Αυτό γίνεται για δύο λόγους. Ένας είναι ότι δεν μπορεί να ξεχωρίσει εύκολα τις λέξεις που να ξεχωρίζουν αμέσως την θεματική της πρότασης, εξού και το CV score 0.3620. Και ο άλλος λόγος είναι, ότι έχουμε μικρές τιμές στο Ήτα που κάνει τις λέξεις να είναι πιο συγκεκριμένες για κάθε θέμα.

Το θετικό με το LDA είναι ότι μπορεί να παρέχει προβλέψεις για όλε τις εγγραφές του dataset σε αντίθεση με ένα μοντέλο μηχανικής μάθησης που σε κάποιες περιπτώσεις μπορεί και να μην καταφέρει να παρέχει κάποια πρόβλεψη.

4.2.2 Αποτελέσματα εξαγωγής θεμάτων με LSTM

Θέλοντας να δούμε και την απόδοση της πρόβλεψης θεμάτων από ένα μοντέλο μηχανικής μάθησης, εκπαιδεύσαμε και ένα μοντέλο LSTM ώστε να μάθει σύμφωνα με την θεματική του κάθε τίτλου. Οι στόχοι του μοντέλου είναι η στήλη “Subjects” του LAIR που οι τιμές της έχουν ομαδοποιηθεί κατά τον Πίνακα 3.1 που έχουμε δει. Αυτό έγινε γιατί οι λέξεις που είχε αυτή η στήλη ήταν πολλές, παραπάνω από 100, και έτσι θα δυσκόλευε το μοντέλο να προσπαθεί να προβλέψει ανάμεσα σε 100 διαφορετικές target labels. Λόγο ότι ένας τίτλος μπορεί να περιγράφεται με παραπάνω από ένα θέμα μας οδηγεί στο να έχουμε multilabel classification. Αυτή η περίπτωση classification είναι όταν έχουμε πολλούς πιθανούς στόχους αλλά δεν είναι απαραίτητο κάθε εγγραφή του dataset να αντιστοιχεί σε μόνο έναν από αυτούς. Αυτό από μόνο του αυξάνει τον βαθμό δυσκολίας του μοντέλου καθώς και δυσκολεύει την αξιολόγηση του με τους κλασικούς τύπους μετρικών. Όπως είχαμε πει ξανά σε προηγούμενο κεφάλαιο στην έρευνα αυτή έχει επιλεχθεί να παραχθούν θέματα από την αρχή για το LAIR. Αυτό έγινε με σκοπό την πιο ρεαλιστική προσέγγιση στο πρόβλημα, γιατί σε κάποιο άλλο dataset μπορεί να μην είχαμε τα θέματα και θα έπρεπε να τα παράξουμε. Επίσης ένας ακόμα λόγος είναι για να μπορέσουμε να συγκρίνουμε τις δυο τεχνικές ανίχνευσης των θεμάτων.

Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν τα dataset του train, validation και test όπως τα παρέχει το LAIR. Το μόνο που άλλαξε από τα δεδομένα είναι η στήλη που αναφερθήκαμε πιο πάνω. Δεν χρησιμοποιήθηκαν άλλα δεδομένα πέρα του τίτλου. Λόγο ότι οι συχνότητες των κλάσεων δεν είναι ίδιες αλλά και το πλήθος των θεμάτων σε κάθε τίτλο δεν ξεπερνάει τις τρεις, παρέχουμε στην συνάρτηση υπολογισμού του σφάλματος κάποια βάρη για

Πίνακας 4.3: Οι παράμετροι που χρησιμοποιήθηκαν για την πρόβλεψη των θεμάτων με το LDA.

Παράμετρος	**Τιμή**
Σύνολο στόχων	13 (multilabel)
Μέθοδος σφάλματος	Binary Cross Entropy
Αλγόριθμος Βελτιστοποίησης	Adam

να εξισορροπήσει την ανισότητα στο πλήθος τον εγγραφών για κάθε κλάση με την Συνάρτηση 4.1, όπου στην συνέχει αυτές οι τιμές κανονικοποιούνται από αυτήν την Συνάρτηση 4.2.

$$W_c = \frac{N}{C \cdot n_c} \tag{4.1}$$

- W_c: Το βάρος για μια συγκεκριμένη κλάση c.
- C: Το πλήθος των κλάσεων.
- n_c: Πόσες εγγραφές περιέχουν την κλάση c.
- N: Το πλήθος των εγγραφών.

$$NW_t = \frac{W_t}{S} \tag{4.2}$$

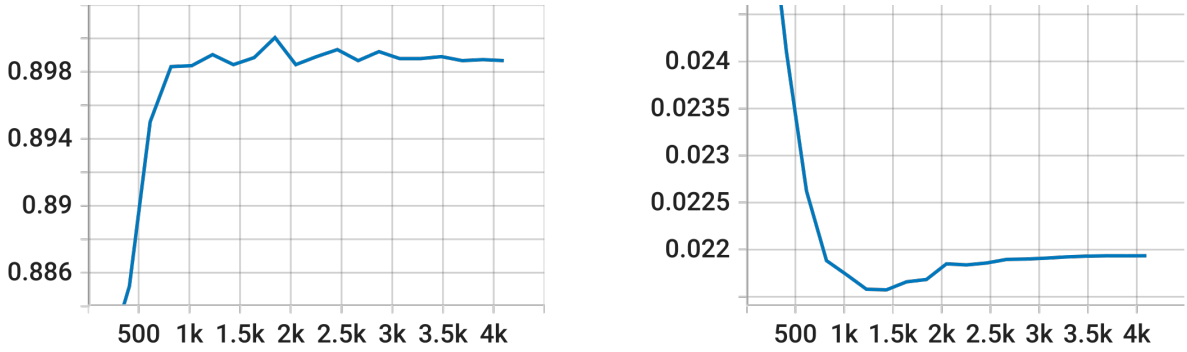
- NW_t: Το κανονικοποιημένο βάρος για έναν στόχο t.
- W_t: Η τιμή W_c για έναν στόχο t.
- s: Το άθροισμα όλων W_c.

Επίσης το θεσιακό βάρος υπολογίστηκε για κάθε κλάση. Το θεσιακό βάρος βοηθάει την συνάρτηση υπολογισμού του σφάλματος να δίνει μεγαλύτερη ποινή όταν το μοντέλο προβλέπει λάθος κλάση. Στην περίπτωση μας ο αριθμός των θεμάτων για μια εγγραφή είναι πολύ μικρός σε σχέση με τον πλήθος όλων τον θεμάτων. Για παράδειγμα για έναν τίτλο μπορεί να υπάρχουν δύο σχετικά θέματα, άρα υπάρχουν άλλα 11 άσχετα θέματα. Άμα το μοντέλο μαντέψει αρκετά από τα άσχετα θέματα, που είναι στατιστικά πιο πιθανό, τότε δεν θα έχει μεγάλη ποινή για τα λάθη που έκανε. Αυτό οδηγεί το μοντέλο εστιάζει στην πρόβλεψη των άσχετων θεμάτων και όχι των σχετικών. Αυτό το πρόβλημα έρχεται να λύσει το θεσιακό βάρος που περιγράφεται από την Συνάρτηση 4.3. Για την εύρεση του καλύτερου μοντέλο έγιναν δοκιμές με και χωρίς το θεσιακού βάρους καθώς και με και χωρίς βάρη για την δυσαναλογία των κλάσεων ώστε να παρατηρήσουμε την επίδραση τους στις τελικές μετρικές.

$$PW_c = \frac{n_{\text{neg},c}}{n_{\text{pos},c}} \tag{4.3}$$

- PW_c: Το θεσιακό βάρος για την κάση c.
- $n_{\text{neg},c}$: Το πλήθος των αρνητικών στόχων για την κλάση c.
- $n_{\text{pos},c}$: Το πλήθος των θετικών στόχων για την κλάση c.

Στον Πίνακα 4.4 φαίνονται τα αποτελέσματα από αυτές τις εκτελέσεις, παρατηρούμε ότι όταν έχουμε ορίσει θεσιακό βάρος τότε το recall είναι μεγαλύτερο από το αντίστοιχο accuracy (γραμμές 1 και 3). Το ανάποδο γίνεται όταν δεν έχουμε ορίσει θεσιακό βάρος αλλά έχουμε



(α) Validation accuracy (β) Validation error

Σχήμα 4.1: Απεικονίζονται δυο διαγράμματα από την εκπαίδευση του LSTM με την εκτέλεση του 2 από τον Πίνακα 4.4. Είναι οι μετρικές του validation accuracy (α) και error (β) σε συνάρτηση με τον χρόνο.

βάρος κλάσεων (γραμμές 2 και 4). Επίσης όταν δεν έχουμε ορίσει θεσιακό βάρος παρατηρούμε ότι το μοντέλο δυσκολεύετε να δώσει προβλέψεις με αποτέλεσμα να έχουμε πολλές εγγραφές χωρίς κάποια πρόβλεψη, που φαίνονται με ποσοστά επί του συνόλου του κάθε set. Σε αυτήν την περίπτωση το μοντέλο είναι σαν να προσπαθεί να δώσει μια πρόβλεψη όταν είναι πιο σίγουρο για το αποτέλεσμα και το αντίθετο γίνεται όταν έχουμε θεσιακό βάρος. Την τελική επίδραση στις προβλέψεις του ψεύδους από αυτές τις δοκιμές τις περιγράφουμε στα αποτελέσματα του τελικού μοντέλου. Σε αυτού του είδους classification και με 13 πιθανές κλάσεις δεν μπορούμε να δώσουμε μεγάλη βαρύτητα στην μετρική του accuracy, γιατί μας παραπλανά η υψηλή τιμή του, που στην πραγματικότητα οφείλεται στο ότι έχουμε πολλές κλάσεις με μικρό αριθμό θετικών κλάσεων ανά εγγραφή και έτσι το μοντέλο ”κερδίζει” accuracy από τις προβλέψεις των true negative. Ο υπολογισμός των precision και recall έγινε χρησιμοποιώντας τις προβλέψεις από όλες τις κλάσεις μαζί και όχι για κάθε κλάση ξεχωριστά και στην συνέχεια να υπολογιστεί ο μέσος όρος της μετρικής. Αυτό βοηθάει στο πρόβλημα με το μικρό πλήθος θετικών κλάσεων ανά εγγραφή, με το να μας δίνει μια πιο σωστή εικόνα για την απόδοση του μοντέλου. Θέλοντας να δούμε τις δυνατότητες του μοντέλου αναζητήθηκαν νέοι τίτλοι από άρθρα της Αμερικής που δεν υπάρχουν στο LAIR και είναι πιο πρόσφατα. Έχοντας 13 πιθανά θέματα επιλέχθηκαν αντίστοιχα και τόσοι νέοι τίτλοι. Προσπαθήσαμε οι τίτλοι να καλύψουν όσο το δυνατός περισσότερο τα θέματα που έχουμε σαν targets στο μοντέλο. Αυτοί οι νέοι τίτλοι περιέχουν και πολλές λέξεις που δεν έχει δει το μοντέλο στο LAIR, αυτό είναι λογικό γιατί το LAIR περιέχει 10000+ που δεν είναι αρκετοί για να καλύψεις μεγάλο εύρος από λέξεις. Παρόλα αυτά το μοντέλο κατάφερε να αποδώσει τα θέματα που περιγράφονται απόπ τον Πίνακα 4.5. Στον πίνακα αυτόν έχουμε τα θέματα που πρόβλεψαν τα μοντέλα με αριθμούς 2 και 3, όπως περιγράφονται στον Πίνακα 4.4. Και εδώ μπορούμε να παρατηρήσουμε ότι το μοντέλο 3 που είχε εκπαιδευτεί με τα θεσιακά βάρη, είναι πιο πιθανό να προβλέψει κάποιο θέμα και μάλιστα πολλά μαζί. Σε αντίθεση με το μοντέλο 2 που δίνει πρόβλεψη μόνο σε λίγους τίτλους, όπου όμως δίνει μια πρόβλεψη έχει περισσότερες πιθανότητες να είναι και σωστή. Το μοντέλο 3 δίνει πολλές προβλέψεις για τον ίδιο τίτλο αλλά έχει αρκετές πιθανότητες κάποιες από αυτές να μην είναι σωστές. Κανένα από τα δυο μοντέλα δεν έχουν καταφέρει να πετύχουν πραγματικά καλά αποτελέσματα. Σε αυτό μπορεί

Πίνακας 4.4: Καταγράφουμε τα αποτελέσματα από τις δοκιμές με διαφορετικά είδη (θεσιακό βάρος, βάρος κλάσεων) από βάρη και την επίδραση που έχουν στις μετρικές αλλά και στην έλλειψη πρόβλεψης.

#	Θεσιακό βάρος	Βάρος κλάσεων	Accuracy	Precission	Recall	N/A train	N/A validation	N/A test
1	✓	✓	0.8459	0.4242	0.5590	1.24%	0.32%	0.3%
2	—	✓	0.8949	0.7651	0.2626	40.1%	6.06%	6.04%
3	✓	—	0.8582	0.4586	0.5784	1.22%	0.29%	0.29%
4	—	—	0.8960	0.7067	0.3257	29.07%	4.01%	3.84%

να οφείλεται ότι οι στόχοι δεν είναι πάντα τόσο ευδιάκριτοι, δηλαδή θα μπορούσαν πολλά θέματα να ταίριαζαν σε έναν τίτλο του dataset. Ένας ακόμα λόγος είναι ότι το LAIR δεν έχει αρκετές εγγραφές για να μπορέσει ένα τέτοιο μοντέλο να μάθει επαρκώς τους στόχους του. Παρατηρούμε όμως την τελική επιρροή που έχουν οι προβλέψεις του LSTM στο τελικό μοντέλο. Στα Σχήμα 4.1 έχουμε τα διαγράμματα από την εκτέλεση 2. Παρατηρούμε γενικά μια απότομη βελτίωση του accuracy και περίπου στο βήμα εκτέλεσης 1250 ξεκινάει το overfit του μοντέλου.

4.3 Αποτελέσματα ανίχνευσης συναισθήματος

Όπως αναλύθηκε στα προηγούμενα κεφάλαια υπάρχουν πολλές έρευνες που έχουν συμπεράνει με δοκιμές ότι το συναίσθημα που εκφράζει μια πρόταση μπορεί να χρησιμοποιηθεί στην ανίχνευση ψεύδους. Αυτό προκύπτει γιατί πολλές είναι οι περιπτώσεις που προσπαθεί κάποιος να εκφράσει μια υπερβολή ώστε να μας πείσει για κάτι που ενδεχομένως είναι ψέμα. Και οι δύο τεχνικές ανίχνευσης συναισθήματος που χρησιμοποιούμε έχουν τρεις πιθανές κατηγορίες, αρνητική, ουδέτερη και θετική. Λόγο ότι δεν έχουμε σαν πληροφορία το συναίσθημα για τις εγγραφές του LAIR δεν μπορούμε να ξέρουμε την απόδοση των δυο τεχνικών που χρησιμοποιήθηκαν στην εργασία αυτήν. Αυτό που μπορούμε να παρατηρήσουμε είναι στο τελικό μοντέλο άμα οι προβλέψεις του συναισθήματος βοηθάνε στην ανίχνευση του ψεύδους ή όχι και συγκρίνοντάς αυτά τα αποτελέσματα ανάμεσα στην χρήση των συναισθημάτων, από VADER και RoBERTa, ώστε να παρατηρήσουμε πιο από τα δυο βοήθησε παραπάνω.

4.3.1 Ανίχνευση συναισθήματος με χρήση λεξικού

Για την ανίχνευση συναισθήματος με λεξικό χρησιμοποιήθηκε το VADER. Το VADER προσπαθεί να προβλέψει το συναίσθημα χρησιμοποιώντας προκαθορισμένες λέξεις που κατέχει και έχουν τοποθετηθεί βάρη σε αυτές με σκοπό να ορίζουν άμα η λέξη είναι θετική η αρνητική. Με αυτόν τον τρόπο βγαίνει ένα τελικό αποτέλεσμα για μια πρόταση. Το πρόβλημα με αυτήν την μεθοδολογία είναι ότι πολλές λέξεις έχουν άλλο πρόσημο ανάλογα με το γενικό νόημα τις πρότασης. Για την καλύτερη πρόβλεψη του συναισθήματος χρησιμοποιήθηκε το pre-processed dataset όπως περιγράφεται στον Πίνακα 3.2. Στον Πίνακα 4.6 περιγράφουμε το

Πίνακας 4.5: Οι τίτλοι αυτοί είναι από άρθρα που δεν υπάρχουν στο LIAR. Οι τελευταίες στήλες είναι οι προβλέψεις των θεμάτων από τα δύο καλύτερα μοντέλα που κατασκευάστηκαν με LSTM.

Πρόταση	**Πρόβλεψη μοντέλο 3**	**Πρόβλεψη μοντέλο 2**
Pam Bondi defends Epstein files handling as House speaker calls for Ghislaine Maxwell to testify before Congress	Politics, Law	Politics
What does Trump's "Big Beautiful Bill" mean for the US economy?	—	—
WHO recommends Gileadś twice-yearly injection for HIV prevention	Government, Infrastructure	—
Smoke from climate-fueled fires in US contributed to 15,000 deaths in 15 years, study finds	Environment, Government, Infrastructure	—
Supreme Court ruling "devastating" for transgender kids: Lawyer	Society, Law	—
Largest teachers union slams "unlawful" cuts to Department of Education after Supreme Court ruling	Society, Education	—
US transportation chief to detail plan to overhaul air traffic control	Environment, Government, Infrastructure	Economy
Supreme Court upholds ban on transgender treatments for minors	Society, Security, Law	—
Hundreds of thousands due refunds as telecoms apologise for overcharging	—	—
Senate debates FAA reauthorization as air traffic control overhaul gains bipartisan momentum	Economy, Government, Business, Infrastructure	Economy
U.S. extradition stance tested as Congress seeks testimony from foreign Epstein associates	Politics, Security, Law	Politics
Philadelphia Eagles and Packers to play NFL's first game in South America	Health, Government, Infrastructure, Sports	—
Solar farming gains ground: USDA backs agrivoltaics pilot projects in Midwes	Environment, Government, Infrastructure	—

πλήθος των εγγραφών που ταξινομήθηκαν στις τρεις κατηγορίες συναισθήματος ανάμεσα στα dataset του LAIR. Παρατηρούμε ότι το VADER στο συγκεκριμένο dataset ταξινομεί περίπου ισόποσα τις εγγραφές ανάμεσα στις κατηγορίες. Αυτό δεν είναι ένα ρεαλιστικό αποτέλεσμα, γιατί είναι σπάνιο να έχει συμβεί κάτι τέτοιο χωρίς να έχουμε επέμβει στα δεδομένα για να έχουμε ισόποσες εγγραφές για κάθε μια κατηγορία συναισθήματος.

Για μια καλύτερη στατιστική κατανόηση των προβλέψεων, μιας και δεν είναι δυνατόν να συγκρίνουμε τις προβλέψεις με τα πραγματικά δεδομένα, υπολογίστηκαν κάποια βασικά στατιστικά στοιχεία όπως το mean, median, min, max, std dev για το entropy και το confidence του VADER, σύμφωνα με το train dataset του LAIR. Το entropy είναι μια τιμή που μας δείχνει κατά πόσο οι προβλέψεις φαίνονται να είναι τυχαίες ή στοχευμένες. Όσο μικρότερος είναι ο αριθμός του entropy τόσο πιο στοχευμένες είναι φαίνονται να είναι οι προβλέψεις, το αντίθετο όσο είναι μεγαλύτερος είναι ο αριθμός. Το confidence πηγάζει από τους αριθμούς που παράγει το VADER για να μπορέσουμε να ταξινομήσουμε τους τίτλους. Όσο μεγαλύτερος είναι ο αριθμός τόσο πιο σίγουρο για την πρόβλεψη του νιώθει το VADER. Στον Πίνακα 4.7 παρατηρούμε ότι το entropy έχει μια σχετικά χαμηλή τιμή. Το std dev δεν είναι και τόσο χαμηλό που σημαίνει ότι δεν είναι πολύ σταθερό το entropy ανάμεσα στις προβλέψεις. Αυτό απεικονίζεται καλύτερα στο Σχήμα 4.2 όπου έχουμε πολλές εγγραφές με μηδενικό entropy αλλά οι περισσότερες είναι μαζεμένες γύρο από τον μέσο όρο. Στον ίδιο πίνακα παρατηρούμε ότι το confidence είναι σχετικά υψηλό αλλά έχει αρκετά χαμηλή τιμή std dev. Αυτό σημαίνει ότι οι προβλέψεις του είναι γενικά υψηλές σε όλο το dataset. Στο Σχήμα 4.3 απεικονίζεται ένα πολυ μεγάλο ποσοστό των εγγραφών να έχουν confidence 1.0 και επίσης οι περισσότερες εγγραφές είναι κοντά στο 0.75 που τείνουν γενικά προς το 1.0. Αυτό δείχνει και με γραφικό τρόπο ότι οι προβλέψεις έγιναν με υψηλό confidence. Βέβαια όλα αυτά δεν αποδεικνύουν ότι το μοντέλο έχει παράξει σωστές προβλέψεις, είναι απλά στατιστικοί αριθμοί που μας δείχνουν με μια γενική οπτική γωνία άμα το μοντέλο έχει επιλέξει τις προβλέψεις έχοντας κάποια λογική και όχι στην τύχη.

4.3.2 Ανίχνευση συναισθήματος με χρήση RoBERTa

Το RoBERTa είναι μια αναβάθμιση του ήδη πετυχημένου μοντέλου BERT που ειδικεύεται πιο πολύ σε κείμενο από social media. Τα κείμενα έχουμε από το LAIR είναι τίτλοι από άρθρα, άρα υποθέτουμε ότι υπάρχει μια σχετική συγγένεια ανάμεσα στους τίτλους αυτούς και τα κείμενα από social media. Γιατί ακόμα και στα social media πολλές φορές γίνονται αναφορές και συζητήσεις σε πολιτικά θέματα. Άρα περιμένουμε να έχει πολύ θετικά αποτελέσματα στις

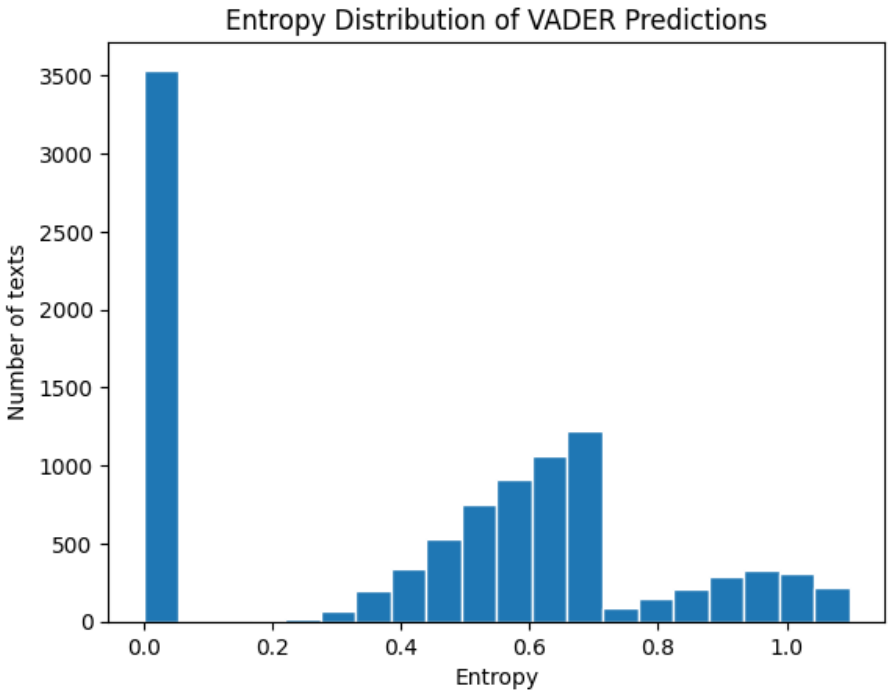
Πίνακας 4.6: Τα αποτελέσματα των προβλέψεων συναισθημάτων για τα dataset με την χρήση του VADER. Παρατηρούμε ότι η ταξινόμηση είναι σχεδόν ισόποση χωρίς εμείς να έχουμε επέμβει με κάποιον τρόπο πάνω σε αυτό.

Dataset	**Αρνητικές**	**Ουδέτερες**	**Θετικές**
Train	3208	3688	3344
Validation	410	427	447
Test	403	427	437

Πίνακας 4.7: Τα στατιστικά δεδομένα για τις προβλέψεις του VADER. Το entropy έχει σχετικά χαμηλή τιμή και το confidence σε σχετικά υψηλή τιμή.

Μετρική	**Mean**	**Median**	**Std Dev**	**Min**	**Max**
Entropy	0.429	0.526	0.346	0.0	1.098
Confidence	0.793	0.783	0.179	0.344	1.0

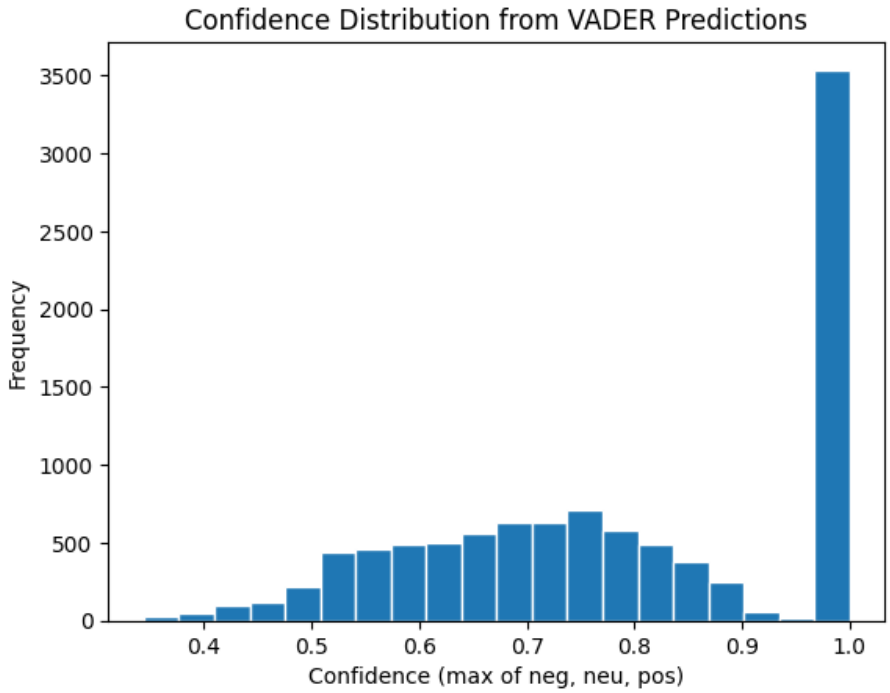
προβλέψεις που θα κάνει για το sentiment. Βέβαια όπως έχουμε πει ήδη τέτοια μεγάλα μοντέλα θέλουν και αρκετό κείμενο για να μπορέσουν να αποδώσου καλά και σε πλήθος κειμένου σε κάθε εγγραφή αλλά και γενικά πολλές εγγραφές, που κάτι τέτοιο δεν συναντάμε στο LAIR. Παρόλα αυτά μπορούμε να δούμε στον Πίνακα 4.8 το πλήθος εγγραφών που ταξινόμησε σε



Σχήμα 4.2: Ιστόγραμμα που δείχνει το πλήθος των εγγραφών σε σχέση με την τιμή entropy που έχουν. Οι περισσότερες εγγραφές έχουν τιμές κοντά στον μέσο όρο του entropy.

Πίνακας 4.8: Τα αποτελέσματα των προβλέψεων συναισθημάτων για τα dataset με την χρήση του RoBERTa. Παρατηρούμε ανισότητες στο πλήθος των ταξινομήσεων που είναι πιο λογική σαν εικόνα σε σχέση με τις προβλέψεις του VADER.

Dataset	**Αρνητικές**	**Ουδέτερες**	**Θετικές**
Train	3834	5560	846
Validation	502	683	99
Test	466	694	107



Σχήμα 4.3: Ιστόγραμμα που δείχνει το πλήθος των εγγραφών σε σχέση με την τιμή entropy που έχουν. Οι περισσότερες εγγραφές έχουν τιμές κοντά στον μέσο όρο του entropy.

Πίνακας 4.9: Τα στατιστικά δεδομένα για τις προβλέψεις του RoBERTa. Το entropy είναι σε σχετικά χαμηλή τιμή και το confidence σε σχετικά υψηλή τιμή. Το std dev είναι μικρός αριθμός που σημαίνει ότι οι περισσότερες τιμές είναι κοντά στο mean.

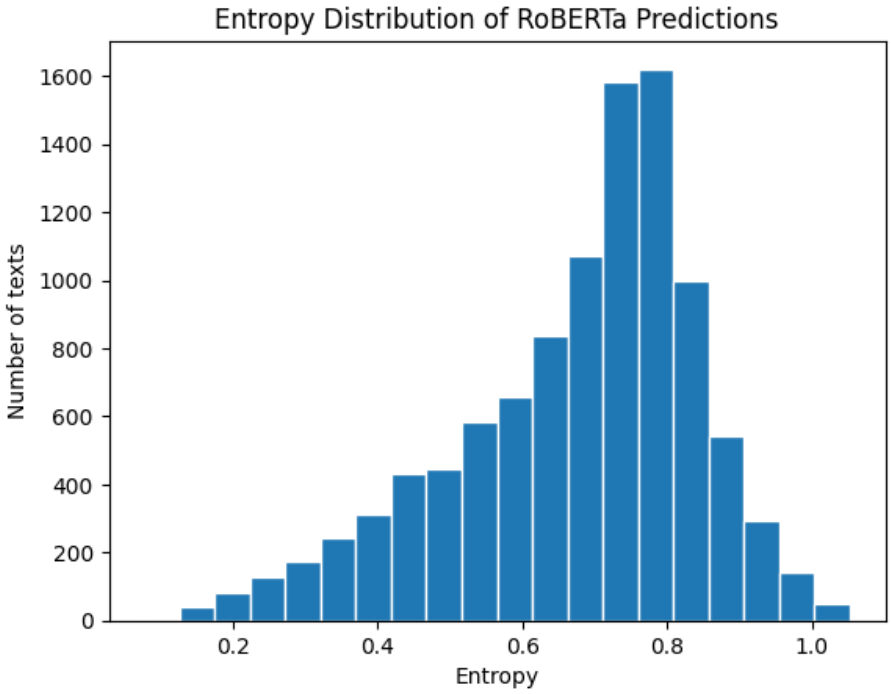
Μετρική	**Mean**	**Median**	**Std Dev**	**Min**	**Max**
Entropy	0.674	0.716	0.170	0.076	1.053
Confidence	0.691	0.683	0.129	0.430	0.986

κάθε κατηγορία για τα τρία dataset του LAIR. Εδώ σε σχέση με το VADER εντοπίζουμε ότι δεν έχει ισόποσα πλήθη από κάθε κατηγορία, αυτό αρχικά δείχνει ότι λογικά έχει αποδώσει καλύτερα από το VADER για τους λόγους που είπαμε πριν στην ανάλυση του VADER. Γενικά από τον Πίνακα 4.8 κατανοούμε ότι οι περισσότερες προβλέψεις είναι ουδέτερες προτάσεις, μετά αρνητικές και τέλος λιγότερες είναι οι θετικές. Στον Πίνακα 4.9 με μια ματιά μπορούμε να παρατηρήσουμε πιο ρεαλιστικές τιμές στατιστικών σε σχέση με το VADER. Συμπεραίνουμε ότι το μοντέλο είναι πιο σταθερό τις προβλέψεις του γιατί τα std dev του entropy αλλά και του confidence είναι σε χαμηλές τιμές. Επίσης το min confidence είναι υψηλό (0.430). Τις τιμές αυτές μπορούμε να της παρατηρήσουμε και στο Σχήμα 4.4 όπου υπάρχει μια τάση του entropy προς τις υψηλές τιμές. Παρόλα αυτά, το γράφημα από μια γενική οπτική γωνία έχει μια πιο φυσιολογική απεικόνιση σε σχέση με το αντίστοιχο του VADER. Το confidence Σχήμα 4.5 δείχνει μια περισσότερο απλωμένη γραφική παράσταση σε σχέση με το VADER με μέση τιμή το 0.691 και έχοντας συγκεντρώσει τις περισσότερες τιμές κοντά στο mean. Βέβαια υπάρχουν και αρκετές τιμές πέρα από το mean προς το 1.0.

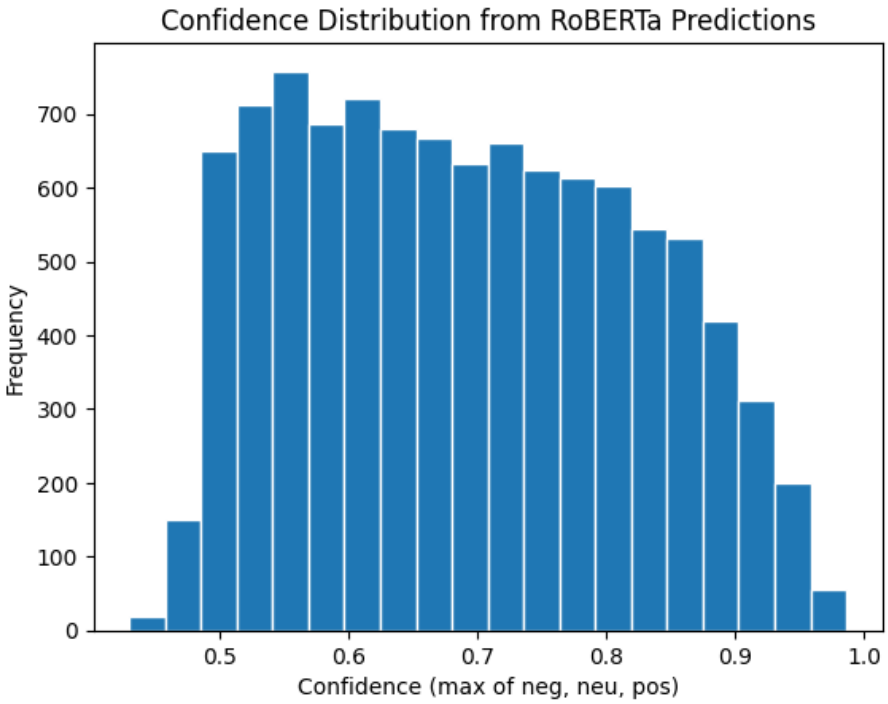
Σε γενικές γραμμές μπορούμε να θέσουμε ότι το RoBERTa παρατηρείται παρέχει πιο φυσιολογικές προβλέψεις σε σχέση με το VADER, είναι πιο ακατάστατες από την μεριά των στατιστικών αλλά πολύ ισόποσες σε πλήθος εγγραφών ανά κλάση. Και πάλι με αυτά τα στατιστικά στοιχεία όσο καλά και να είναι δεν μας δείχνουν πόσο σωστές είναι οι προβλέψεις του μοντέλου.

4.4 Αποτελέσματα τελικού μοντέλου

Με το τελικό μοντέλο, που την δομή του την έχουμε περιγράψει σε προηγούμενο κεφάλαιο, έχουν εκτελεστεί όλοι οι πιθανοί συνδυασμοί των παραπάνω δεδομένων, θεμάτων και της ανάλυσης συναισθήματος. Αυτό έγινε με σκοπό να κατανοήσουμε ποια πληροφορία είναι αυτή που βοηθάει περισσότερο το μοντέλο να προβλέψει το ψεύδος. Επίσης ειδικά για το LSTM εκτελέστηκαν οι προβλέψεις του με και χωρίς το θεσιακό βάρος και πάλι για να δούμε άμα έχουμε καλύτερες προβλέψεις όταν τα θέματα έχουν παραχθεί από το LSTM μοντέλο με θεσιακό βάρος ή χωρίς θεσιακό βάρος. Λόγο ότι είναι multiclass classification πρόβλημα κρατήθηκαν πολλές μετρικές για να μπορέσουμε να έχουμε καλύτερη εικόνα της επίδοσης του μοντέλου. Παρατηρώντας το validation dataset κρατήθηκαν snapshots του εκπαιδευόμενου μοντέλου με το υψηλότερο accuracy και το χαμηλότερο error. Παρατηρούμε ότι κάποια μοντέλα παράγουν καλύτερα αποτελέσματα στο test dataset όταν έχουν επιλεχθεί με τον γνώμονα του καλύτερου accuracy και άλλοτε με το καλύτερο error. Έτσι δεν χρειαζόταν να σταματάμε την εκπαίδευση του μοντέλου πρόωρα για να μην πάθει overfit γιατί θα έχει κρατήσει το



Σχήμα 4.4: Ιστόγραμμα που δείχνει το πλήθος των εγγραφών σε σχέση με την τιμή entropy που έχουν. Οι περισσότερες εγγραφές έχουν τιμές κοντά στον μέσο όρο του entropy.



Σχήμα 4.5: Ιστόγραμμα που δείχνει το πλήθος των εγγραφών σε σχέση με την τιμή entropy που έχουν. Οι περισσότερες εγγραφές έχουν τιμές κοντά στον μέσο όρο του entropy.

παλιό snapshot πριν ξεκινήσει το overfit. Στον Πίνακα 4.10 καταγράψαμε τις προβλέψεις που έχουν κάνει οι ερευνητές και δημιουργοί του LIAR dataset, που αυτά τα score είναι τα καλύτερα που πέτυχαν και έγιναν με χρήση όλων τον δεδομένων του dataset και έχοντας σαν στόχο όλες τις κλάσεις του ψεύδους. Όλα τα αποτελέσματα στον πίνακα αυτόν είναι για το test dataset και μόνο. Στην δημοσίευση του LAIR δεν παρέχουν άλλες μετρικές πέρα του accuracy. Αυτό δεν βοηθάει και πολύ στην σύγκριση των μοντέλων μας γιατί δεν παρέχει αρκετή πληροφορία από μόνη της αυτή η μετρική. Εμείς για να έχουμε μια ολοκληρωμένη εικόνα για την απόδοση του μοντέλου μας παρέχουμε πολλές μετρικές. Το kappa score μας δείχνει κατά πόσο υπάρχει κάποια πραγματική πρόβλεψη των στόχων του μοντέλου γιατί συνυπολογίζει την πιθανότητα το μοντέλο να πετύχει από τύχη την σωστή κλάση. Βέβαια στο συγκεκριμένο dataset είναι πιο απίθανο να συμβεί αυτό γιατί οι κλάσεις είναι έξι. Στην πρώτη γραμμή του πίνακα υπάρχουν τα αποτελέσματα του μοντέλου χωρίς να του έχουμε παρέχει κάποια επιπλέον πληροφορία από θέματα ή συναισθήματα. Αυτά τα αποτελέσματα είναι το σημείο αναφοράς για όλες τις επόμενες εκτελέσεις που έγιναν και τις συγκρίνουμε με αυτά ώστε να καταλάβουμε άμα βοήθησαν περισσότερο ή όχι. Από Πίνακα 4.10 μπορούμε να παρατηρούμε ότι το τελικό μοντέλο χωρίς κανένα από τα παραπάνω δεδομένα, όπως τα θέματα ή τα συναισθήματα, πετυχαίνει καλύτερο accuracy από τη δημοσίευση του LAIR. Εμείς περιμέναμε τα καλύτερα αποτελέσματα να τα πετύχουμε με τον συνδυασμό θεμάτων και συναισθημάτων. Αλλά στην πραγματικότητα πετύχαμε το καλύτερο accuracy (0.2867) με τα συναισθήματα που παράχθηκαν από το RoBERTa και χωρίς την πληροφορία των θεμάτων. Αυτός ο συνδυασμός έχει και το καλύτερο recall και το δεύτερο καλύτερο precision, όλα αυτά μαζί αποδεικνύουν ότι έχει σωστές προβλέψεις και καλύπτει καλά τους στόχους του dataset.

Πίνακας 4.10: Τα αποτελέσματα από όλους του συνδυασμούς των δεδομένων και σε σύγκριση με το καλύτερο accuracy από την δημοσίευση του LAIR dataset, με μέθοδο σφάλματος το cross entropy.

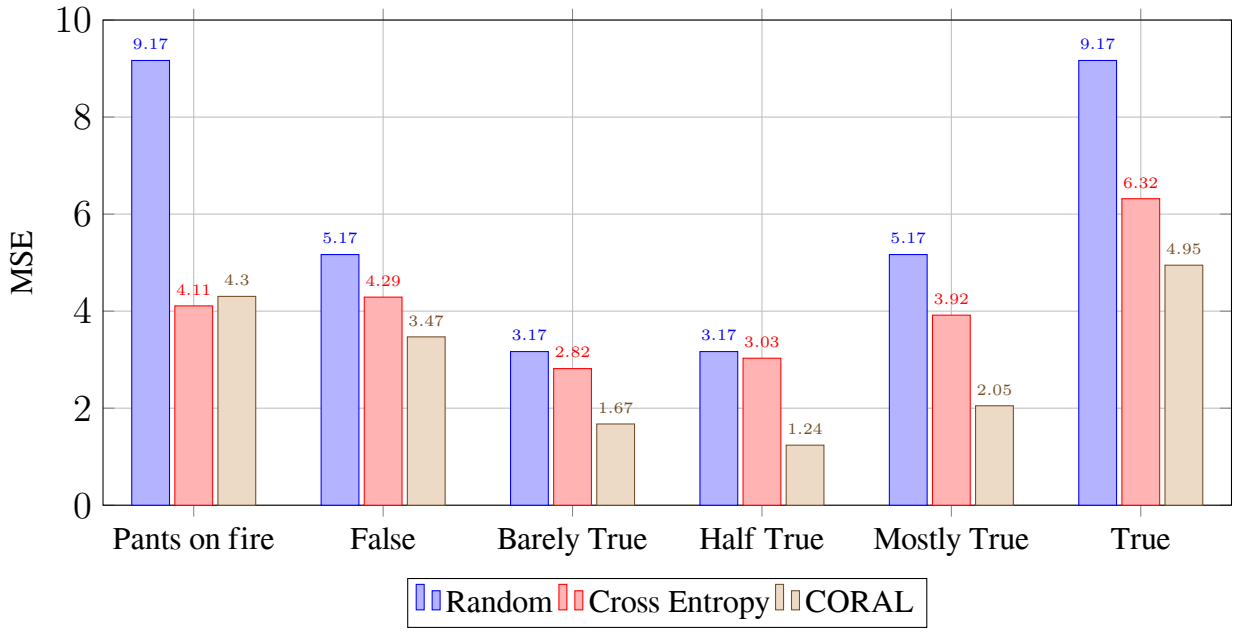
Θέματα		**Συναισθήματα**		**Μετρικές**					
LDA	**LSTM**	**VADER**	**RoBERTa**	**Accuracy**	**F1**	**Precision**	**Recall**	**Kappa**	**MSE**
—	—	—	—	0.2773	**0.2508**	0.2499	0.2773	0.2614	4.3317
✓	—	—	—	0.2814	0.2354	**0.2578**	0.2814	**0.2759**	4.5313
—	✓	—	—	0.2621	0.2201	0.2241	0.2621	0.2557	4.6607
—	—	✓	—	0.2700	0.2230	0.2447	0.2700	0.2389	4.7108
—	—	—	✓	**0.2867**	0.2438	0.2509	**0.2867**	0.2620	4.5127
✓	—	✓	—	0.2667	0.2255	0.2343	0.2667	0.2300	4.7533
✓	—	—	✓	0.2644	0.2154	0.2368	0.2644	0.2431	5.1959
—	✓	✓	—	0.2667	0.2036	0.2167	0.2667	0.2437	5.1406
—	✓	—	✓	0.2643	0.2445	0.2454	0.2643	0.2701	**4.0797**
Αποτελέσματα LAIR				0.2740	—	—	—	—	—

Μέσα στις μετρικές έχουμε και το μέσο τετραγωνικό σφάλμα (Mean Squared Error – MSE). Αυτό υπάρχει γιατί οι κλάσεις στην πραγματικότητα είναι κλιμακωτές και έτσι μπορούμε να υποθέσουμε ότι το μοντέλο μπορεί να μην έχει πετύχει ακριβώς την κλάση που έπρεπε, αλλά την προσέγγισε προβλέποντας μια κοντινή. Ανάλογα με την απόκλιση αυτή παράγεται και το MSE· όσο μεγαλύτερο, τόσο μεγαλύτερη η απόκλιση από τον στόχο, άρα και το σφάλμα. Έτσι, το καλύτερο MSE έχει ο συνδυασμός πληροφορίας από LSTM θέματα και συναισθήματα από RoBERTa, παρόλο που έχει χειρότερο accuracy και από το βασικό μοντέλο χωρίς περαιτέρω πληροφορίες. Επίσης, αυτός ο συνδυασμός έχει το δεύτερο καλύτερο F1 score, σχετικά καλό precision και μέτριο recall. Αυτό, κατά κάποιον τρόπο, επιβεβαιώνει την πρόβλεψή μας ότι ο συνδυασμός των θεμάτων και των συναισθημάτων είναι αυτός που θα έχει τα καλύτερα αποτελέσματα, απλά το κάνει με πιο γενικευμένο τρόπο και όχι με υψηλό accuracy. Παρατηρούμε ότι δεν πετυχαίνει τόσο καλά τις κλάσεις, αλλά γενικά οι προβλέψεις του τις προσεγγίζουν πιο πολύ από όλους τους άλλους συνδυασμούς. Γενικά, για το συγκεκριμένο dataset και όταν οι στόχοι είναι και οι έξι μαζί, είναι φυσιολογικό οι μετρικές του accuracy, precision, recall να είναι κοντά στο 0.27, γιατί όπως συμπεραίναμε και στα προηγούμενα κεφάλαια, είναι δύσκολο dataset χωρίς να χρησιμοποιηθεί παραπάνω πληροφορία από αυτήν που παρέχεται. Στον Πίνακα 4.11 περιγράφουμε τα αποτελέσματα κατά τα ίδια δεδομένα όπως στον Πίνακα 4.10, με την διαφορά ότι η συνάρτηση εξόδου και η μέθοδος υπολογισμού του σφάλματος είναι το CORAL. Όπως αναλύθηκε σε προηγούμενο κεφάλαιο, το CORAL αναμένεται να οδηγεί σε χαμηλότερη τιμή MSE. Γιατί ο σκοπός του δεν είναι να πετύχει ακριβώς τον στόχο της κλάσης αλλά να την προσεγγίσει όσο περισσότερο γίνεται. Το CORAL καθοδηγείται προς την κατεύθυνση αυτή, γιατί σχεδιάστηκε να μπορεί να δώσει προβλέψεις σε multiclass προβλήματα όπου οι κλάσεις είναι κλιμακωτές. Με το CORAL παρατηρούμε τελείως διαφορετικά αποτελέσματα σε σχέση με τα προηγούμενα από το cross entropy. Παρατηρούμε ανάμεσα στα αποτελέσματά του μια εκτέλεση που ξεχωρίζει και έχει παράξει τις καλύτερες μετρικές. Βρίσκεται στην πρώτη γραμμή του πίνακα και δεν

περιέχει καμιά πληροφορία από θέματα ή συναισθήματα. Αυτό οξύμωρο στην αρχή — πως γίνεται, όταν παρέχουμε λιγότερη πληροφορία σε ένα μοντέλο, αυτό να αποδίδει καλύτερα; Στην πραγματικότητα, αυτό το φαινόμενο υπάρχει και στις εκτελέσεις με cross entropy, απλά βρίσκεται σε μικρότερο βαθμό και δεν είναι τόσο ευδιάκριτο. Συμπερασματικά, είναι ο ίδιος λόγος που στο cross entropy, όταν τα θέματα και τα συναισθήματα από μόνα τους παράγουν καλύτερα αποτελέσματα από όταν συνδυάζονται μαζί. Το ίδιο θέμα υπάρχει και εδώ με το

Πίνακας 4.11: Τα αποτελέσματα από όλες τις εκτελέσεις στο LIAR, αλλά με συνάρτηση εξόδου και μέθοδο σφάλματος το CORAL.

Θέματα		**Συναισθήματα**		**Μετρικές**					
LDA	**LSTM**	**VADER**	**RoBERTa**	**Accuracy**	**F1**	**Precision**	**Recall**	**Kappa**	**MSE**
—	—	—	—	**0.2615**	**0.2435**	0.2795	**0.2615**	**0.3104**	**2.9472**
✓	—	—	—	0.2469	0.2146	0.2596	0.2469	0.3013	2.9607
—	✓	—	—	0.2475	0.2120	**0.2879**	0.2475	0.2739	3.1767
—	—	✓	—	0.2514	0.2203	0.2677	0.2514	0.2646	3.1747
—	—	—	✓	0.2454	0.2251	0.2599	0.2454	0.2825	3.1927
✓	—	✓	—	0.2554	0.2233	0.2706	0.2554	0.2758	3.0154
✓	—	—	✓	0.2541	0.2251	0.2582	0.2541	0.2779	3.1586
—	✓	✓	—	0.2495	0.2267	0.2612	0.2495	0.2873	3.0770
—	✓	—	✓	0.2403	0.2238	0.2564	0.2403	0.3015	3.0719
Αποτελέσματα LAIR				0.2740	—	—	—	—	—



Σχήμα 4.6: Σύγκριση ανά κλάση των καλύτερων MSE ανάμεσα σε cross entropy και CORAL. Επίσης σαν αναφορά έχουμε και τις τυχαίες προβλέψεις. Εντοπίζουμε ότι το CORAL έχει πολυ χαμηλότερα MSE σε όλες τις κλάσεις εκτός της κλάσης pants on fire.

CORAL, αλλά σε μεγαλύτερο βαθμό. Στο CORAL έχουμε το καλύτερο precision με τα δεδομένα των θεμάτων από το LSTM· πέρα από αυτό το αποτέλεσμα, όλα τα άλλα είναι χειρότερα από την εκτέλεση χωρίς τα παραπάνω δεδομένα θεμάτων και συναισθημάτων. Ένας λόγος που μπορεί να γίνεται αυτό είναι ότι τα παραπάνω δεδομένα δεν βοηθούν το μοντέλο να ξεχωρίσει κάθε κλάση σε καλό βαθμό, αλλά δημιουργούν μεροληψίες που κάνουν το μοντέλο να επικεντρώνεται σε συγκεκριμένες κλάσεις. Η προηγούμενη υπόθεσή μας, ότι με το CORAL θα έχουμε χαμηλότερο MSE, αποτυπώνεται σε όλες τις εκτελέσεις του CORAL. Από όλους τους συνδυασμούς πληροφοριών που έχουμε δοκιμάσει, το CORAL παράγει αρκετά καλύτερο MSE από όλες τις εκτελέσεις του cross entropy. Στο Σχήμα 4.6 περιγράφουμε τις δύο καλύτερες εκτελέσεις σύμφωνα με το MSE από cross entropy και το CORAL. Στο σχήμα έχουμε αναλυτικά το MSE για κάθε κλάση. Το average, σαν τιμή — και ας είναι χαμηλό — δεν μας δείχνει αν η τιμή του προέκυψε από μια καλή κλάση ή από μια ομοιομορφία στις προβλέψεις σε όλες τις κλάσεις. Γενικά, σε ένα μοντέλο ψάχνουμε να πετυχαίνει καλά όλες τις κλάσεις μαζί και όχι μόνο μερικές από αυτές, για να έχει καλή γενίκευση και όχι ειδίκευση. Από τα αποτελέσματα του σχήματος είναι εμφανές ότι το CORAL παράγει καλύτερο MSE σε όλες τις κλάσεις σε σχέση με το cross entropy. Στο CORAL υπάρχει μια καλύτερη απόδοση στις κεντρικές κλάσεις, αλλά είναι σχετικά "απλωμένη" αριστερά και δεξιά της κεντρικής κλάσης, ενώ στο cross entropy έχουμε μια καλή απόδοση σε μια από τις κεντρικές κλάσεις, αλλά έχει μεγάλη διαφορά σε σχέση με τις γειτονικές κλάσεις. Στο ίδιο σχήμα, σαν αναφορά, υπάρχουν και τα MSE που θα προβλέπαμε αν οι προβλέψεις μας ήταν τελείως τυχαίες και το LAIR είχε ακριβώς ισόποσες τις κλάσεις του. Αυτό το κάναμε για να έχουμε ένα σημείο αναφοράς σαν βάση και να συγκρίνουμε τις αποδόσεις σε σχέση με αυτό. Επίσης, μας δείχνει ότι και στην περίπτωση των τυχαίων προβλέψεων έχουμε ένα καλύτερο MSE στις κεντρικές κλάσεις και όσο πλησιάζουμε προς τις ακριανές κλάσεις γίνεται χειρότερο. Άρα είναι αναμενόμενο να δούμε μια τέτοια εικόνα και στις προβλέψεις από τα μοντέλα μας. Βέβαια, όσο καλύτερο είναι ένα μοντέλο, τόσο αυτό το φαινόμενο θα εμφανίζεται σε μικρότερο βαθμό και θα μπορεί να προβλέπει εξίσου καλά και τις ακριανές κλάσεις.

4.5 Επίλογος

Στο κεφάλαιο αυτό αναλύσαμε όλες τις δοκιμές που έγιναν με όλα τα μοντέλα για την εξαγωγή των επιπλέον πληροφοριών, όπως τα θέματα αλλά και τα συναισθήματα. Οι προβλέψεις των μοντέλων για την ανάλυση συναισθήματος δεν είχαν στόχους που να παρέχει το LAIR και έτσι έγινε στατιστική ανάλυση αυτών. Περιγράψαμε πως το LDA κατασκευάζει τα θέματα του, οργανώνοντας τις λέξεις σε σύνολα, και το LSTM πως καταφέρνει να προβλέψει θέματα για τίτλους που δεν έχει εκπαιδευτεί σε αυτούς. Τα αποτελέσματα του τελικού μοντέλου, μέχρι ένα σημείο, ήταν αναμενόμενα, αλλά υπήρξαν και πολλές αναπάντεχες καταστάσεις, όπως όταν το μοντέλο με παραπάνω δεδομένα είχε χαμηλότερη απόδοση. Τέλος, συγκρίναμε τις δύο μορφές του τελικού μοντέλου, τη μία εκπαιδευόμενη με το cross entropy και την άλλη με το CORAL. Το cross entropy πετυχαίνει γενικά καλύτερα αποτελέσματα στις μετρικές, αλλά αντιμετωπίζει τις κλάσεις ως διακριτές μεταξύ τους και έτσι μειώνεται το MSE του. Το CORAL κάνει το ακριβώς αντίθετο· είναι σχεδιασμένο για multiclass όπου οι κλάσεις μπορούν να τεθούν σε μια λογική σειρά, όπως είναι στην περίπτωση του LAIR. Αυτό το οδηγεί στο να μην παράγει καλές μετρικές όπως το cross entropy, αλλά πετυχαίνει πολύ καλύτερο MSE.

Κεφάλαιο 5

Συμπεράσματα

Όπως προκύπτει από τις έρευνες που έχουν γίνει, αλλά και με την κοινή λογική, μπορούμε να κατανοήσουμε ότι οι ψευδείς ειδήσεις έχουν μεγάλη επίπτωση στη ζωή των ανθρώπων αλλά και στις αποφάσεις που θα πάρουν, μικρές ή μεγάλες. Για αυτόν τον λόγο είναι σημαντικό να αναπτυχθούν τεχνικές και μεθοδολογίες που μπορούν να τις ανιχνεύσουν ώστε να προλάβουμε την εξάπλωση τους πριν δημιουργήσουν μεγάλο κακό. Η τεχνητή νοημοσύνη είναι το μέσο που χρησιμοποιούν οι περισσότεροι για την ανίχνευση του ψευδούς λόγου, αν και το ίδιο μέσο είναι αυτό που πολλές φορές χρησιμοποιούν και αυτοί που θέλουν να παράξουν μια ψευδή είδηση. Η παρούσα εργασία επικεντρώνεται στην ανίχνευση του ψεύδους στον πολιτικό λόγο, μέσα από τα άρθρα που αναρτώνται στις ιστοσελίδες. Πιο συγκεκριμένα, μέσα από το LAIR dataset, που περιέχει πληροφορίες από άρθρα της Αμερικής για θέματα όπως η πολιτική και η οικονομία, μαζί με διάφορες δημόσιες καταθέσεις που έχουν διατυπωθεί από πολιτικά πρόσωπα στην Αμερική. Ο λόγος που οι περισσότεροι στρέφονται στη χρήση της τεχνητής νοημοσύνης για την ανίχνευση του ψεύδους είναι ότι έχει αποδείξει έμπρακτα, αλλά και μέσα από έρευνες, ότι μπορεί να αναλύσει με αποτελεσματικότητα κείμενο φυσικής γλώσσας και να κατανοήσει το νόημα του, να αντλήσει τα συναισθήματα που εκφράζει, καθώς και να το ταξινομήσει σε κατηγορίες. Από τη φύση της, η πρόβλεψη ψεύδους στον γραπτό λόγο είναι πολύ δύσκολη. Για να μπορέσει κάποιος να προβλέψει μια πρόταση αν είναι ψευδής ή αληθής, χωρίς να του παρέχουμε κάποια παραπάνω πληροφορία, θα πρέπει να είναι γνώστης του θέματος και να γνωρίζει ήδη την αλήθεια ή να έχει πληροφορίες για την πηγή, ποιος το έγραψε και το παρελθόν του, ώστε να κάνει μια πρόβλεψη. Επειδή το να έχουμε όλες τις πιθανές γνώσεις που μπορεί να χρειαστούν είναι πολύ δύσκολο να συμβεί, για αυτόν τον λόγο εστιάζουμε σε πληροφορίες που σχετίζονται με το ιστορικό του προσώπου που το έγραψε, τα συναισθήματα που εκφράζει και το θέμα που σχετίζεται. Αυτό κάναμε και σε αυτήν την εργασία: το LAIR παρέχει ήδη κάποιες πληροφορίες σχετικές με τον τίτλο του άρθρου, αλλά εμείς, στοχεύοντας στο να παράξουμε καλύτερα αποτελέσματα, εξάγαμε το θέμα του τίτλου μαζί με το συναίσθημα, θέλοντας να αυξήσουμε την απόδοση του μοντέλου μας. Είναι δύσκολο για ένα μοντέλο μηχανικής μάθησης να προβλέψει το ψεύδος για τους λόγους που είπαμε, αλλά και πιο συγκεκριμένα για το LAIR, που δεν προβλέπει απλά αν είναι ψευδής ή αληθής, αλλά και κατά πόσο ψευδής ή αληθής είναι — κάτι που πολλές φορές θα μπορούσαμε να πούμε ότι είναι υποκειμενικό μετά από ένα σημείο. Άρα, τα αποτελέσματα που περιμέναμε να δούμε από τα μοντέλα μας, ξέραμε ότι δεν θα είναι τόσο καλά όσο θα έβλεπε κάποιος σε άλλα προβλήματα που λύνει η τεχνητή νοημοσύνη.

Από τα τελικά αποτελέσματα μπορούμε να συμπεράνουμε ότι έχουμε πετύχει μια βελτίωση στις προβλέψεις σε σχέση με τη δημοσίευση του LAIR. Όπου εμείς με το βασικό μοντέλο και με την επιπλέον πληροφορία του συναισθήματος από το RoBERTa πετύχαμε accuracy 0.2867 ενώ οι ερευνητές του LIAR 0.2740. Η διαφορά αυτή μπορεί να είναι μικρή αλλά στην πραγματικότητα δεν είναι λόγο της πολυπλοκότητας του, το μικρό του μέγεθος και το μικρό πλήθος λέξεων που έχει στους τίτλους. Αυτά είναι που το καθιστά πολύ δύσκολο για ένα μοντέλο να προβλέψει με μεγάλη ακρίβεια τους στόχους του. Για αυτό και πολλοί ερευνητές εμπλουτίζουν τους τίτλους και τους κάνουν παραγράφους, εστιάζουν σε κάποιες μόνο κατηγορίες και μετατρέπουν το πρόβλημα σε δυαδικό. Για να έχουμε μια πλήρη εικόνα των επιπτώσεων που έχουν οι πληροφορίες των θεμάτων και των συναισθημάτων στην ανίχνευση του ψεύδους, έγιναν δοκιμές με όλους τους πιθανούς συνδυασμούς, έχοντας σημείο αναφοράς τα αποτελέσματα του μοντέλου χωρίς να του έχουμε παρέχει καμία επιπλέον πληροφορία πέρα από αυτές που παρέχει το LAIR. Από τα αποτελέσματα, συμπεραίναμε ότι το LDA κατάφερε να βοηθήσει παραπάνω από το LSTM το τελικό μοντέλο στην ανίχνευση του ψεύδους. Το αποτέλεσμα αυτό πηγάζει από το γεγονός ότι το LSTM αδυνατεί να παρέχει προβλέψεις θέματος για όλες τις εγγραφές και αυτό οδηγεί σε μεροληψία του μοντέλου. Αν μπορούσαμε να επενδύσουμε περισσότερο χρόνο σε πειραματισμό και βελτίωση του LSTM και καταφέρναμε να έχουμε καλύτερες προβλέψεις των θεμάτων με αυτό, ίσως και να βοηθούσε περισσότερο από το LDA. Για την ανίχνευση του συναισθήματος, ήδη από τη στατιστική ανάλυση φαινόταν ότι το λεκτικό μοντέλο VADER δεν θα βοηθούσε πολύ στην πρόβλεψη του ψεύδους. Αντιθέτως, χειροτέρεψε τις προβλέψεις, έχοντας σαν βασικά αποτελέσματα της εκτέλεσης αυτά χωρίς επιπλέον πληροφορία από τα θέματα ή τα συναισθήματα. Όπως φάνηκε και στα τελικά αποτελέσματα, η υποψία μας βγήκε σωστή: το λεκτικό μοντέλο δεν βοήθησε τόσο όσο το RoBERTa, όπου μόνο με αυτό πετύχαμε τα καλύτερα αποτελέσματα σε accuracy και precision. Ο συνδυασμός που αναμέναμε να έχει τα τα καλύτερα αποτελέσματα, είναι όταν παρείχαμε στο μοντέλο τα θέματα μαζί με τα συναισθήματα από το RoBERTa. Από τα αποτελέσματα όμως παρατηρούμε ότι με αυτόν τον συνδυασμό πετύχαμε χειρότερα αποτελέσματα ακόμα και από τη βασική εκτέλεση. Αυτό μας δείχνει ότι το μοντέλο εστιάζει περισσότερο σε συγκεκριμένες εγγραφές και να χάνει τη γενίκευση του, όταν του παρέχουμε όλες αυτές τις πληροφορίες. Στη συνέχεια περιγράψαμε τα αποτελέσματα του μοντέλου όταν χρησιμοποιήθηκε το CORAL για συνάρτηση σφάλματος. Αυτό που ήταν αναμενόμενο και συνέβη είναι ότι με το CORAL το μοντέλο είχε χαμηλότερο MSE σε σύγκριση με το cross entropy. Αυτό μπορεί να συμβαίνει στο συγκεκριμένο dataset, γιατί οι κλάσεις που παρέχει μπορούμε να τις βάλουμε σε μια λογική σειρά και να τις δούμε σαν κλιμακωτές κατηγορίες και όχι σαν απλές κατηγορίες που είναι διακριτές μεταξύ τους. Με το CORAL, καμία από τις επιπλέον πληροφορίες δεν το βοήθησε να αυξήσει τις μετρικές του — πιθανώς για τον ίδιο λόγο που δεν βοήθησε ο συνδυασμός θεμάτων και συναισθημάτων με το cross entropy. Όπως παρουσιάστηκε και σε προηγούμενα κεφάλαια, οι κατηγορίες Half-true και Mostly-true μαζί περιέχονται στις περισσότερες εγγραφές του dataset. Αυτό αποτυπώνεται στο Σχήμα 4.6 έχοντας σε αυτές τις κατηγορίες το μικρότερο MSE, παρόλο που έχει χρησιμοποιηθεί διάνυσμα με βάρη, στην συνάρτηση σφάλματος, για την εξισορρόπηση αυτής της ανισότητας.

Ανάλογα με τα αποτελέσματα που θα θέλαμε να έχει το μοντέλο, θα πρέπει να επιλέξουμε ανάμεσα σε αυτό που είχε τις καλύτερες μετρικές με το cross entropy ή με το CORAL. Αν θέλαμε το μοντέλο να παρέχει πιο ακριβείς προβλέψεις αλλά σε αρκετές περιπτώσεις να

αστοχεί με μεγάλη απόκλιση την πραγματικότητα, τότε θα επιλέγαμε το cross entropy. Αν θέλαμε να παρέχει προβλέψεις που αρκετές φορές δεν είναι ακριβώς ο πραγματικός στόχος, αλλά κοντά σε αυτόν, τότε θα επιλέγαμε το CORAL. Βέβαια, όλες αυτές οι έρευνες και τα dataset που γίνονται είναι benchmark, αυτό σημαίνει ότι στην πραγματικότητα θα είναι ακόμα πιο δύσκολες οι προβλέψεις, γιατί δεν θα είναι εύκολο να συλλέξει κάποιος όλη αυτήν τη μεταπληροφορία που σχετίζεται με το κείμενο, όπως ποιος το είπε, πότε το είπε, πόσες άλλες φορές έχει πει ψέματα και άλλα.

Σε μελλοντική έρευνα θα μπορούσαμε να δοκιμάσουμε και άλλα dataset ή την επαύξηση του LAIR με επιπλέον εγγραφές και πληροφορίες. Στη συνέχεια, ένα μοντέλο όπως το BERT θα μπορούσε να αποδώσει καλύτερα από ό,τι μπορεί τώρα να κάνει στο LAIR και έτσι να πετύχουμε πολύ καλά αποτελέσματα. Θα άξιζε επίσης να δοκιμαστεί η εκπαίδευση μοντέλων που ειδικεύονται στην ανίχνευση συγκεκριμένης κατηγορίας ψεύδους και, στο τέλος, με ένα μοντέλο επιτροπής, να παράγεται η τελική πρόβλεψη.

Βιβλιογραφία

[1] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan και S. Liu, "Fake news on social media: The impact on society", *Information Systems Frontiers*, τόμ. 26, αρθμ. 2, σσ. 443–458, 1 Απρ. 2024, issn: 1572-9419. doi: 10.1007/s10796-022-10242-z. επίσκεψη 27 Απρ. 2025. διεύθν.: https://doi.org/10.1007/s10796-022-10242-z.

[2] W. Y. Wang, *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*, 1 Μάι. 2017. doi: 10.48550/arXiv.1705.00648. arXiv: 1705.00648. επίσκεψη 22 Οκτ. 2024. διεύθν.: http://arxiv.org/abs/1705.00648.

[3] H. Jelodar κ.ά., "Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey", *Multimedia Tools and Applications*, τόμ. 78, αρθμ. 11, σσ. 15 169–15 211, 1 Ιούν. 2019, issn: 1573-7721. doi: 10.1007/s11042-018-6894-4. επίσκεψη 27 Απρ. 2025. διεύθν.: https://doi.org/10.1007/s11042-018-6894-4.

[4] S. Hochreiter και J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, τόμ. 9, αρθμ. 8, σσ. 1735–1780, 15 Νοέ. 1997, issn: 0899-7667. doi: 10.1162/neco.1997.9.8.1735. επίσκεψη 27 Απρ. 2025. διεύθν.: https://doi.org/10.1162/neco.1997.9.8.1735.

[5] Y. LeCun, Y. Bengio και G. Hinton, "Deep learning", *Nature*, τόμ. 521, αρθμ. 7553, σσ. 436–444, Μάι. 2015, Publisher: Nature Publishing Group, issn: 1476-4687. doi: 10.1038/nature14539. επίσκεψη 27 Απρ. 2025. διεύθν.: https://www.nature.com/articles/nature14539.

[6] C. Hutto και E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text", *Proceedings of the International AAAI Conference on Web and Social Media*, τόμ. 8, αρθμ. 1, σσ. 216–225, 16 Μάι. 2014, Number: 1, issn: 2334-0770. doi: 10.1609/icwsm.v8i1.14550. επίσκεψη 27 Απρ. 2025. διεύθν.: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[7] Y. Liu κ.ά., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 26 Ιούλ. 2019. doi: 10.48550/arXiv.1907.11692. arXiv: 1907.11692[cs]. επίσκεψη 27 Απρ. 2025. διεύθν.: http://arxiv.org/abs/1907.11692.

[8] J. Devlin, M.-W. Chang, K. Lee και K. Toutanova, "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, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran και T. Solorio, επιμελητές, Minneapolis, Minnesota: Association for Computational Linguistics,

Ιούν. 2019, σσ. 4171–4186. doi: 10.18653/v1/N19-1423. επίσκεψη 27 Απρ. 2025. διεύθν.: https://aclanthology.org/N19-1423/.

[9] Y. Lecun, L. Bottou, Y. Bengio και P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, τόμ. 86, αρθμ. 11, σσ. 2278–2324, Νοέ. 1998, issn: 1558-2256. doi: 10.1109/5.726791. επίσκεψη 27 Απρ. 2025. διεύθν.: https://ieeexplore.ieee.org/document/726791.

[10] L. Hu, S. Wei, Z. Zhao και B. Wu, "Deep learning for fake news detection: A comprehensive survey", *AI Open*, τόμ. 3, σσ. 133–155, 1 Ιαν. 2022, issn: 2666-6510. doi: 10.1016/j.aiopen.2022.09.001. επίσκεψη 29 Απρ. 2025. διεύθν.: https://www.sciencedirect.com/science/article/pii/S2666651022000134.

[11] M. Elhadad, K. F. Li και F. Gebali, "Detecting misleading information on COVID-19", *IEEE Access*, τόμ. 8, σσ. 165 201–165 215, 9 Σεπτ. 2020. doi: 10.1109/ACCESS.2020.3022867.

[12] J. Su, C. Cardie και P. Nakov, *Adapting Fake News Detection to the Era of Large Language Models*, 13 Απρ. 2024. doi: 10.48550/arXiv.2311.04917. arXiv: 2311.04917[cs]. επίσκεψη 29 Απρ. 2025. διεύθν.: http://arxiv.org/abs/2311.04917.

[13] K. Shu, A. Sliva, S. Wang, J. Tang και H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective", *SIGKDD Explor. Newsl.*, τόμ. 19, αρθμ. 1, σσ. 22–36, 1 Σεπτ. 2017, issn: 1931-0145. doi: 10.1145/3137597.3137600. επίσκεψη 29 Απρ. 2025. διεύθν.: https://doi.org/10.1145/3137597.3137600.

[14] S. Raza, "Automatic Fake News Detection in Political Platforms - A Transformer-based Approach", στο *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, A. Hürriyetoğlu, επιμελητής, Online: Association for Computational Linguistics, Αύγ. 2021, σσ. 68–78. doi: 10.18653/v1/2021.case-1.10. επίσκεψη 29 Απρ. 2025. διεύθν.: https://aclanthology.org/2021.case-1.10/.

[15] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin και A. Iqbal, "A Benchmark Study of Machine Learning Models for Online Fake News Detection", *Machine Learning with Applications*, τόμ. 4, σ. 100 032, Ιούν. 2021, issn: 26668270. doi: 10.1016/j.mlwa.2021.100032. arXiv: 1905.04749[cs]. επίσκεψη 29 Απρ. 2025. διεύθν.: http://arxiv.org/abs/1905.04749.

[16] D. J. Sude, G. Sharon και S. Dvir-Gvirsman, "True, justified, belief? Partisanship weakens the positive effect of news media literacy on fake news detection", *Frontiers in Psychology*, τόμ. 14, σ. 1 242 865, 2023, issn: 1664-1078. doi: 10.3389/fpsyg.2023.1242865.

[17] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez και J. Vilares, "Sentiment analysis for fake news detection", *Electronics*, τόμ. 10, αρθμ. 11, σ. 1348, Ιαν. 2021, Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, issn: 2079-9292. doi: 10.3390/electronics10111348. επίσκεψη 2 Μάι. 2025. διεύθν.: https://www.mdpi.com/2079-9292/10/11/1348.

[18] B. Bhutani, N. Rastogi, P. Sehgal και A. Purwar, "Fake News Detection Using Sentiment Analysis", στο *2019 Twelfth International Conference on Contemporary Computing (IC3)*, ISSN: 2572-6129, Αύγ. 2019, σσ. 1–5. doi: 10.1109/IC3.2019.8844880. επίσκεψη 2 Μάι. 2025. διεύθν.: https://ieeexplore.ieee.org/abstract/document/8844880.

[19] I. G. S. Mas Diyasa, N. M. I. Marini Mandenni, M. I. Fachrurrozi, S. I. Pradika, K. R. Nur Manab και N. R. Sasmita, "Twitter sentiment analysis as an evaluation and service base on python textblob", *IOP Conference Series: Materials Science and Engineering*, τόμ. 1125, αρθμ. 1, σ. 012 034, Μάι. 2021, Publisher: IOP Publishing, issn: 1757-899X. doi: 10.1088/1757-899X/1125/1/012034. επίσκεψη 2 Μάι. 2025. διεύθν.: https://dx.doi.org/10.1088/1757-899X/1125/1/012034.

[20] Y. Li, X. Li, Y. Zhai, D. Wang και C. Hon, "Misinformation Features Detection in Weibo: Unsupervised Learning, Latent Dirichlet Allocation, and Network Structure", *IEEE Access*, τόμ. 12, σσ. 166 977–166 987, 2024, issn: 2169-3536. doi: 10.1109/ACCESS.2024.3494015. επίσκεψη 3 Μάι. 2025. διεύθν.: https://ieeexplore.ieee.org/abstract/document/10747342.

[21] M. Hosseini, A. Javadian Sabet, S. He και D. Aguiar, "Interpretable fake news detection with topic and deep variational models", *Online Social Networks and Media*, τόμ. 36, σ. 100 249, 1 Ιούλ. 2023, issn: 2468-6964. doi: 10.1016/j.osnem.2023.100249. επίσκεψη 3 Μάι. 2025. διεύθν.: https://www.sciencedirect.com/science/article/pii/S2468696423000083.

[22] T. Mikolov, K. Chen, G. Corrado και J. Dean, *Efficient Estimation of Word Representations in Vector Space*, 7 Σεπτ. 2013. doi: 10.48550/arXiv.1301.3781. arXiv: 1301.3781[cs]. επίσκεψη 3 Μάι. 2025. διεύθν.: http://arxiv.org/abs/1301.3781.

[23] Y. Bengio, R. Ducharme, P. Vincent και C. Jauvin, "A neural probabilistic language model",

[24] H. Thakar και B. Bhatt, *Fine-Tuning RoBERTa for Truthfulness Classification: A Case Study on the LIAR Dataset*. 11 Ιούλ. 2024. doi: 10.21203/rs.3.rs-4721974/v1.

[25] B. Upadhayay και V. Behzadan, *Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification*, 22 Οκτ. 2020. doi: 10.48550/arXiv.2009.01047. arXiv: 2009.01047[cs]. επίσκεψη 2 Ιούν. 2025. διεύθν.: http://arxiv.org/abs/2009.01047.

[26] C. Whitehouse, T. Weyde, P. Madhyastha και N. Komninos, *Evaluation of Fake News Detection with Knowledge-Enhanced Language Models*, 13 Φεβ. 2023. doi: 10.48550/arXiv.2204.00458. arXiv: 2204.00458[cs]. επίσκεψη 2 Ιούν. 2025. διεύθν.: http://arxiv.org/abs/2204.00458.

[27] P. Patwa κ.ά., "Fighting an Infodemic: COVID-19 Fake News Dataset", στο τόμ. 1402, 2021, σσ. 21–29. doi: 10.1007/978-3-030-73696-5_3. arXiv: 2011.03327[cs]. επίσκεψη 2 Ιούν. 2025. διεύθν.: http://arxiv.org/abs/2011.03327.

[28] N. Aslam, I. Ullah Khan, F. S. Alotaibi, L. A. Aldaej και A. K. Aldubaikil, "Fake detect: A deep learning ensemble model for fake news detection", *Complexity*, τόμ. 2021, αρθμ. 1, σ. 5 557 784, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/5557784, issn: 1099-0526. doi: 10.1155/2021/5557784. επίσκεψη 30 Μαρ. 2025. διεύθν.: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5557784.

[29] A. M. Ali, F. A. Ghaleb, M. S. Mohammed, F. J. Alsolami και A. I. Khan, "Web-informed-augmented fake news detection model using stacked layers of convolutional neural network and deep autoencoder", *Mathematics*, τόμ. 11, αρθμ. 9, σ. 1992, Ιαν. 2023, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, issn: 2227-7390. doi: 10.3390/math11091992. επίσκεψη 3 Μάι. 2025. διεύθν.: https://www.mdpi.com/2227-7390/11/9/1992.

[30] "(PDF) glove: Global vectors for word representation", στο *ResearchGate*. doi: 10.3115/v1/D14-1162. επίσκεψη 2 Ιούν. 2025. διεύθν.: https://www.researchgate.net/publication/284576917_Glove_Global_Vectors_for_Word_Representation.

[31] M. Röder, A. Both και A. Hinneburg, "Exploring the Space of Topic Coherence Measures", στο *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, σειρά WSDM '15, New York, NY, USA: Association for Computing Machinery, 2 Φεβ. 2015, σσ. 399–408, isbn: 978-1-4503-3317-7. doi: 10.1145/2684822.2685324. επίσκεψη 21 Μάι. 2025. διεύθν.: https://doi.org/10.1145/2684822.2685324.

[32] D. P. Kingma και J. Ba, *Adam: A Method for Stochastic Optimization*, 30 Ιαν. 2017. doi: 10.48550/arXiv.1412.6980. arXiv: 1412.6980[cs]. επίσκεψη 24 Μάι. 2025. διεύθν.: http://arxiv.org/abs/1412.6980.

[33] W. Cao, V. Mirjalili και S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation", *Pattern Recognition Letters*, τόμ. 140, σσ. 325–331, 1 Δεκ. 2020, issn: 0167-8655. doi: 10.1016/j.patrec.2020.11.008. επίσκεψη 25 Μάι. 2025. διεύθν.: https://www.sciencedirect.com/science/article/pii/S016786552030413X.