ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Επεξεργασίας Φυσικής Γλώσσας για Εντοπισμό και Αποφυγή Ψευδών Ειδήσεων στα Μέσα Κοινωνικής Δικτύωσης

Θεόδωρος – Αλέξανδρος Α. Χανδρινός

Θεόδωρος Ε. Ζαμπάτης

ΕΠΙΒΛΕΠΟΥΣΑ: **Δρ. Τσαλγατίδου Αφροδίτη**, Αναπληρώτρια Καθηγήτρια

ΤΕΧΝΙΚΗ ΥΠΟΣΤΗΡΙΞΗ: **Δρ. Ελένη Κουτρούλη**, Μεταδιδακτορική Ερευνήτρια
Κα. Μαρίζα Κατικαρίδη, Υποψ. Διδάκτωρ

ΑΘΗΝΑ

ΙΑΝΟΥΑΡΙΟΣ 2021

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Επεξεργασίας Φυσικής Γλώσσας για Εντοπισμό και Αποφυγή Ψευδών Ειδήσεων στα Μέσα Κοινωνικής Δικτύωσης

Θεόδωρος – Αλέξανδρος Α. Χανδρινός
Α.Μ.: 1115201300198

Θεόδωρος Ε. Ζαμπάτης
Α.Μ.: 1115201300045

ΕΠΙΒΛΕΠΟΥΣΑ: **Δρ. Τσαλγατίδου Αφροδίτη**, Αναπληρώτρια Καθηγήτρια

ΤΕΧΝΙΚΗ ΥΠΟΣΤΗΡΙΞΗ: **Δρ. Ελένη Κουτρούλη**, Μεταδιδακτορική Ερευνήτρια
Κα. Μαρίζα Κατικαρίδη, Υποψ. Διδάκτωρ

ΠΕΡΙΛΗΨΗ

Στην εργασία μας, διερευνούμε την ανίχνευση ψευδών tweets στο Twitter χρησιμοποιώντας την επεξεργασία φυσικής γλώσσας (NLP) με τη γλώσσα προγραμματισμού Python μέσω της εποπτευόμενης μηχανικής μάθησης. Μελετήσαμε μια ποικιλία προσεγγίσεων στο θέμα από διάφορες πηγές και συγγραφείς. Αυτό μας ενέπνευσε να συνδυάσουμε αυτές τις προσεγγίσεις με στόχο να μάθουμε ποιοι συνδυασμοί λειτουργούν καλύτερα.

Για αυτό τον σκοπό, έχουμε αναπτύξει ένα εργαλείο λογισμικού, το οποίο ελέγχει το ποσοστό επιτυχίας τεσσάρων (4) διαφορετικών συστημάτων για την ανίχνευση ψευδών ειδήσεων χρησιμοποιώντας τέσσερα (4) διαφορετικά σύνολα δεδομένων, με αποτέλεσμα συνολικά δεκαέξι (16) ποσοστά επιτυχίας, ένα για κάθε συνδυασμό.

Για τη δημιουργία του παραπάνω εργαλείου, χρησιμοποιήσαμε το σύνολο δεδομένων PHEME [15], το οποίο περιλαμβάνει χιλιάδες πραγματικά προ-επεξεργασμένα tweets με ετικέτα που εξάγονται μέσω του TweeterAPI [16]. Δημιουργήσαμε ένα πρόγραμμα python, το οποίο αναλύει το προαναφερθέν σύνολο δεδομένων και αποθηκεύει όλα τα tweets από αυτό σε αρχεία της μορφής .tsv. Έχουμε τέσσερα (4) διαφορετικά σύνολα δεδομένων που διαφοροποιούνται βάσει των ακόλουθων χαρακτηριστικών:

1. Πρέπει να αποδεχτούμε την ύπαρξη διπλών tweet: Μερικά από τα ίδια tweets μπορεί να έχουν κοινοποιηθεί από διάφορους χρήστες / προφίλ.
2. Πρέπει να αποδεχτούμε μια τρίτη ετικέτα για την εγκυρότητα των tweets εκτός από το "true" (αληθής είδηση) ή "false" (ψευδής είδηση) , το οποίο είναι το "undefined" (είδηση απροσδιόριστης εγκυρότητας).

Αφού επιλεγεί ένα αρχείο .tsv, πραγματοποιείται η ανάλυση συναισθήματος σε κάθε tweet με τη χρήση του αλγορίθμου Sentiment Intensity Analyzer [17]. Στη συνέχεια, επεξεργάζονται τα αποτελέσματα αυτής της ανάλυσης και αποφασίζεται εάν ένα tweet θα πρέπει να επισημαίνεται ως θετικό ή αρνητικό.

Στη συνέχεια, χρησιμοποιούμε ένα pipeline στο οποίο εκτελούνται κατά σειρά τα ακόλουθα βήματα:

1. Αναγνώριση λεξικών μονάδων (Tokenization) και λημματοποίηση (Lemmatization) σε αναπαράσταση σάρωσης λέξεων (bag of words) χρησιμοποιώντας NLTK
2. Διανυσματοποίηση (Vectorization) χρησιμοποιώντας τον απαριθμητή διανυσμάτων (Count Vectorizer) ή διανυσματοποιητή συχνότητας όρου – άνισης κατανομής του όρου (TF-IDF Vectorizer) μέσω της βιβλιοθήκης Scikit-Learn.
3. Ταξινόμηση (Classification) χρησιμοποιώντας Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machine) ή τον Πολυωνυμικό Απλοϊκό Ταξινομητή Bayes (Multinomial Naive Bayes) μέσω της βιβλιοθήκης Scikit-Learn.

Ο συνδυασμός της γλωσσικής και της συναισθηματικής επεξεργασίας εξάγει διαφορετικά αποτελέσματα με βάση την επιλογή του αρχείου, του διανυσματοποιητή και του ταξινομητή. Το ποσοστό επιτυχίας των σωστά επισημασμένων με ετικέτα δεδομένων κυμαίνεται μεταξύ 54,6% και 99,8%.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εντοπισμός Ψευδών Ειδήσεων με Τεχνικές NLP

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ψευδείς Ειδήσεις, Επεξεργασία Φυσικής Γλώσσας, Αλγόριθμος, Twitter, Python, Εποπτευόμενη Μηχανική Μάθηση

ABSTRACT

The advancement of social networks has facilitated the sharing and spread of news among people all over the world. With the growth of these networks and of the volume of news shared daily, the phenomena of fake news have become stronger and widely spread. Over the past few years, big social networks like Twitter admit that fake and duplicate accounts, fake news and fake likes exist in their networks. This stems from the fact that the social network account owners have the ability to distribute false information, to support or attack an idea or a product, to promote or demote an election candidate, as well as to influence real network users in their decision making. Therefore, misinformation detection in enhancing public trust and society stability becomes of critical importance. Along these lines, detection of misinformation is still a challenging problem for the Natural Language Processing community.

In our work, we have utilized natural language processing and supervised machine learning in order to detect fake tweets using Python. We have studied a variety of approaches on the subject from various sources and authors. This inspired us to combine these approaches with the goal to find out which combinations work better. Therefore, we have developed a software tool, which checks the success ratio of four (4) different systems for fake news detection using four (4) different datasets, resulting in a total of sixteen (16) ratios, one for each combination.

SUBJECT AREA: Natural Language Processing

KEYWORDS: Fake News, Twitter, Algorithm, Machine Learning, Classifiers, Natural Language Processing, Vectorizers

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Πτυχιακής Εργασίας, θα θέλαμε να ευχαριστήσουμε την επιβλέπουσα καθηγήτρια κα. Τσαλγατίδου Αφροδίτη, καθώς και τις συνεργάτιδές της Δρ. Ελένη Κουτρούλη και κα. Μαρίζα Κατικαρίδη, Υπ. Δρ. για τη συνεργασία και την πολύτιμη συμβολή τους στην ολοκλήρωσή της.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

Τα μέσα κοινωνικής δικτύωσης διαδραματίζουν ολοένα και αυξανόμενο παρεμβατικό ρόλο στην καθημερινότητά μας και κατά συνέπεια η αξιολόγηση των ειδήσεων που προβάλλουν αποκτά όλο και μεγαλύτερη σημασία.

Οι ψευδείς ειδήσεις (fake news) είναι ιστορίες οι οποίες παρουσιάζονται κυρίως ως δημοσιογραφικές, είναι όμως κατασκευασμένες εσκεμμένα για να εξυπηρετήσουν κάποιο σκοπό. Οι ψευδείς ειδήσεις μπορούν να περιέχονται σε ψηφιακό ή έντυπο περιεχόμενο. Επίσης, μπορεί να λάβουν τη μορφή ολόκληρων ιστοσελίδων που έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να μοιάζουν με αξιόπιστα ειδησεογραφικά sites. Ο σκοπός των “ψευδών ειδήσεων” είναι είτε εμπορικός όπως η προώθηση ενός προϊόντος ή η δημιουργία κίνησης (traffic) προς μία ιστοσελίδα (clickbait), είτε πολιτικός (παραπληροφόρηση-διαμόρφωση κοινής γνώμης).

Σήμερα, στην εποχή της τεχνολογίας, η χρήση των κοινωνικών δικτύων ως πηγή κατανάλωσης ειδήσεων και άλλων πληροφοριών είναι ένα δίκοπο σπαθί. Από τη μία πλευρά, οι άνθρωποι μπορούν να προσεγγίζουν ειδήσεις στα κοινωνικά δίκτυα εύκολα, φθηνότερα από τα παραδοσιακά μέσα ενημέρωσης και μπορούν να διαδώσουν τις πληροφορίες τους πιο γρήγορα, ενώ από την άλλη, τα κοινωνικά δίκτυα παρέχουν ένα τέλειο περιβάλλον για τη δημιουργία και διάδοση ψευδών ειδήσεων. Σημαντικό πρόβλημα αποτελεί επίσης η συχνότατη απουσία ελέγχου αξιοπιστίας των ειδήσεων από το μεγαλύτερο ποσοστό των ανθρώπων. Αυτό με τη σειρά του προκαλεί πολλά προβλήματα για άτομα, οργανισμούς, ακόμη και κυβερνήσεις. Οι ψεύτικες ειδήσεις μπορούν να οδηγήσουν σε μεγάλες (αρνητικές) αλλαγές στην κοινωνία, ψευδείς σκέψεις και απόψεις, συλλογική υστερία ή άλλες σοβαρές συνέπειες. Χαρακτηριστικό παράδειγμα είναι αυτό που συνέβη στην Αραβία κατά τη διάρκεια της “Αραβικής άνοιξης” [18],όπου τα ψεύτικα νέα που διαδόθηκαν στα κοινωνικά δίκτυα χρησιμοποιήθηκαν ως όπλο ψυχολογικού πολέμου εναντίον ατόμων και κυβερνήσεων.

Για να βοηθήσουν στην αντιμετώπιση αυτού του ζητήματος, ερευνητές και ειδικοί στα μέσα ενημέρωσης πρότειναν ανιχνευτές ψευδών ειδήσεων που χρησιμοποιούν την επεξεργασία φυσικής γλώσσας (NLP) για την ανάλυση προτύπων λέξεων και στατιστικών συσχετίσεων ειδησεογραφικών άρθρων.

Στόχος αυτής της εργασίας είναι:

- Η μελέτη και η συγκριτική παρουσίαση μεθόδων βασισμένων στην επεξεργασία φυσικής γλώσσας (NLP) για τον εντοπισμό ψευδών ειδήσεων
- Η αξιολόγηση της αξιοπιστίας του περιεχομένου των μέσων κοινωνικής δικτύωσης και ο σχεδιασμός και υλοποίηση μεθόδων που βασίζονται σε επεξεργασία φυσικής γλώσσας (NLP) για τον εντοπισμό ψευδών ειδήσεων.

Στην ενότητα 2 θα εξηγήσουμε κάποιες βασικές έννοιες που σχετίζονται με τις ψευδείς ειδήσεις καθώς και τις τεχνικές επεξεργασίας φυσικής γλώσσας. Στην ενότητα 3 παρουσιάζουμε δέκα (10) διαφορετικά συστήματα εντοπισμού ψευδών ειδήσεων όπου γίνεται χρήση τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) και τέλος στην ενότητα 4 παρουσιάζουμε ένα σύστημα που υλοποιήσαμε για εντοπισμό ψευδών ειδήσεων με χρήση τεχνικών επεξεργασίας φυσικής γλώσσας που ενσωματώνουν τεχνικής ανάλυσης συναισθήματος.

2. ΠΑΡΑΠΛΗΡΟΦΟΡΗΣΗ ΚΑΙ ΤΕΧΝΙΚΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ - ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Τι είναι η επεξεργασία φυσικής γλώσσας;

Η επεξεργασία φυσικής γλώσσας (NLP) . ξεκίνησε τη δεκαετία του 1950 ως τομή της τεχνητής νοημοσύνης και της γλωσσολογίας. Ορίζεται ευρέως ως ο αυτόματος χειρισμός φυσικής γλώσσας, όπως η ομιλία ή το κείμενο, από λογισμικό. Η NLP επιτρέπει στους υπολογιστές να κατανοούν τη φυσική γλώσσα όπως οι άνθρωποι. Είτε η γλώσσα μιλιέται είτε γράφεται, η επεξεργασία φυσικής γλώσσας χρησιμοποιεί τεχνητή νοημοσύνη για να λάβει πληροφορίες από τον πραγματικό κόσμο, να τις επεξεργαστεί και να τις κατανοήσει με τρόπο που μπορεί να καταλάβει ένας υπολογιστής.

Γιατί είναι σημαντική η επεξεργασία φυσικής γλώσσας;

Η επεξεργασία φυσικής γλώσσας βοηθά τους υπολογιστές να επικοινωνούν με τους ανθρώπους στη γλώσσα τους και χρησιμοποιείται και σε άλλες εργασίες που σχετίζονται με τη γλώσσα. Για παράδειγμα, η NLP επιτρέπει στους υπολογιστές να διαβάζουν κείμενα, να ακούν ομιλίες, να τις ερμηνεύουν, να μετρούν το συναίσθημα και να καθορίζουν ποια μέρη είναι σημαντικά και ποιά όχι.

Οι σημερινές μηχανές μπορούν να αναλύσουν περισσότερα δεδομένα που βασίζονται στη γλώσσα από τους ανθρώπους, χωρίς κόπο, με συνέπεια και αμερόληπτο τρόπο. Λαμβάνοντας υπόψη την εκπληκτική μεγάλη ποσότητα μη δομημένων δεδομένων που παράγονται καθημερινά, από ιατρικά αρχεία έως τα μέσα κοινωνικής δικτύωσης, η αυτοματοποιημένη διαδικασία θα είναι κρίσιμος παράγοντας για την πλήρη αποτελεσματική ανάλυση δεδομένων κειμένου και ομιλίας.

Πώς λειτουργεί η επεξεργασία φυσικής γλώσσας;

Υπάρχουν δύο κύριες φάσεις στην επεξεργασία φυσικής γλώσσας: η προεπεξεργασία δεδομένων και η ανάπτυξη αλγορίθμου.

Η προεπεξεργασία δεδομένων περιλαμβάνει την προετοιμασία και τον "καθαρισμό" των δεδομένων του κειμένου για να μπορούν τα μηχανήματα να τα αναλύσουν. Η προεπεξεργασία θέτει δεδομένα σε λειτουργική μορφή και επισημαίνει χαρακτηριστικά στο κείμενο με τα οποία μπορεί να λειτουργήσει ένας αλγόριθμος. Υπάρχουν διάφοροι τρόποι που μπορεί να γίνουν, μεταξύ των οποίων:

- Αναγνώριση Λεξικών Μονάδων (Tokenization). Αυτό συμβαίνει όταν το κείμενο αναλύεται σε μικρότερες μονάδες (λέξεις, σημεία στίξης) για επεξεργασία. (βλ. Ενότητα 4.4 – Περιγραφή Όρων)
- Λεμματοποίηση (Lemmatization). Αυτό συμβαίνει όταν οι λέξεις επαναφέρονται στις ρίζες τους για καλύτερη επεξεργασία. (βλ. Ενότητα 4.4 – Περιγραφή Όρων)
- Ανάλυση Συναισθήματος. Κατά της διαδικασία αυτή επεξεργαζόμαστε το κείμενο αναλύοντας τις φράσεις, τις λέξεις ή τα σημεία στίξης που χρησιμοποιεί ο συγγραφέας και καταλήγουμε στον βαθμό συναισθηματικής φόρτισης του κειμένου. (βλ. Ενότητα 4.4 – Περιγραφή Όρων)

Μετά την προεπεξεργασία των δεδομένων, αναπτύσσεται ένας αλγόριθμος για την επεξεργασία τους. Υπάρχουν πολλοί διαφορετικοί αλγόριθμοι επεξεργασίας φυσικής γλώσσας, αλλά συνήθως χρησιμοποιούνται δύο κύριοι τύποι:

- Σύστημα βασισμένο σε κανόνες. Αυτό το σύστημα χρησιμοποιεί προσεκτικά σχεδιασμένους γλωσσικούς κανόνες. Αυτή η προσέγγιση χρησιμοποιήθηκε νωρίς στην ανάπτυξη της επεξεργασίας φυσικής γλώσσας και εξακολουθεί να χρησιμοποιείται.
- Σύστημα βασισμένο στη μηχανική μάθηση. Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν στατιστικές μεθόδους. Μαθαίνουν να εκτελούν εργασίες με βάση τα δεδομένα εκπαίδευσης που τους δίνονται και προσαρμόζουν τις μεθόδους τους καθώς επεξεργάζονται περισσότερα δεδομένα. Χρησιμοποιώντας έναν συνδυασμό μηχανικής μάθησης και νευρωνικών δικτύων, οι αλγόριθμοι επεξεργασίας φυσικής γλώσσας βελτιώνουν τους δικούς τους κανόνες μέσω της επαναλαμβανόμενης επεξεργασίας και εκμάθησης..

Η NLP είναι το μέλλον

Με την αυξανόμενη ποσότητα δεδομένων κειμένου που δημιουργείται κάθε μέρα, η NLP θα γίνεται όλο και πιο σημαντική για να αποκτούν νόημα τα δεδομένα και να χρησιμοποιούνται σε πολλές άλλες εφαρμογές.

Πιθανότατα έχετε ήδη χρησιμοποιήσει μερικές από τις πιο ισχυρές εφαρμογές NLP αλλά δεν το γνωρίζετε. Για παράδειγμα η Μετάφραση της Google[68] ή τα chatbots[69].

Η NLP άλλαξε τον τρόπο που αλληλεπιδρούμε με τους υπολογιστές και θα συνεχίσει να το κάνει στο μέλλον. Αυτές οι τεχνολογίες τεχνητής νοημοσύνης θα αποτελέσουν τη βασική δύναμη για τις προσπάθειες μετασχηματισμού των δεδομένα σε υπηρεσίες που βασίζονται στη νοημοσύνη, και έτσι συνεχώς θα διαμορφώνουν και θα βελτιώνουν την τεχνολογία επικοινωνίας.

Βασικές έννοιες

Η **παραπληροφόρηση (misinformation)** αναφέρεται στην εσφαλμένη πληροφόρηση, συμπεριλαμβανομένων όλων των κατασκευασμένων, παραπλανητικών, λανθασμένων, ψευδών, απατηλών ή παραμορφωμένων πληροφοριών. Συνήθως δημιουργούνται με κακόβουλη πρόθεση από δημιουργούς πληροφοριών. Ως εκ τούτου, η αξιοπιστία των πληροφοριών συνήθως υπονομεύεται. Κάτω από την κοινή ομπρέλα της μεταφοράς παραποιημένων πληροφοριών, περιέχονται πολλές παρόμοιες έννοιες, όπως ψεύτικες ειδήσεις (fake news), φήμες (rumors), εξαπάτηση (deception), απάτες (hoaxes), ανεπιθύμητη γνώμη (spam opinion) κ.λπ. Παρά το γεγονός ότι εν πρώτοις φαίνονται -αν όχι ίδιες- παρόμοιες έννοιες, υπάρχουν σημαντικές διαφορές μεταξύ τους ως προς το μέγεθος του λάθους, τη χρήση του περιεχομένου ή τη λειτουργία που αυτό εξυπηρετεί.

Μεταξύ των διαφόρων τύπων παραπληροφόρησης, η **εξαπάτηση (deception)** ορίζεται γενικά η ως εκ προθέσεως παραπλανητική δήλωση. Μια συμπεριφορά εξαπάτησης συνήθως δείχνει τα ακόλουθα δύο χαρακτηριστικά: (α) ο εξαπατητής μεταδίδει ένα ψεύτικο μήνυμα (ενώ κρύβει αληθινές πληροφορίες) και (β) η πράξη είναι σκόπιμη. Σημειώνεται ότι η ακούσια συμπεριφορά που οδηγεί σε αναληθή πεποίθηση, όπως ειλικρινά λάθη ή λανθασμένες αναμνήσεις, δεν θεωρείται εξαπάτηση.

Οι **ψευδείς ειδήσεις (fake news)**, διαφοροποιούνται από το περιεχόμενο που μιμείται τα μέσα ενημέρωσης σε μορφή αλλά όχι σε διαδικασίες σύνταξης. Αυτός ο ορισμός υπογραμμίζει δύο κύρια χαρακτηριστικά των ψεύτικων ειδήσεων: το ψευδές περιεχόμενο των ειδήσεων και την έλλειψη συντακτικών κανόνων και διαδικασιών για τον έλεγχο αξιοπιστίας. Τα ψεύτικα νέα μπορεί να είναι παραπλανητικά ή ακόμη και επιβλαβή, ειδικά όταν αποσυνδέονται από τις αρχικές πηγές και τα περιβάλλοντά τους.

Τα **ανεπιθύμητα μηνύματα (spam opinion)**, που ονομάζονται επίσης ανεπιθύμητα σχόλια, είναι κατασκευασμένες κριτικές που κυμαίνονται από τις αυτοπροωθήσεις εως τις ψευδείς ανακοινώσεις του προϊόντος που αξιολογούν, με σκοπό να παρασύρουν σκόπιμα τους καταναλωτές να αγοράσουν ή να αποφύγουν το προϊόν. Υπάρχουν δύο παραλλαγές: το υπερ spam, όπου δίνονται θετικά σχόλια σε προϊόντα με σκοπό την παραπλανητική προώθησή τους και το δυσφημιστικό spam, το οποίο δίνει αδικαιολόγητα αρνητικά σχόλια σε ανταγωνιστικά προϊόντα, προκειμένου να βλάψουν τη φήμη τους στους καταναλωτές.

Μια **φήμη (rumor)** ορίζεται ως ένα κομμάτι πληροφοριών που κυκλοφορούν, των οποίων η κατάσταση ακρίβειας δεν έχει ακόμη επαληθευτεί κατά τη στιγμή της διάδοσης. Η λειτουργία μιας φήμης είναι να κατανοήσει μια διφορούμενη κατάσταση και η αξία της αλήθειας που προωθεί θα μπορούσε να είναι αληθινή, ψευδής ή μη επαληθευμένη. Διαφέρει από τις ψεύτικες ειδήσεις, οι οποίες συνήθως αναφέρονται σε δημόσια γεγονότα που μπορούν να επαληθευτούν ως αληθή ή ψευδή, ενώ οι φήμες μπορεί να περιλαμβάνουν μακροπρόθεσμες φήμες (όπως θεωρίες συνωμοσίας), ή/και βραχυπρόθεσμες αναδυόμενες φήμες.

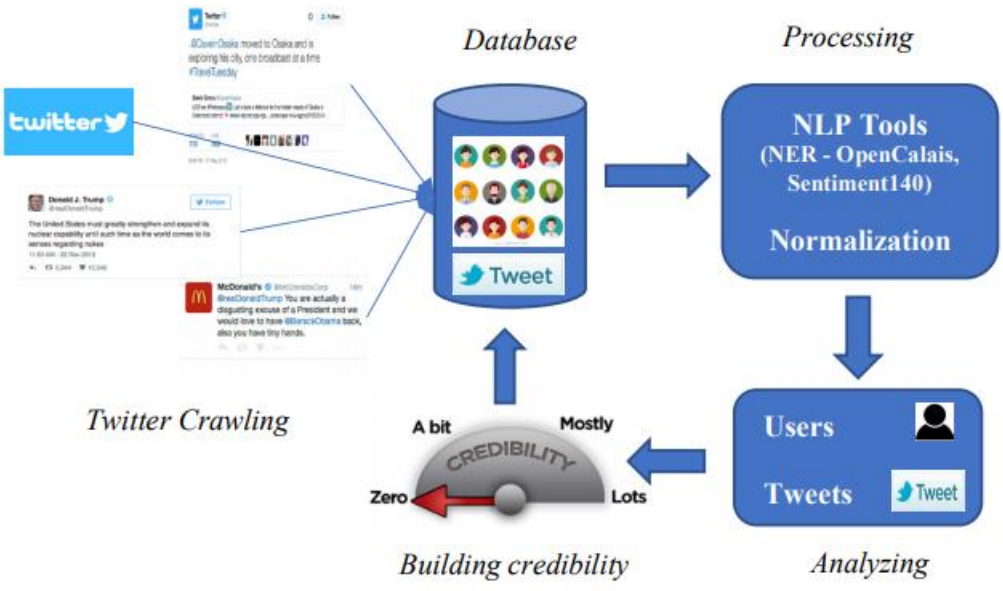
3. ΣΥΣΤΗΜΑΤΑ ΕΝΤΟΠΙΣΜΟΥ ΨΕΥΔΩΝ ΕΙΔΗΣΕΩΝ ΣΤΑ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Στην ενότητα αυτή παρουσιάζουμε δέκα (10) διαφορετικά συστήματα εντοπισμού ψευδών ειδήσεων όπου γίνεται χρήση τεχνικών επεξεργασίας φυσικής γλώσσας (NLP), με σκοπό να παρουσιαστεί η υφιστάμενη κατάσταση (state-of-the-art) των τεχνολογιών και των τεχνικών που χρησιμοποιούνται ευρέως και παρουσιάζουν τα καλύτερα αποτελέσματα. Οι δημοφιλέστερες από τις παρακάτω τεχνικές έχουν χρησιμοποιηθεί στο δικό μας σύστημα, έτσι ώστε να βρεθούν οι αποτελεσματικότεροι συνδυασμοί τους υπό τις ανάλογες προϋποθέσεις όπως αναλύονται στην ενότητα 4.

3.1 Εντοπισμός ψευδών ειδήσεων και ψεύτικων χρηστών στο Twitter (Atodiresei et al. 2018 [55])

Σε αυτή τη μελέτη προτείνεται ένα μοντέλο ανίχνευσης ψευδών ειδήσεων στο κοινωνικό δίκτυο του Twitter. Η εφαρμογή που δημιουργήθηκε λαμβάνει μία δημοσίευση στο Twitter και αναλύει την εγκυρότητά της μέσω της σύγκρισής της με άλλες αξιόπιστες πηγές και βάσει της μέχρι στιγμής αξιοπιστίας, του χρήστη που έκανε τη δημοσίευση. Συμπληρωματικά ελέγχει επίσης και άλλες στατιστικές, όπως το συναίσθημα της δημοσίευσης, τα emojis, τα hashtags κ.α.

Η αρχιτεκτονική του συστήματος παρουσιάζεται στη παρακάτω εικόνα. Υπάρχει, αρχικά, ένας Twitter Ανιχνευτής (Twitter crawler), ο οποίος συλλέγει δημοσιεύσεις και τις προσθέτει στην βάση δεδομένων (database). Η βάση δεδομένων επίσης περιέχει και τις δημοσιεύσεις οι οποίες είναι αξιόπιστες, τις οποίες θα τις χρησιμοποιήσουμε αργότερα στη σύγκριση. Μετά, έχουμε τη Μονάδα Επεξεργασίας (Processing module): όταν ένας χρήστης θέλει να ελέγξει την εγκυρότητα μίας δημοσίευσης, εισάγει έναν σύνδεσμο (link) στην πλατφόρμα. Στη συνέχεια ο αλγόριθμος χρησιμοποιεί μία μονάδα Αναγνώρισης Ονοματικών Οντοτήτων (Named Entity Recognition [NER]), η οποία διαχωρίζει το κείμενο στα τμήματα που το αποτελούν: αφαιρεί τι οντότητες (γενικά, τα ουσιαστικά και τη σχετική βαρύτητα που έχουν ως προς το περιεχόμενο), τα θέματα, τα σύμβολα των μέσων κοινωνικής δικτύωσης, το συνολικό συναίσθημα του περιεχομένου της δημοσίευσης και το συναίσθημα του hashtag.



Σχήμα 1: Αρχιτεκτονική Συστήματος

Για την ανάλυση που εκτελεί η μονάδα NER, χρησιμοποιήθηκε μία δημόσια προγραμματιστική διεπαφή εφαρμογής (Application Programming Interface – API) που ονομάζεται OpenCalais [1], ενώ για τους υπολογισμούς του συναισθήματος και των hashtags χρησιμοποιήθηκε το Sentiment140 [2].

Μετέπειτα φθάνουμε στην μονάδα Ανάλυσης (Analyzing module) όπου: η δημοσίευση που έχει εισαχθεί συγκρίνεται με αξιόπιστες πηγές που έχουν αποθηκευθεί τοπικά στη βάση δεδομένων του συστήματος. Παρακάτω, συναντούμε την Μονάδα Δημιουργίας Αξιοπιστίας (Building credibility module): εδώ, αναζητούμε στην βάση αξιόπιστων δεδομένων, δημοσιεύσεις οι οποίες είναι κοινού περιεχομένου με την εξεταζόμενη. Αν βρεθούν παρόμοιες δημοσιεύσεις, ο συνολικός βαθμός αξιοπιστίας της αυξάνεται, άρα αποκτά μεγαλύτερη πιθανότητα αξιοπιστίας για το σύστημα.

Ταυτόχρονα ελέγχεται και η αξιοπιστία του χρήστη μέσω ενός δεύτερου συστήματος βαθμολόγησης. Αρχικά, ο χρήστης ξεκινά με βαθμό αξιοπιστίας το 0· καθότι θεωρείται αναξιόπιστος. Ωστόσο, όσες πιο πολλές δημοσιεύσεις κάνει, για τις οποίες βρίσκονται πανομοιότυπες και αξιόλογες στο σύστημα, τόσο πιο πολύ αυτός ο ατομικός βαθμός αυξάνεται.

Η εφαρμογή εξάγει ως αποτελέσματα τον ατομικό βαθμό, τον βαθμό της δημοσίευσης και ένα μήνυμα το οποίο περιγράφει τη δημοσίευση ως αληθή, ψευδή ή ότι αδυνατεί να την επαληθεύσει.

Σχετικά με της λεπτομέρειες ανάπτυξης της εφαρμογής, ακολουθήθηκε μία υπηρεσιοστραφής αρχιτεκτονική (service-oriented architecture), όπου δημιουργήθηκαν διαφορετικές οντότητες για να αλληλεπιδρούν και να παράγουν ένα βελτιωμένο αποτέλεσμα. Η κύρια προγραμματιστική γλώσσα που χρησιμοποιήθηκε ήταν η Java. Το σύστημα χρησιμοποιεί τη NoSQL, βάση δεδομένων, MongoDb για να αποθηκεύει και να οργανώνει τις πληροφορίες χωρίζοντάς τες σε δύο κατηγορίες: τις δημοσιεύσεις και τους χρήστες.

Η οντότητα που χειρίζεται το χώρο αποθήκευσης δημιουργήθηκε με την ασύγχρονη μηχανή Node.js, που ενεργοποιείται μόλις συμβεί ένα γεγονός. Έτσι όταν η εφαρμογή κάνει μία αίτηση, δεν θα παγώσει το σύστημα μέχρι να ληφθεί η απάντηση, αλλά θα συνεχίσει να εκτελεί άλλες χρήσιμες διεργασίες, δίνοντας έτσι τη δυνατότητα ταυτόχρονης και παράλληλης εξυπηρέτησης.

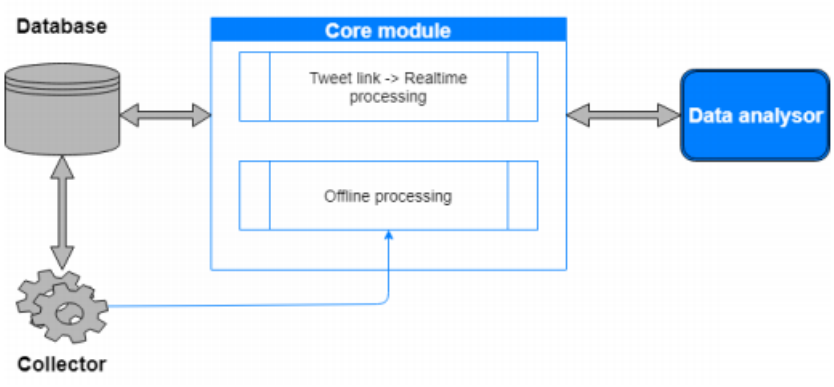
3.1.1 Βασικές Οντότητες

Οντότητα πυρήνα (Core module): Εδώ γίνεται ο συγχρονισμός και η ενορχήστρωση των διαδικασιών. Υπάρχει η online λειτουργία, κατά την οποία δεδομένου ενός URL υπάρχει απευθείας απόφαση για την αξιοπιστία του καθώς και η offline λειτουργία η οποία σχετίζεται με την αξιολόγηση των χρηστών.

Οντότητα συλλογέα (Collector module): Η συλλογή δεδομένων είναι ζωτική σημασίας για την εφαρμογή καθώς ένα μη ενημερωμένο σύστημα είναι άχρηστο και μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα αξιοπιστίας. Ο συλλογέας αρχικοποιείται με αξιόπιστα δεδομένα και συνεχίζει συλλέγοντας συνεχώς δεδομένα δημοσιεύσεων και χρηστών τα οποία προωθεί στο σύστημα για offline επεξεργασία.

Οντότητα βάσης δεδομένων (Database module): Χειρίζεται αιτήματα για προσθήκη, αφαίρεση, ενημέρωση και ανάκτηση δεδομένων.

Οντότητα ανάλυσης (Analyzing module): Αποτελεί μία από τις σημαντικότερες οντότητες του συστήματος καθώς εδώ γίνεται η ανάλυση και η σύγκριση των δημοσιεύσεων.



Σχήμα 2: Αλληλεπιδράσεις μεταξύ των κύριων ενοτήτων

3.1.2 Υπολογισμός Βαθμού Αξιοπιστίας

Ο αλγόριθμος ακολουθεί τα εξής βήματα:

1. Ανακτά τις πληροφορίες για το χρήστη από τη βάση δεδομένων.
2. Εάν δεν υπάρχει ο χρήστης στη βάση, λαμβάνει τις πληροφορίες του χρήστη από το Twitter και αρχικοποιεί το βαθμό αξιοπιστίας του σε 0.
3. Λαμβάνει τη δημοσίευση από το Twitter.
4. Υπολογίζει το βαθμό αξιοπιστίας της δημοσίευσης μέσω τον APIs που χρησιμοποιούνται κατά την σύγκριση των παρόμοιων δημοσιεύσεων.
5. Υπολογίζει ένα βαθμό αξιοπιστίας χρήστη υπολογίζοντας τον μέσο όρο του προϋπάρχοντα βαθμού και του καινούργιου.

3.1.2.1 Σχετικά με τον Βαθμό Αξιοπιστίας της Δημοσίευσης

1. Ξεκινά ως 0 και εξάγονται οι οντότητες (εταιρίες, άνθρωποι, μέρη, προϊόντα κτλ) με το OpenCalais (αποσύνθεση με NER).
2. Αν η δημοσίευση δεν περιέχει οντότητες, απαντάται : “Δεν δυνάμεθα να απαντήσουμε με ακρίβεια εάν είναι ή όχι αξιόπιστη η δημοσίευση.”(We cannot tell with precision whether the tweet is fake or not). Ο βαθμός γίνεται -500.
3. Το σύστημα αναζητά παρόμοιες δημοσιεύσεις από αξιόπιστους χρήστες. (Ομοιότητα κειμένου άνω του 0.25)
4. Εάν δεν βρεθούν, απαντάται : “Δεν δυνάμεθα να απαντήσουμε με ακρίβεια εάν είναι ή όχι αξιόπιστη η δημοσίευση.”(We cannot tell with precision whether the tweet is fake or not). Ο βαθμός γίνεται -500.
5. Για κάθε παρόμοια δημοσίευση που βρίσκεται, εξάγονται οι οντότητες με το OpenCalais και προστίθενται 10 πόντου στο βαθμό. Αν η ομοιότητα είναι ίση με 1 τότε προστίθενται 40 πόντοι στο βαθμό. Αν η ομοιότητα είναι μεταξύ 0.5 και <1, τότε προστίθενται 10 πόντοι στο βαθμό. Αν η ομοιότητα είναι μεταξύ 0.1 και 0.5, τότε προστίθενται 5 πόντοι στο βαθμό, αλλιώς δεν υπάρχει επιπλέον βαθμολόγηση.
6. Αν κατά τη διάρκεια του ελέγχου των δημοσιεύσεων ο βαθμός ξεπεράσει του 150 πόντους, τότε διακόπτεται η διαδικασία και ο βαθμός μετατρέπεται σε 100. Το αυτό και αν ο βαθμός υπερβεί τους 100 πόντους μέχρι το τέλος της διαδικασίας.
7. Τέλος, αν ο βαθμός είναι μεγαλύτερος ή ίσος του 50, απαντάται: “Η δημοσίευση δεν είναι ψευδής. Αξιοπιστία: βαθμός.” (Tweet is not fake. Confidence: score),

αλλιώς, αν ο βαθμός εί-ναι μικρότερος του 50, απαντάται: “ Η δημοσίευση είναι ψευδής. Αξιοπιστία: 50- βαθ-μός.”(Tweet is fake. Confidence: 100 – score).

8. Εάν βρεθεί μόνο μία δημοσίευση, ο βαθμός θα είναι -βαθμός (αρνητικός) και η απάντηση: “Βρέθηκε μόνο μία πηγή.” (Only one source found.).
9. Αν κάτι πήγε λάθος, ο βαθμός θα είναι -1000.

Συμπερασματικά, η πιθανή βαθμολογία μίας δημοσίευσης μπορεί να λάβει τιμές: [-1000, 500] ή [-50,100].

3.1.2.2 Σχετικά με τον Βαθμό Αξιοπιστίας του Χρήστη

Ο βαθμός του χρήστη εξαρτάται από τις τελευταίες δημοσιεύσεις του. Ένας χρήστης με μία δημοφιλή δημοσίευση, θα λάβει πολύ υψηλό βαθμό αξιοπιστίας.

Αφότου υπολογιστεί ο βαθμό αξιοπιστίας της δημοσίευσης, χρησιμοποιείται για τον υπολογισμό ενός βαθμού t, ο οποίος μετέπειτα χρησιμοποιείται για τον υπολογισμό του βαθμού αξιοπιστίας του χρήστη. Πιο αναλυτικά:

- Αρχικά t=0.
- Εάν η *βαθμολογία δημοσίευσης* είναι -1000 ή -500, τότε *t* παραμένει 0.
- Εάν η *βαθμολογία δημοσίευσης* είναι μεταξύ -50 και 50, τότε *t* =1.
- Εάν η *βαθμολογία δημοσίευσης* είναι μεταξύ -51 και 70, τότε *t* =9.
- Εάν η *βαθμολογία δημοσίευσης* είναι μεταξύ -71 και 90, τότε *t* =10.
- Εάν η *βαθμολογία δημοσίευσης* είναι μεταξύ -91 και 99, τότε *t* =11.
- Εάν η *βαθμολογία δημοσίευσης* είναι 100, τότε *t* =12.

Τώρα με βάσει με τη μεταβλητή t, υπολογίζεται ο βαθμό αξιοπιστίας του χρήστη:

- u = 0 (αρχικά – 0 αναλυμένες δημοσιεύσεις)
- u = (u +t)/2 t/2(μετά από 1 αναλυμένη δημοσίευση)
- u = (u +t)/2 t/2(μετά από 2 αναλυμένες δημοσιεύσεις)

Όσο πιο υψηλή βαθμολογία λάβει ο χρήστης τόσο πιο αξιόπιστος είναι. Για να επιτύχει υψηλή βαθμολογία ένας χρήστης θα πρέπει να έχει υψηλή βαθμολογία σε όλες τις δημοσιεύσεις του. Το σύστημα έχει προβλέψει να σταθεροποιείται η βαθμολογία ενός χρήστη στη περίπτωση που η ίδια δημοσίευση ελεγχθεί 2-3 φορές, για να την αποφυγή χειραγώγησης του συστήματος προς όφελος ενός χρήστη.

Εν κατακλείδι, το σύστημα προσπαθεί να εντοπίσει ψευδείς δημοσιεύσεις μέσω της επεξεργασίας του κειμένου και της αξιοπιστίας του χρήστη. Δεν είναι το ιδανικό μοντέλο, αλλά βρίσκεται ακόμα υπό ανάπτυξη με σκοπό την περαιτέρω βελτίωση.

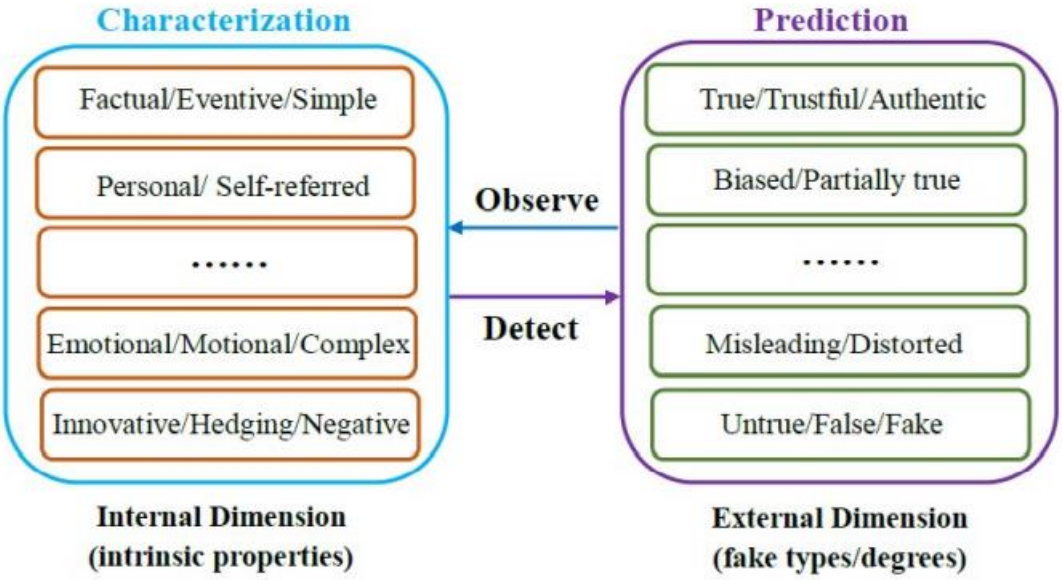
3.2 Κίνητρα, μέθοδοι και μετρικές αντιμετώπισης της παραπληροφόρησης μέσω της επεξεργασίας φυσικής γλώσσας. (Su et al. 2020 [56])

Αυτό το άρθρο ασχολείται με τα κύρια ζητήματα της παραπληροφόρησης και την ανίχνευσή τους. Παρουσιάζει, αναλύει και συγκρίνει μεθόδους ανίχνευσης, τεχνικές αναπαράστασης χαρακτήρων, τεχνικές αξιολόγησης και βάσεις δεδομένων. Αρχικά εισάγει το υπόβαθρο της έρευνας και τη σημασία της ανίχνευσης παραπληροφόρησης και στη συνέχεια επικεντρώνεται στην δισδιάστατη όψη αυτής της αποστολής: η εσωτερική διάσταση της βάσει-περιεχομένου ανάλυσης (δηλαδή ο χαρακτηρισμός των πληροφοριών χαμηλής αξιοπιστίας) και την εξωτερική διάσταση της προγνωστικής μοντελοποίησης (δηλαδή την αυτόματη ανίχνευση της παραπληροφόρησης).

3.2.1 Υπόβαθρο

3.2.1.1 Μια δισδιάτατη οπτική στον τρόπο μελέτης της παραπληροφόρησης

1. Η εσωτερική διάσταση υπογραμμίζει τη διαδικασία παρατήρησης του χαρακτηρισμού των εγγενών ιδιοτήτων της παραπληροφόρησης σε σύγκριση με τις πραγματικές πληροφορίες.
2. Η εξωτερική διάσταση υπογραμμίζει τη διαδικασία ανίχνευσης της πρόβλεψης των ψεύτικων τύπων / βαθμών με τη μοντελοποίηση διαφόρων αναπαραστάσεων πληροφοριών.



Σχήμα 3: Μια δισδιάστατη άποψη της μελέτης της παραπληροφόρησης

Η διαδικασία παρατήρησης και του χαρακτηρισμού

Κατά τη διαδικασία της παρατήρησης εστιάζουμε στον εντοπισμό των διαφόρων “διαρροών στη γλώσσα” επεξεργαζόμενοι τις μοναδικές ιδιότητες της γλώσσας, τα χαρακτηριστικά του περιεχομένου, τα μοτίβα διάδοσης της παραπληροφόρησης σε σύγκριση με τις αληθινές πληροφορίες. Η διαδικασία αυτή είναι απαιτητική, έτσι ο στοχεύουμε στα “προγνωστικά εξαπάτησης” μέσω των ενδείξεων που βρίσκονται στο περιεχόμενο ενός μηνύματος. Καταλαβαίνουμε λοιπόν ότι η επιλογή των κατάλληλων

εργαλείων για την επεξεργασία των λέξεων ενός κειμένου είναι πολύ κρίσιμη για το σύστημά μας.

Η διαδικασία ανίχνευσης και πρόβλεψης

Το κύριο πλαίσιο για την ανίχνευση παραπληροφόρησης συνήθως περιλαμβάνει δύο φάσεις:

1. Εξαγωγή των χαρακτηριστικών. Η φάση αυτή στοχεύει στην αναπαράσταση των πληροφοριών του περιεχομένου και των σχετικών βοηθητικών πληροφοριών σε μία μαθηματική δομή.
2. Κατασκευή μοντέλων. Η φάση αυτή χτίζει περαιτέρω μοντέλα μηχανικής μάθησης για να διαφοροποιήσει καλύτερα την παραπληροφόρηση από τις πληροφορίες υψηλής αξιοπιστίας βάσει των αναπαραστάσεων των χαρακτηριστικών.

3.2.2 Προσεγγίσεις Ανίχνευσης

Οι προσεγγίσεις για την ανίχνευση της παραπληροφόρησης χωρίζονται γενικά σε δύο κατηγορίες: αυτές που βασίζονται στο περιεχόμενο του κειμένου και σε αυτές που βασίζονται στον τρόπο διάδοσης της πληροφορίας. Ενώ οι προσεγγίσεις που βασίζονται στη διάδοση βασίζονται κυρίως σε κοινωνικά πλαίσια και κοινωνικές δεσμεύσεις χρηστών όπως: η σχέση εκδότη-ειδήσεων, η σχέση ειδήσεων-χρηστών, τα κοινωνικά δίκτυα μεταξύ των χρηστών και τα προφίλ των χρηστών, οι τεχνικές και οι μέθοδοι NLP εφαρμόζονται κυρίως για τη διερεύνηση του περιεχομένου του κειμένου για παραπληροφόρηση. Σε αυτές τις μεθόδους και τις τεχνικές επικεντρώνεται το συγκεκριμένο άρθρο.

3.2.2.1 Διατύπωση του προβλήματος

Οι τρεις βασικές κατηγορίες διατύπωσης του προβλήματος για τον εντοπισμό παραπληροφόρησης μέσω των NLP τεχνικών είναι οι εξής:

Ταξινόμηση: Η απλούστερη προσέγγιση για την αντιμετώπιση της παραπληροφόρησης θα μπορούσε να είναι η προσέγγισή της ως ένα δυαδικό πρόβλημα ταξινόμησης ή ένας χαρακτηρισμός της ως ψευδής ή αληθινή. Αυτή η προσέγγιση όμως δεν θα μπορούσε να είναι αποτελεσματική καθώς δεν περιλαμβάνει τις περιπτώσεις που οι πληροφορίες είναι εν μέρει πραγματικές ή εν μέρει ψευδείς. Για την αντιμετώπιση αυτού του προβλήματος, η ανίχνευση παραπληροφόρησης μπορεί επίσης να διατυπωθεί ως ένα πρόβλημα πολλαπλής ταξινόμησης προσθέτοντας επιπλέον κλάσεις (και ετικέτες) στα σύνολα δεδομένων (π.χ. false, barely-true, half-true, mostly-true, και true).

Παλινδρόμηση: Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Η αξιολόγηση μπορεί να γίνει υπολογίζοντας τη διαφορά μεταξύ των προβλεπόμενων βαθμολογιών και των βαθμολογιών επιτόπιας αλήθειας(ground truth scores) ή χρησιμοποιώντας τους συντελεστές συσχέτισης των Pearson / Spearman. Εντούτοις, δεδομένου ότι οι διαθέσιμες βάσεις δεδομένων έχουν ετικέτες επιτόπιας αλήθειας, η διαμόρφωση παλινδρόμησης καθίσταται προβληματική καθώς ο τρόπος μετατροπής των διακριτών ετικετών σε αριθμητικές βαθμολογίες φαίνεται ότι αποτελεί ένα δύσκολο ζήτημα.

Ομαδοποίηση: Η ανίχνευση της παραπληροφόρησης συνήθως διατυπώνεται ως εποπτευόμενο μαθησιακό πρόβλημα, δεδομένου ενός προϋπάρχοντος συνόλου δεδομένων με ετικέτες. Ωστόσο, τα δεδομένα του πραγματικού κόσμου είναι πιο συχνά χωρίς ετικέτες. Ως εκ τούτου, προτείνονται ημιεποπτευόμενες και μη εποπτευόμενες μέθοδοι για την ανάπτυξη συστημάτων ανίχνευσης παραπληροφόρησης, διατυπώνοντάς το ως πρόβλημα ομαδοποίησης.

Παρά τις παραπάνω τρεις διατυπώσεις, οι περισσότερες προσεγγίσεις για την ανίχνευση παραπληροφόρησης βασίζονται στην εποπτευόμενη λειτουργία, η οποία απαιτεί ένα τυποποιημένο σύνολο δεδομένων με έγκυρες ετικέτες για την εκπαίδευση ενός μοντέλου. Η εγκυρότητα όμως αυτή, για να είναι αξιόλογη, απαιτεί χρόνο και αρκετή ανθρώπινη εργασία. Συμπεραίνουμε λοιπόν ότι, σε ένα ρεαλιστικό σενάριο, είναι πιο πρακτική η χρήση ημι-εποπτευόμενων ή μη εποπτευόμενων μοντέλων.

3.2.3 Μοντέλα και Αλγόριθμοι

Πολλές μελέτες στην ανίχνευση παραπληροφόρησης έχουν χρησιμοποιήσει αλγόριθμους μηχανικής μάθησης. Η κύρια τεχνική μπορεί να γενικευτεί ως εργασία πολλαπλής ταξινόμησης με τη χρήση προγνωστικής μοντελοποίησης σε διανύσματα χαρακτηριστικών βάσει περιεχομένου σε πολλαπλά επίπεδα. Οι διάφορες κατηγορίες των μοντέλων ανίχνευσης παρουσιάζονται στις ακόλουθες υποενότητες.

3.2.3.1 Στατιστικά Μοντέλα

Τα πιο συχνά χρησιμοποιούμενα στατιστικά μοντέλα στην ανίχνευση της παραπληροφόρησης είναι οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machine [SVM]) και ο Ταξινομητής Naive Bayes (Naive Bayes Classifier [NBC]). Όταν ένα μαθηματικό μοντέλο έχει εκπαιδευτεί επαρκώς από παραδείγματα με προϋπάρχουσα ετικέτα σε μία από τις δύο κατηγορίες, μπορεί να προβλέψει περιπτώσεις μελλοντικής εξαπάτησης βάσει αριθμητικής ομαδοποίησης και αποστάσεων. Η χρήση διαφορετικών μεθόδων ομαδοποίησης και συναρτήσεων απόστασης μεταξύ σημείων δεδομένων διαμορφώνουν την ακρίβεια του SVM. Οι αλγόριθμοι Naive Bayes κάνουν ταξινομήσεις με βάση συσσωρευμένες αποδείξεις για τη συσχέτιση μεταξύ μιας δεδομένης μεταβλητής (π.χ. n-άδες [n-gram]) και των άλλων μεταβλητών που υπάρχουν στο μοντέλο. Άλλα στατιστικά μοντέλα με ξεχωριστά προτερήματα το καθένα, πέρα των SVM και NBC είναι: η Λογιστική Παλινδρόμηση (Logistic Regression [LR]), οι K-κοντινότεροι γείτονες (K-Nearest-Neighbours [KNN]), τα Δέντρα Αποφάσης (Decision Trees) και ο Ταξινομητής Τυχαίων Δασών (Random Forest Classifier [RFC]). Στην παρακάτω εικόνα θα βρείτε μια στατιστική μελέτη για τις διάφορες προσπάθειες ερευνητών να συνδυάσουν διαφορετικά στατιστικά μοντέλα.

Πίνακας 1: Μελέτες σχετικά με τη χρήση στατιστικών μοντέλων

Study	Dataset	Best Feature	Best Model	Performance
Gilda [33]	Data from Signal Media	TF-IDF of bi-grams	Stochastic Gradient Descent	72% Acc.
Feng, Banerjee, and Choi [34]	Online hotel review corpora	Syntactic stylometry	SVM classifier	91% Acc.
Feng and Hirst [35]	opspamv1.3	Syntactic+semantic 'object:descriptor' pairs	SVM^{perf}	91.3% Acc.
Rubin, Lukoianova, and Tatiana [36]	A sample of 36 elicited personal stories	Rhetorical structure	Vector Space Model	67% agreement to human raters
Ciampaglia et al. [37]	Three datasets of RDF triples	Knowledge graph extracted from Wikipedia	K-Nearest-Neighborhood	0.65 AUC

3.2.3.2 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα μας δίνουν τη δυνατότητα να αποφύγουμε την χρονοβόρα διαδικασία της χειροποίητης εξαγωγής χαρακτηριστικών για την εκπαίδευση ενός μοντέλου μηχανικής μάθησης, με σκοπό την εγκυρότητα στον εντοπισμό των ψευδών ειδήσεων. Πλέον τα μοντέλα Βαθιάς Μάθησης (deep learning) ξεπερνούν σε απόδοση τα παραδοσιακά μοντέλα Μηχανικής Μάθησης (machine learning). Ωστόσο, ένα μεγάλο πρόβλημα των μοντέλων βαθιάς μάθησης είναι ότι συχνά απαιτούν μαζικά δεδομένα και σημαντικό χρόνο εκπαίδευσης και ρύθμισης παραμέτρων, ενώ η απόδοσή τους είναι συνήθως δύσκολο να ερμηνευθεί. Στην παρακάτων εικόνα συγκρίνονται τα χαρακτηριστικά, τα μοντέλα βαθιάς μάθησης και οι κορυφαίες επιδόσεις σε ορισμένα αντιπροσωπευτικά έργα.

Πίνακας 2: Μελέτες για τη χρήση μοντέλων deep learning

Study	Dataset	Best Feature	Best Model	Performance
Rashkin et al. [26]	A political fact-checking database available at PolitiFact	Text + lexicons in LIWC	LSTM	0.57 F1 (2-class), 0.22 F1 (6-class)
Ma et al. [38]	Two public Twitter datasets	Structural and textural properties	A recursive neural networks (RvNN)	0.835 F1
Zhang et al. [40]	PolitiFact	Textual content information and latent feature	Deep diffusive network model	0.63 F1 (2-class), 0.28 F1 (multi-class)
Yang et al. [41]	A news dataset scrapped from the web	Text and image information	Convolutional Neural Network	0.92-0.93 F1
Liu and Wu [42]	Three real-world data collections: Weibo, Twitter15 and Twitter16	A multivariate time series of user characteristics	A combination of recurrent and convolutional networks	85% Acc. on Twitter and and 92% Acc. on Sina Weibo

3.2.4 Χαρακτηριστικά και Αναπαραστάσεις

Έχει παρατηρηθεί ότι οι απλές ετικέτες-περιεχομένου με n-grams και οι επιπόλαιες ετικέτες με Part-of-Speech (POS) είναι ανεπαρκείς για την διεργασία ταξινόμησης και συχνά αποτυγχάνουν να λάβουν υπόψιν σημαντικές πληροφορίες του περιεχομένου. Εντούτοις, φαίνονται πολύ πιο αποδοτικές σε συνδυασμό με πιο περίπλοκες μεθόδους ανάλυσης. Για παράδειγμα η μέθοδος των Ν-άδων (n-grams) φαίνεται να συνδυάζεται ιδανικά με την σε βάθος ανάλυση του συντακτικού χρησιμοποιώντας πιθανοτική γραμματική χωρίς περιεχόμενο (Probabilistic Context Free Grammars [PCFG]). Ακολουθούν σχολιασμοί για τις πιο κοινά χρησιμοποιούμενες αναπαραστάσεις χαρακτηριστικών.

3.2.4.1 Συχνότητα Όρου – Άνιση κατανομή του όρου (TF-IDF)

Αποτελεί την πιο συχνή επιλογή. Μας δείχνει το πόσο συχνά εμφανίζεται μία συγκεκριμένη Ν-άδα σε ένα έγγραφο σε σχέση με το πόσο συχνά εμφανίζεται η Ν-άδα σε όλα τα έγγραφα ενός σώματος. Συνήθως, το TF-IDF υπολογίζεται για κάθε n-gram σε κάθε έγγραφο και δημιουργεί για ένα πίνακα βάσει των προσκοπτόντων χαρακτηριστικών.

3.2.4.2 Αναπαραστάσεις Γνώσεως

Εκτός από τις N-άδες, οι πληροφορίες μπορούν να αναπαρασταθούν από τις γνώσεις που εξάγονται από τη πρόταση. Η γνώση εδώ συνήθως ορίζεται ως ένα σύνολο τριπλής πλειάδας που περιέχει Υποκείμενο, Αντικείμενο, Κατηγορούμενο. Για παράδειγμα η πλειάδα (DonaldTrump, Profession, President) είναι εξαγόμενη γνώση από την πρόταση “Donald Trump is the president of the U.S” και μπορεί να αντιπροσωπεύσει το περιεχόμενο της πρότασης. Ουσιαστικά λοιπόν συγκρίνουμε το περιεχόμενο της είδησης με Γράφους και Βάσεις Γνώσεως, θεωρώντας το περιεχόμενο αυτών των δύο ως απόλυτη αλήθεια καθώς την έχουμε επεξεργαστεί χειροκίνητα. Ωστόσο, υπάρχουν ορισμένες ελλείψεις της προσέγγισης που βασίζεται στη γνώση για τον εντοπισμό παραπληροφόρησης. Το χειροκίνητα επεξεργασμένο περιεχόμενο αποτελεί και το μεγαλύτερο πρόβλημα καθώς δεν μπορούμε να είμαστε απόλυτα σίγουροι για το αν η δική μας κρίση είναι σωστή και επιπλέον, η πληροφορίες αλλάζουν συνεχώς οπότε θα πρέπει να διατηρείται ενημερωμένη η βάση των δεδομένων μας.

3.2.5 Γλωσσικά χαρακτηριστικά του τομέα

Σε αυτή τη κατηγορία ασχολούμαστε με τον τομέα των ειδήσεων αναζητώντας αναφερόμενες λέξεις, εξωτερικούς συνδέσμους, αριθμούς γραφημάτων, το μέσο μήκος γραφημάτων κ.λπ. Επιπλέον, άλλες δυνατότητες εμπνευσμένες από ψυχολογικές θεωρίες μπορούν να σχεδιαστούν ειδικά για να αποτυπώσουν τα παραπλανητικά στοιχεία στον τρόπο γραφής για να διαφοροποιήσουν τα ψεύτικα νέα από τα αληθινά. Έτσι βάσει του τρόπου γραφής του, αναπαριστάται το περιεχόμενο σε πολυ-επιπέδα, τα οποία έπειτα χρησιμοποιούνται ως χαρακτηριστικά για τη πρόβλεψη μέσω της μεθόδου μηχανικής μάθησης.

3.2.6 Γλωσσικά Στοιχεία

Λόγω της προσπάθειας να παραπληροφορήσουν το κοινό, οι εκδότες των ψευδών ειδήσεων χρησιμοποιούν εμπρηστική γλώσσα με χαρακτηριστικά που θα οδηγήσουν τον χρήστη να διαβάσει το άρθρο (clickbait) ή για να δημιουργήσουν σύγχυση. Αυτό ωστόσο μας δίνει τη δυνατότητα να αναγνωρίσουμε τέτοιες γλωσσικές διαφοροποιήσεις και να τις χρησιμοποιήσουμε για να εντοπίσουμε αναξιόπιστες ειδήσεις. Κοινές τεχνικές εντοπισμού των γλωσσικών αυτών στοιχείων είναι οι εξής:

3.2.6.1 Λεξικά Χαρακτηριστικά

Τα λεξικά χαρακτηριστικά περιλαμβάνουν χαρακτηριστικά επιπέδου χαρακτήρων και πιπέδου λέξεων, όπως συνολικές λέξεις, χαρακτήρες ανά λέξη, συχνότητα μεγάλων λέξεων και μοναδικές λέξεις. Τέτοιες δυνατότητες συνήθως εξάγονται με TF-IDF, Γλωσσική Έρευνα Καταμέτρηση Λέξεων (Linguistic Enquiry Word Count [LIWC]) και n-grams. Για ακολουθίες λέξεων, χρησιμοποιούνται συνήθως εκπαιδευμένα διανύσματα ενσωμάτωσης λέξεων όπως word2vec και GloVe. Για τα γλωσσικά χαρακτηριστικά, η πιο απλή αναπαράσταση κειμένων είναι η προσέγγιση με αναπαράσταση-σάρωσης-λέξεων (bag of words). Η απλότητα όμως αυτή αποτελεί και το πρόβλημά της, καθώς βασίζεται αποκλειστικά στη γλώσσα και στις απομονωμένες N-άδες (n-grams), οι οποίες συχνά είναι διαχωρισμένες από το πραγματικά χρήσιμο περιεχόμενο του κειμένου. Για αυτό συνίσταται η τεχνική αυτή να συνδυάζεται με άλλες για πιο ακριβές αποτέλεσμα.

3.2.6.2 Συντακτικά Χαρακτηριστικά

Τα συντακτικά χαρακτηριστικά επικεντρώνονται στο συντακτικό του κειμένου εξετάζοντας, συχνότητα λέξεων, φράσεων ή σημείων στίξης και τοποθετούν ετικέτες part-of-speech (POS). Τα παραπάνω αποτελούν επιφανεικά συντακτικά χαρακτηριστικά, ενώ ακολουθεί βαθύτερη συντακτική ανάλυση όπως η επανεγγραφή των λέξεων. Αυτή λαμβάνεται μέσω δέντρων συντακτικής ανάλυσης βασισμένα στην πιθανοτική γραμματική χωρίς περιεχόμενο (Probabilistic Context Free Grammars [PCFG]). Έπειτα υπολογίζονται τα στατικά της επανεγγραφής με TF-IDF. Για παράδειγμα, ουσιαστικά και ρήματα, τα οποία με τη σειρά τους ξαναγράφονται από τα συντακτικά μέρη τους. Εργαλεία τρίτων, όπως ο Stanford Parser, η συντακτική ανάλυση AutoSlog-TS κ.α. βοηθούν στην αυτοματοποίηση. Μόνη της η ανάλυση σύνταξης μπορεί να μην είναι επαρκώς ικανή να εντοπίσει την εξαπάτηση για αυτό οι μελέτες συχνά συνδυάζουν αυτήν την προσέγγιση με άλλες γλωσσικές ή τεχνικές ανάλυσης (πχ. δικτύου).

3.2.6.3 Συναισθηματικά Χαρακτηριστικά

Εμπνευσμένο από θεμελιώδεις θεωρίες κοινωνικής ψυχολογίας, τα χαρακτηριστικά του τρόπου γραφής σε συναισθηματικό επίπεδο διερευνούν ορισμένα ψυχο-γλωσσικά χαρακτηριστικά αναλύοντας το συναίσθημα, την ανεπισημότητα, την ποικιλομορφία, την υποκειμενικότητα, τις γνωστικές και αντιληπτικές διαδικασίες στο περιεχόμενο των ειδήσεων, καθώς και την ποσότητα συγκεκριμένων χαρακτήρων, λέξεων, προτάσεων ή ακόμα και παραγράφων. Αυτά τα χαρακτηριστικά εξάγονται ως χαρακτηριστικά τρόπου γραφής για την ανίχνευση ψευδών πληροφοριών στα κείμενα ειδήσεων.

3.2.6.4 Χαρακτηριστικά λόγου

Μία ρητορική προσέγγιση εφαρμόζεται συνήθως για την εξαγωγή χαρακτηριστικών σε επίπεδο λόγου με βάση τη Θεωρία της Ρητορικής Δομής (Rhetorical Structure Theory [RST]), η οποία είναι ένα αναλυτικό πλαίσιο για την εξέταση της συνοχής μιας ιστορίας. Σε συνδυασμό με το Μοντέλο Διανυσματικού Χώρου (Vector Space Model [VSM]) αποτελεί συχνή τεχνική για τον εντοπισμό των ψευδών ειδήσεων.

3.2.6.5 Λανθάνουσα Αναπαράσταση

Μέσω παραγοντοποίησης πινάκων ή τανυστών ή μοντέλων νευρωνικών δικτύων στη βαθιά μάθηση (π.χ. CNN, RNN, LSTM, δίκτυο προσοχής, δίκτυο μνήμης κ.λπ.), τα λανθάνοντα χαρακτηριστικά του περιεχομένου των ειδήσεων μπορούν να εξαχθούν αυτόματα χωρίς χειροποίητη εργασία. Τέτοιες μέθοδοι αναπαράστασης εξοικονομούν χρόνο και εργασία για τον χειροποίητο σχεδιασμό των χαρακτηριστικών του κειμένου, αλλά η επιλογή ή η εξαγωγή των λανθανόντων χαρακτηριστικών συχνά επέρχεται μέσω της εμπειρίας ή μέσω τεχνικών χωρίς θεωρητική βάση. Επομένως, είναι δύσκολο να κατανοήσουμε τα παραγόμενα χαρακτηριστικά με μεθόδους λανθάνουσας αναπαράστασης. Η παρακάτω εικόνα δείχνει μερικά πρόσφατα πειράματα που χρησιμοποιούν διαφορετικές μεθόδους για ανίχνευση της παραπληροφόρησης.

Πίνακας 3: Πρόσφατα πειράματα σχετικά με την ανίχνευση της παραπληροφόρησης

Paper	Dataset	Method	Acc.
Karimi and Tang (2019)	FakeNewsNet	N-grams	72.37
	Fake or Real News	LIWC	70.26
		RST	67.68
		BiGRNN-CNN	77.06
		LSTM[w+s]	80.54
		LSTM[s]	73.63
		HDSF	**82.19**
Wu et al. (2019)	PHEME	SVM	72.18
		CNN	59.23
		TE	65.22
		DeClarE	67.87
		MTL-LSTM	74.94
		TRNN	78.65
		Bayesian-DL	80.33
		sifted MTL	**81.27**
Yang et al. (2019)	LIAR	Major voting	58.6
		TruthFinder	63.4
		LTM	64.1
		CRH	63.9
		UFD	**75.9**
Qian et al. (2018)	Weibo	LIWC	66.06
		POS-gram	74.77
		1-gram	84.76
		CNN	86.23
		TCNN	88.08
		TCNN-URG	**89.84**
Karimi et al. (2018)	LIAR	SVM	29.98
		RandomForests	27.01
		NN	29.12
		MMFD	**38.81**
Roy et al. (2018)	LIAR	hybrid CNN	27.4
		hybrid LSTM	41.5
		Bi-LSTM	42.65
		CNN	42.89
		RNN-CNN	**44.87**

3.2.7 Πολυπλοκότητα

Στον πραγματικό κόσμο, οι εκδότες των ψευδών ειδήσεων μπορούν να ελέγξουν και να περιορίσουν κάποιο στρατηγικά λεκτικά στοιχεία τα οποία θα τους αποκάλυπταν σε έναν έλεγχο για ύποπτη, αναξιόπιστη και ψευδή συμπεριφορά. Επομένως, πολλές ενδείξεις μπορεί να καταστούν λιγότερο αποτελεσματικές και να παρακαμφθούν. Αντίθετα ορισμένα στοιχεία όπως οι λέξεις που φανερώνουν συναίσθημα, τα συντακτικά μοτίβα κα., μπορούν να διαρρεύσουν ακόμα και από τους συγγραφείς ψευδών ειδήσεων που γνωρίζουν πώς να αποφεύγουν τις λεκτικές παγίδες. Αυτό δίνει υψηλότερες πιθανότητες επιτυχίας στον εντοπισμό της εξαπάτησης εντοπίζοντας ενδείξεις πολυπλοκότητας γλώσσας. Πρόσφατες μελέτες αναφέρουν ότι, οι ψευδείς δηλώσεις είναι πιο περίπλοκες από τις αληθινές δηλώσεις, καθώς η εξαπάτηση θεωρείται ότι είναι γνωστικά πιο απαιτητική από το να πεις την αλήθεια.

3.2.8 Βάσεις Δεδομένων

Η κατασκευή βάσεων δεδομένων με στοιχεία για την εγκυρότητα δημοσιεύσεων είναι μια από τις σημαντικότερες προκλήσεις για την αυτόματη ανίχνευση ψευδών ειδήσεων λόγω της ποσότητας και της ποιότητας των δεδομένων καθώς και του κόστους για σχολιασμό, ειδικά για την εποπτευόμενη μάθηση. Γενικά, οι σχολιασμοί σε δεδομένα ειδήσεων μπορούν να γίνουν με τους ακόλουθους τρόπους: ειδικοί δημοσιογράφοι, ιστότοποι ελέγχου γεγονότων (π.χ. PolitiFact, Snopes), ανιχνευτές της βιομηχανίας και εργαζόμενοι που προέρχονται από το πλήθος. Η παρακάτω εικόνα δείχνει μια σύνοψη των βάσεων δεδομένων για ανίχνευση ψεύτικων ειδήσεων.

Πίνακας 4: Δημόσια διαθέσιμες βάσεις δεδομένων για τον εντοπισμό ψευδών ειδήσεων:

Dataset	Main Input	Data Size	Label	Annotation
LIAR	Short claims	12,836	Six-class	PolitiFact
FEVER	Short claims	185,445	Three-class	Trained annotators
BuzzFeedNews	Facebook post	2,282	Four-class	Journalists
BuzzFace	Facebook post	2,263	Four-class	Journalists
PHEME	Tweet	6,425 (threads)	Three-class	Journalists
RumourEval	Tweet	325 (treads)	Three-class	Journalists
CREDBANK	Tweet	60 million	30-dimension vector	Crowdsourcing workers
BS Detector	Web Post	12,999	Three-class	BS Detector
FakeNewsNet	News Articles	23,921	Fake or Real	Editors
Fake or Real News	News Articles	7,800	Fake or Real	-

3.2.8.1 Βάσεις Δεδομένων με Σύντομες Δηλώσεις

LIAR: Αυτή η βάση δεδομένων συλλέγεται από τον ιστότοπο PolitiFact που ελέγχει τα γεγονότα μέσω του API του [5]. Σχολιάζεται με έξι εκλεπτυσμένα μαθήματα και περιλαμβάνει 12.836 σχολιασμένες σύντομες δηλώσεις (δειγματοληψία από διάφορα περιβάλλοντα, όπως δελτία ειδήσεων, τηλεοπτικές ή ραδιοφωνικές συνεντεύξεις, ομιλίες εκστρατείας κ.λπ.) που αναφέρθηκαν κατά τη διάρκεια του έτους 2007 έως 2016 μαζί με διάφορες πληροφορίες σχετικά με τον ομιλητή. Σε αυτή τη βάση δεδομένων, κάθε σειρά δεδομένων περιέχει μια σύντομη δήλωση, μια ετικέτα της δήλωσης, το θέμα, το πλαίσιο της δήλωσης και 10 άλλες στήλες που αντιστοιχούν σε διάφορες πληροφορίες σχετικά με τον ομιλητή, όπως το ιστορικό δηλώσεων του ομιλητή και η συνεργασία με τα κόμματα.

FEVER: Αυτή η βάση δεδομένων παρέχει σχετικές αποδείξεις για σύντομες δηλώσεις με σκοπό τον εντοπισμό ψευδών ειδήσεων. Περιέχει 185.445 αξιώσεις που συλλέχθηκαν από τη Wikipedia. Κάθε αξίωση χαρακτηρίζεται ως “Υποστηριζόμενη”, “Απορριφθείσα” ή “Μη επαρκείς πληροφορίες”. Η βάση δεδομένων επισημαίνει επίσης τις προτάσεις από τη Wikipedia που χρησιμοποιεί ως απόδειξη μιας αξίωσης. Ωστόσο, ο τύπος των γεγονότων και των αποδεικτικών στοιχείων από τη Βικιπαίδεια μπορεί να εμφανίζει ορισμένες στυλιστικές διαφορές από αυτές σε ρεαλιστικά σενάρια και δεν μπορεί να εφαρμοστεί πλήρως σε δεδομένα πραγματικού κόσμου.

3.2.8.2 Βάσεις Δεδομένων με Αναρτήσεις σε SNSs

BuzzFeedNews: Αυτό το σύνολο δεδομένων συλλέγει 2.282 δημοσιεύσεις που δημοσιεύθηκαν στο Facebook από 9 πρακτορεία ειδήσεων κατά τις εκλογές των ΗΠΑ το 2016. Κάθε αξίωση σε κάθε ανάρτηση ελέγχεται από 5 δημοσιογράφους του BuzzFeed. Αυτό το σύνολο δεδομένων εμπλουτίζεται περαιτέρω προσθέτοντας τα συνδεδεμένα άρθρα, τα συνημμένα μέσα και τα σχετικά μεταδεδομένα. Περιέχει 1.627 άρθρα: 826 κέντρου, 356 αριστερά και 545 δεξιά.

BuzzFace: Οι Santia and Williams [3] επεκτείνουν το σύνολο δεδομένων BuzzFeed με τα σχόλια που σχετίζονται με άρθρα ειδήσεων στο Facebook. Περιλαμβάνει 2.263 άρθρα ειδήσεων και 1,6 εκατομμύρια σχόλια.

PHEME: Αυτή η βάση δεδομένων συλλέγεται από νήματα συνομιλίας Twitter, συμπεριλαμβανομένων 6.425 νήματα Twitter και καλύπτει εννέα ειδήσεις όπως η αναταραχή Ferguson, οι πυροβολισμοί στον Τσάρλι Χέμπντο κ.λπ. Ένα νήμα συνομιλίας αποτελείται από ένα tweet που κάνει έναν αληθινό και έναν ψευδή ισχυρισμό, και μια σειρά απαντήσεων. Έτσι, το σύνολο δεδομένων έχει διαφορετικά επίπεδα σχολιασμών, συμπεριλαμβανομένου του επιπέδου νήματος και του επιπέδου tweet. Οι ετικέτες σχολιασμού είναι “αληθείς”, “ψευδείς” ή “μη επαληθευμένες”.

RumourEval: Αυτό το σύνολο δεδομένων είναι παρόμοιο με το PHEME όσον αφορά τη δομή δεδομένων, το περιεχόμενο που καλύπτει και το σχήμα σχολιασμών. Παρόμοια με το PHEME, το σύνολο δεδομένων περιέχει νήματα συνομιλίας Twitter που σχετίζονται με διαφορετικά αξιοσημείωτα γεγονότα. Ωστόσο, το RumourEval περιέχει μόνο 325 νήματα Twitter που συζητούν φήμες.

CREDBANK: Πρόκειται για μία βάση δεδομένων μεγάλης κλίμακας από περίπου 60 εκατομμύρια tweets που καλύπτουν 96 ημέρες από τον Οκτώβριο του 2015. Τα tweets στη βάση δεδομένων καλύπτούν πάνω από 1.000 γεγονότα ειδήσεων, με κάθε συμβάν να αξιολογείται για αξιοπιστία από 30 σχολιαστές από την Amazon Mechanical Turk [4].

BS Detector: Αυτή η βάση δεδομένων συλλέγεται από μια επέκταση προγράμματος περιήγησης που ονομάζεται BS detector που αναπτύχθηκε για τον έλεγχο της πραγματικότητας των ειδήσεων. Το σύνολο δεδομένων περιέχει κείμενο και μεταδεδομένα από 244 ιστότοπους και αντιπροσωπεύει 12,999 αναρτήσεις συνολικά από τις τελευταίες 30 ημέρες. Κάνει αναζήτηση σε όλους τους συνδέσμους σε μια δεδομένη ιστοσελίδα για αναφορές σε αναξιόπιστες πηγές, ελέγχοντας μια λίστα τομέων που συμμορφώνονται με μη αυτόματο τρόπο. Οι ετικέτες είναι οι έξοδοι του ανιχνευτή BS και όχι οι ανθρώπινοι σχολιαστές.

3.2.8.3 Βάσεις Δεδομένων με Ολόκληρα Άρθρα Ειδήσεων

FakeNewsNet: Πρόκειται για ένα συνεχιζόμενο έργο συλλογής δεδομένων για ανίχνευση πλαστών ειδήσεων. Αποτελείται από πρωτοσέλιδα και βασικά κείμενα ψεύτικων ειδήσεων από το BuzzFeed και το PolitiFact. Συλλέγει επίσης πληροφορίες σχετικά με τις κοινωνικές δεσμεύσεις αυτών των άρθρων από το Twitter, όπως οι σχέσεις χρήστη-ειδήσεων, κοινωνικά δίκτυα χρηστών-χρηστών και προφίλ χρηστών κ.λπ.

Fake or Real News: Αυτό η βάση δεδομένων αναπτύχθηκε από τον George McIntire και το αποθετήριο της βάσης δεδομένων στο GitHub περιλαμβάνει περίπου 7,8 χιλιάδες άρθρα ειδήσεων με ίση διανομή ψεύτικων και αληθινών ειδήσεων και τα μισά από τα νέα προέρχονται από τον πολιτικό τομέα. Το τμήμα ψεύτικων ειδήσεων αυτής της βάσης δεδομένων συλλέγεται από τη βάση δεδομένων ψεύτικων ειδήσεων Kaggle που

περιλαμβάνει ειδήσεις για τις εκλογές στις ΗΠΑ για το 2016. Το πραγματικό τμήμα ειδήσεων συλλέγεται από οργανισμούς μέσων όπως οι New York Times, WSJ, Bloomberg, NPR και ο Guardian κατά τη διάρκεια του 2015 και του 2016.

Αυτές είναι οι αντιπροσωπευτικές δημόσια διαθέσιμες βάσεις δεδομένων για ανίχνευση πλαστών ειδήσεων τα τελευταία χρόνια. Τα νέα που συλλέχθηκαν σε αυτά τα σύνολα δεδομένων έχουν επαληθευτεί ως προς το αληθές περιεχόμενό τους. Ωστόσο, εξακολουθούν να υπάρχουν ορισμένοι περιορισμοί. Για παράδειγμα, τα σύνολα δεδομένων που περιέχουν αναρτήσεις σε SNS περιορίζονται σε ένα μικρό εύρος θεμάτων και μπορεί να περιλαμβάνουν αναρτήσεις ή tweets που δεν έχουν σχέση με τις ειδήσεις.

3.2.9 Προκλήσεις και Προοπτικές

Η προσπάθεια για εντοπισμό των ψευδών ειδήσεων είχε μεγάλη επιτυχία τα τελευταία χρόνια. Ωστόσο, εξακολουθούν να υπάρχουν ορισμένα δύσκολα προβλήματα, τα οποία εντοπίζονται και σχολιάζονται παρακάτω.

3.2.9.1 Πρώιμη ανίχνευση παραπληροφόρησης

Η ανίχνευση της παραπληροφόρησης σε πρώιμο στάδιο διάδοσης είναι ένα σημαντικό βήμα για τον μετριασμό και την αντιμετώπισή της. Η δυσκολία είναι ότι υπάρχει μια αντιστάθμιση μεταξύ της ελαχιστοποίησης του χρονικού κενού και της μεγιστοποίησης της ακρίβειας. Ο Ramezani προτείνει ως λύση τη χρήση Επαναλαμβανόμενου Νευρωνικού Δικτύου (Recurrent Neural Network - RNN) με μια νέα λειτουργία απώλειας και έναν νέο κανόνα διακοπής. Πρώτον, το πλαίσιο των ειδήσεων είναι ενσωματωμένο σε μια αναπαράσταση κειμένου για συγκεκριμένη τάξη, έπειτα το μοντέλο χρησιμοποιεί το διαθέσιμο δημόσιο προφίλ των χρηστών και την ταχύτητα διάδοσης των ειδήσεων για την έγκαιρη επισήμανση παραπληροφόρησης.

3.2.9.2 Εύρεση Κατάλληλου Περιεχομένου για Έλεγχο

Με μαζικές πληροφορίες που δημιουργούνται σε πλατφόρμες κοινωνικών μέσων, ο προσδιορισμός αξιόπιστου περιεχομένου μπορεί να βελτιώσει την αποτελεσματικότητα της ανίχνευσης παραπληροφόρησης. Για να προσδιορίσετε εάν ένα συγκεκριμένο θέμα ή ειδήσεις αξίζει τον έλεγχο, η ανάλυση παραπληροφόρησης μεταξύ τομέων, θέματος, ιστότοπου, γλώσσας, πολιτισμού είναι μια πιθανή ερευνητική κατεύθυνση για τον εντοπισμό παραπληροφόρησης.

3.2.9.3 Ενσωμάτωση Δεδομένων από Πολλαπλές πηγές

Οι περισσότερες προηγούμενες έρευνες σχετικά με την ανίχνευση παραπληροφόρησης βασίζονται κυρίως στο περιεχόμενο ως εισαγωγή, αλλά πρόσφατες μελέτες έχουν δείξει ότι η ενσωμάτωση πρόσθετων πληροφοριών, όπως το προφίλ του ομιλητή ή τα δεδομένα κοινωνικής σχέσης μπορεί να βελτιώσει περαιτέρω την ακρίβεια των συστημάτων ανίχνευσης. Ως προς την αξιοπιστία του ομιλητή έρευνες όπως των Kirilin και Strube [6] και του Long [7] δείχνουν ότι πολλές φορές η αξιοπιστία και μόνο του ομιλητή οδηγεί σε ασφαλή συμπεράσματα. Επιπλέον, τα δεδομένα κοινωνικών σχέσεων δείχνουν επίσης ότι είναι αποτελεσματικά για τον εντοπισμό παραπληροφόρησης. Ωστόσο, ένα πρόβλημα είναι ότι η εξάρτηση από κρίσεις σε ομιλητές, εκδότες ή κοινωνικά δίκτυα μπορεί να προκαλέσει ορισμένα ρίσκα. Άλλωστε η πιο επικίνδυνη παραπληροφόρηση έρχεται από τα πρόσωπα που εμπιστευόμαστε.

3.2.9.4 Πολυμορφική Αναπαράσταση

Οι πληροφορίες κοινωνικών μέσων συχνά περιέχουν τόσο κείμενο όσο και οπτικό περιεχόμενο (π.χ. εικόνα και βίντεο) και το καθένα έχει τόσο εστιασμένες όσο και συμπληρωματικές πληροφορίες. Επομένως, για την ανίχνευση παραπληροφόρησης, είναι απαραίτητο να χρησιμοποιήσουμε την προσέγγιση πολλαπλών τρόπων ανίχνευσης ενσωματώνοντας τις πληροφορίες κειμένου και εικόνας για να αξιολογήσουμε την αλήθεια των ειδήσεων.

Στο “Multimodal fusion with recurrent neural networks for rumor detection on microblogs” εισάγονται πολυτροπικές πληροφορίες για εντοπισμό ψευδών ειδήσεων για πρώτη φορά μέσα από βαθιά νευρωνικά δίκτυα. Προτείνουν ένα μοντέλο βασισμένο σε RNN(attRNN]) [8] με μηχανισμό προσοχής για την ενσωμάτωση κειμένων και οπτικών πληροφοριών. Το μοντέλο εξάγει τις πολυτροπικές πληροφορίες σε δύο μέρη: το ένα είναι να εξαγάγει τα σημασιολογικά χαρακτηριστικά των εικόνων από το δίκτυο VGG-19 και το άλλο είναι να χρησιμοποιήσει τον μηχανισμό προσοχής για την εξαγωγή βασικών πληροφοριών στο κείμενο και στα πλαίσια.

Στο “Eann: Event adversarial neural networks for multi-modal fake news detection” [9] προτείνεται ένα μοντέλο από άκρο σε άκρο βασισμένο σε εχθρικά δίκτυα. Το κίνητρο είναι ότι πολλά τρέχοντα μοντέλα μαθαίνουν χαρακτηριστικά που σχετίζονται με γεγονότα που είναι δύσκολο να αξιοποιηθούν σε νέα γεγονότα. Κατά συνέπεια, μειώνεται η ικανότητα γενίκευσης του μοντέλου. Αντ 'αυτού, το μοντέλο πρέπει να μάθει περισσότερα ανεξάρτητα από τα γεγονότα χαρακτηριστικά για να βελτιώσει τη δυνατότητα γενίκευσης. Σε αυτήν τη μέθοδο, το TextCNN εφαρμόζεται για την εξαγωγή σημασιολογικών χαρακτηριστικών του κειμένου και το VGG-19 χρησιμοποιείται για την εξαγωγή σημασιολογικών χαρακτηριστικών οπτικού περιεχομένου. Τα πολυτροπικά χαρακτηριστικά στη συνέχεια συνενώνονται για να αντιπροσωπεύουν το περιεχόμενο της παραπληροφόρησης. Από τη μία πλευρά, αυτές οι λειτουργίες χρησιμοποιούνται για να προσδιοριστεί εάν τα νέα είναι ψεύτικα, ενώ από την άλλη πλευρά, χρησιμοποιούνται για να προσδιορίσουν σε ποιο συμβάν σχετίζονται τα νέα.

Οι Khattar et al. [10] υποστηρίζουν ότι η απλή συνένωση κειμένων και οπτικών χαρακτηριστικών δεν επαρκεί για να εκφράσει πλήρως την αλληλεπίδραση και τη σχέση μεταξύ των δύο τροπικών πληροφοριών. Επομένως, προτείνουν μια μέθοδο κωδικοποίησης-αποκωδικοποίησης για τη κατασκευή μίας πολυτροπικής αναπαράστασης χαρακτηριστικών. Σε αυτό το μοντέλο, τα συνενωμένα χαρακτηριστικά των κειμένων και του οπτικού περιεχομένου κωδικοποιούνται ως ενδιάμεση έκφραση και μια ανακατασκευασμένη απώλεια χρησιμοποιείται για να εξασφαλιστεί ότι η κωδικοποιημένη ενδιάμεση έκφραση μπορεί να αποκωδικοποιηθεί πίσω στην αρχική κατάσταση και στη συνέχεια ο ενδιάμεσος φορέας έκφρασης χρησιμοποιείται για ανίχνευση της παραπληροφόρησης.

Ωστόσο, η τρέχουσα πολυτροπική προσέγγιση για την ανίχνευση παραπληροφόρησης αντιμετωπίζει δύο μεγάλες προκλήσεις. Η πρώτη πρόκληση είναι ότι οι υψηλής ποιότητας σχολιασμένες βάσεις δεδομένων πολλαπλών τρόπων παραπληροφόρησης είναι περιορισμένες. Η άλλη πρόκληση είναι ότι παρά τη δημιουργία μεγαλύτερων βάσεων δεδομένων, πρέπει να αναπτυχθούν μη εποπτευόμενες ή ημι-εποπτευόμενες μέθοδοι για την ανίχνευση της παραπληροφόρησης κατά την χρήση δεδομένων χωρίς ετικέτα.

3.2.10 Συμπεράσματα

Η έρευνα των Qi Su [14] παρείχε μια ολοκληρωμένη ανασκόπηση, περίληψη και αξιολόγηση της πρόσφατης έρευνας σχετικά με την ανίχνευση παραπληροφόρησης

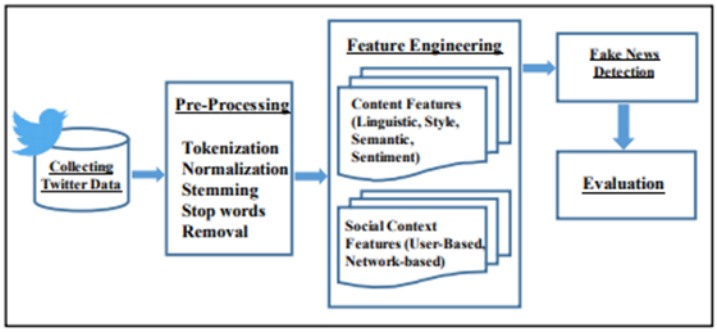
από την οπτική γωνία του NLP. Περιγράφηκαν και συγκρίθηκαν οι υπάρχουσες προσεγγίσεις, τα χαρακτηριστικά, οι επιδόσεις, οι βάσεις δεδομένων, οι προκλήσεις και οι προοπτικές για ανίχνευση παραπληροφόρησης με δύο διαστάσεις.

3.3 Μοντέλο χαρακτηριστικών των μέσων κοινωνικής δικτύωσης για την ανίχνευση ψευδών ειδήσεων στο Twitter (Hussein et al. 2019 [57])

3.3.1 Παρατηρήσεις επί των Απολογισμών

Σε αυτή τη μελέτη προτείνεται ένα μοντέλο ανίχνευσης ψευδών ειδήσεων του Twitter. Η τεχνική ανίχνευσης αυτή, είναι βασισμένη στη χρήση ενός μη εποπτευόμενου συστήματος που βασίζεται σε έναν γράφο για ομαδοποίηση. Στο σύστημα γίνεται ανάλυση τόσο του περιεχόμενου του κείμενου όσο και μια πιο γενική ανάλυση που έχει σχέση με το διαδικτυακό προφίλ του πομπού (ατόμου ή της σελίδας) που κοινοποιεί την είδηση.

Πιο συγκεκριμένα, θα μπορούσαμε να διακρίνουμε πέντε βήματα στο συγκεκριμένο μοντέλο. (Σχήμα 4).



Σχήμα 4: Μοντέλο Ανίχνευσης Ψευδών Ειδήσεων

1. **Συλλογή Δεδομένων:** Το πρώτο βήμα είναι η συλλογή των δεδομένων όπου δημοσιεύσεις του Twitter συλλέγονται και τοποθετούνται σε μια βάση δεδομένων.
2. **Προ-επεξεργασία των Δεδομένων:** Δεύτερο βήμα αποτελεί η προ-επεξεργασία των δεδομένων. Στο στάδιο αυτό περιλαμβάνονται οι εξής διαδικασίες για την διαγραφή δεδομένων που δημιουργούν θόρυβο στην μετέπειτα επεξεργασία. Αρχικά αποσπούμε δεδομένα όπως, το URL, οι ειδικοί χαρακτήρες, κοινότυπες λέξεις και διαφημίσεις (αν υπάρχουν). Μετέπειτα η επεξεργασία του κειμένου περιλαμβάνει την αναγνώριση των λεξικών μονάδων(tokenization) του κειμένου όπου κάθε Tweet χωρίζεται στις μεμονωμένες λέξεις του, στην κανονικοποίηση (normalization) όπου ελέγχεται και τακτοποιείται το ζήτημα των περιττών γραμμάτων (για παράδειγμα κάποιοι χρήστες πληκτρολογούν “thankssss” όπου φυσικά μετατρέπεται σε “thanks”). Στη συνέχεια μέσω της αποκοπής (stemming), αφαιρούνται τα προθέματα ή/και οι καταλήξεις των λέξεων, ενώ στο τέλος αφαιρούνται οι κοινότυπες λέξεις όπως “the”, “a”, “an”, “in” κ.α.
3. **Επεξεργασία Χαρακτηριστικών:** Τα προϊόντα αυτού του σταδίου είναι αυτά που θα επεξεργαστούμε στο τρίτο στάδιο. Εδώ, βρίσκουμε τις δύο κατηγορίες που έχουμε ήδη αναφέρει, τα χαρακτηριστικά του περιεχομένου και τα χαρακτηριστικά του πομπού.

Προσεγγίζοντας αρχικά την επεξεργασία του περιεχομένου συναντάμε πέντε βασικές κατηγορίες χαρακτηριστικών. 1)Γλωσσικά χαρακτηριστικά (Linguistic features). 2) Τρόπος γραφής (Writing style features). 3) Σημασιολογικά χαρακτηριστικά (Semantic features). 4) Χαρακτηριστικά Συναισθημάτων (Sentiment features). 5) Χαρακτηριστικά οπτικού υλικού (Visual-Based features).

Τα Γλωσσικά χαρακτηριστικά δείχνουν τη βασική γλωσσική κατανόηση και τη δομή της πρότασης της είδησης. Στην επεξεργασία του τρόπου γραφής εντοπίζουμε τις διαφορετικές μεθόδους γραφής του εκδότη των ψευδών ειδήσεων. Παρά το γεγονός ότι οι εκδότες αυτοί, προσπαθούν να μιμηθούν τη δομή των αληθινών ειδήσεων, πάντα υπάρχουν γλωσσικές παρατυπίες στις οποίες υποπίπτουν και αποτελούν σημαντική ένδειξη για εμάς. Ωστόσο στον τρόπου γραφής δεν περιέχεται και το συναισθηματικό κομμάτι. Την επεξεργασία αυτού του κομματιού την έχουν υλοποιήσει λίγες ερευνητικές ομάδες, εντούτοις τα αποτελέσματα δείχνουν ότι είναι ένα σημαντικό και αποτελεσματικό εργαλείο στην αντιμετώπιση των ψευδών ειδήσεων. Υπάρχουν πολλές μέθοδοι για τη χρήση της επεξεργασίας των συναισθηματικών χαρακτηριστικών όπως : επικροτούμενη βαθμολογία σθένους διέγερσης (arousal valence dominance score), εντοπισμός χαράς και ανάλυση (happiness detection and analysis), ανάλυση συναισθημάτων(emotion analysis) και ανάλυση πόλωσης και αντοχής (analysis of polarization and strength). Τέλος, τα οπτικά χαρακτηριστικά έχουν να κάνουν με τη πιθανή χρήση εικόνων ή βίντεο που συνοδεύουν μία είδηση και είναι κατασκευασμένα ώστε να την υποστηρίξουν.

Προσεγγίζοντας τώρα την επεξεργασία και ανάλυση του δικτύου του πομπού διακρίνουμε ότι ενώ είναι μία πολύ σημαντική και αξιοσημείωτη τεχνική στη προσπάθεια αντιμετώπισης των ψευδών ειδήσεων, λίγες ερευνητικές ομάδες επιλέγουν να την χρησιμοποιήσουν, προτιμώντας καθαρά την επεξεργασία και ανάλυση του κειμένου. Οι δύο βασικές κατηγορίες της επεξεργασίας του δικτύου είναι η ανάλυση των χαρακτηριστικών χρήστη-πομπού και η ανάλυση του δικτύου αυτού καθαυτό.

Η ανάλυση των χαρακτηριστικών του χρήστη βοηθάει στον εντοπισμό και την καταγραφή των κανονικών ή εικονικών (bots) χρηστών που διαδίδουν ψευδείς ειδήσεις και μπορεί να διασπαστεί σε τρεις κατηγορίες. 1) Ανάλυση του λογαριασμού του χρήστη (user profiling features analysis). 2) Συμπεριφορά και χρονικού πλαισίου κοινοποίησης (posting behavior and temporal features analysis). 3) Ανάλυση αξιοπιστίας (credibility features analysis).

Η ανάλυση του λογαριασμού του χρήστη αποτελείται από την επεξεργασία βασικών χαρακτηριστικών όπως η γλώσσα που χρησιμοποιεί, η γεωγραφική του θέση, η ώρα δημιουργίας του λογαριασμού, η επαλήθευση του λογαριασμού, ο αριθμός των κοινοποιήσεων κ.α.

Η ανάλυση της συμπεριφοράς και του χρονικού πλαισίου κοινοποίησης έχει να κάνει με τον εντοπισμό διαφόρων μοτίβων κατά την κοινοποίηση δύο διαδοχικών δημοσιεύσεων. Ψεύτικοι ή εικονικοί χρήστες (bots, cyborgs) χρησιμοποιούν συνήθων συγκεκριμένα χρονικά μοτίβα μεταξύ των δημοσιεύσεων τους.

Η ανάλυση της αξιοπιστίας των ειδήσεων που έχει δημοσιεύσει ένας χρήστης επικεντρώνεται στην επεξεργασία και ανάλυση των χρηστών που έχουν σχολιάσει ή πατήσει “Μου αρέσει” στην εν λόγω δημοσίευση. Η αξιοπιστία της δημοσίευσης κρίνεται από την αξιοπιστία των χρηστών που αλληλεπιδρούν με αυτή. Αν οι χρήστες είναι αναξιόπιστοι τότε το ίδιο είναι και η δημοσίευση.

Η ανάλυση του δικτύου μελετά την διάδοση των δεδομένων στο κοινωνικό δίκτυο και τη διαδρομή που ακολούθησαν. Η μελέτη των διαδρομών αυτών είναι πολύ χρήσιμη. Δύο ξεχωριστά δίκτυα μπορούν να αντιπροσωπεύσουν την αλληλεπίδραση μεταξύ χρηστών και δημοσιεύσεων. 1) Ομογενή. 2). Ανομοιογενή. Καθένα από αυτά έχει διαφορετικά χαρακτηριστικά.

Στα ομογενή δίκτυα υπάρχουν. 1) Τα φιλικά δίκτυα (Friendship Networks) όπου δείχνουν τη δομή των χρηστών που κάποιος ακολουθεί ή τον ακολουθούν. 2)Τα δίκτυα

διάχυσης (Diffusion Networks) όπου ακολουθούν τη διαδρομή των ειδήσεων και πώς αυτές προωθούνται από χρήστη σε χρήστη με κάθε μία να αποτελεί έναν κόμβο στη διαδρομή μετάδοσης της είδησης στο δίκτυο (retweet). 3) Δίκτυα αξιοπιστίας (Credibility Networks) τα οποία μπορούν να αναπαρασταθούν ως μη κατευθυνόμενοι γράφοι Gc=(V , Ec , s). Όπου V το σύνολο των αναρτήσεων του χρήση με αντίστοιχη βαθμολογία αξιοπιστίας, E η σύνδεση αξιοπιστίας μεταξύ των δύο χρηστών/κόμβων.

Στα ανομοιογενή δίκτυα διακρίνουμε πάλι τρεις κατηγορίες. 1) Δίκτυα Γνώσης (Knowledge Networks) όπου μπορεί κι αυτό να αναπαρασταθεί από έναν γράφο Gk=(I , Ei , R). Όπου I το σύνολο των κόμβων, E η σχέση μεταξύ των κόμβων, R τα σύνολα σχέσεων. 2) Δίκτυα στάσεων (Stance Networks) τα οποία αποτελούνται από ένα σύνολο κόμβων που υποδεικνύουν τις δημοσιεύσεις και ένα σύνολο των ακμών που αντιπροσωπεύουν τα βάρη των στάσεων. 3) Δίκτυα αλληλεπίδρασης (Interaction Networks) όπου μπορεί κι αυτό να αναπαρασταθεί από έναν γράφο Gi=((P,U,V),Ei). Όπου P ο εκδότης της πληροφορίας, U το σύνολο του χρήστη, V η πληροφορία και Ei η αλληλεπίδραση μεταξύ των προηγούμενων στοιχείων.

4. **Εντοπισμός ψευδών ειδήσεων:** Στο τέταρτο βήμα έχουμε τον εντοπισμό πλέον των ψευδών ειδήσεων. Τα αποτελέσματα των εποπτευόμενων αλγορίθμων μηχανικής μάθησης βασίζονται στην ποιότητα των συνόλων δεδομένων με ετικέτα και αν αυτή είναι σωστή. Αυτό απαιτεί μεγάλη προσπάθεια και ανάλυση καθώς πρέπει να οργανωθούν και να καθαριστούν μεγάλοι όγκοι δεδομένων. Η διαδικασία αυτή είναι χρονοβόρα και δύσκολη. Για αυτό το λόγο, οι ενδείξεις κάνουν λόγο για χρήση μόνο μη εποπτευόμενων συστημάτων όπου δεν απαιτείται εκπαίδευση των ταξινομητών. Τα δεδομένα συγκεντρώνονται και κατηγοριοποιούνται βάσει των ομοιοτήτων και των διαφορών τους. Σε αυτή τη προσέγγιση υπάρχουν πέντε τρόποι ανάλυσης. 1) Ανάλυση ομαδοποίησης (clustering analysis). 2) Μη εποπτευόμενη ενσωμάτωση ειδήσεων (unsupervised news embedding). 3) Εξωτερική ανάλυση(outlier analysis). 4) Ανάλυση σημασιολογικής ομοιότητας (semantic similarity analysis). 5) Ανάλυση διάχυσης πληροφοριών (information diffusion analysis) με δύο διαφορετικές προσεγγίσεις: την προσέγγιση που σχετίζεται με γράφο και αυτή που δεν κάνει χρήση γράφου.

5. Αξιολόγηση:

Εν κατακλείδι, σε αυτό το μοντέλο έγινε χρήση γράφων με μη εποπτευόμενη προσέγγιση στον εντοπισμό ψευδών ειδήσεων και υποστηρίζεται ότι η ανάλυση του περιεχόμενου του κείμενου όσο και η ανάλυση του δικτύου δίνουν μεγαλύτερη ακρίβεια στο αποτέλεσμα, ενώ το μοντέλο είναι πρακτικό για μελλοντική χρήση, λόγω της μη εποπτευόμενης φύσης του.

3.4 Μια απλή αλλά δύσκολη να την κερδίσεις βάση για τον διαγωνισμό ανίχνευσης ψευδών ειδήσεων (Riedel et al. 2018 [58])

Σε αυτή τη μελέτη προτείνεται ένα μοντέλο ανίχνευσης ψευδών ειδήσεων το οποίο δημιουργήθηκε με σκοπό τη συμμετοχή στον διαγωνισμό του Fake News Challenge, όπου και κατέλαβε την τρίτη θέση.

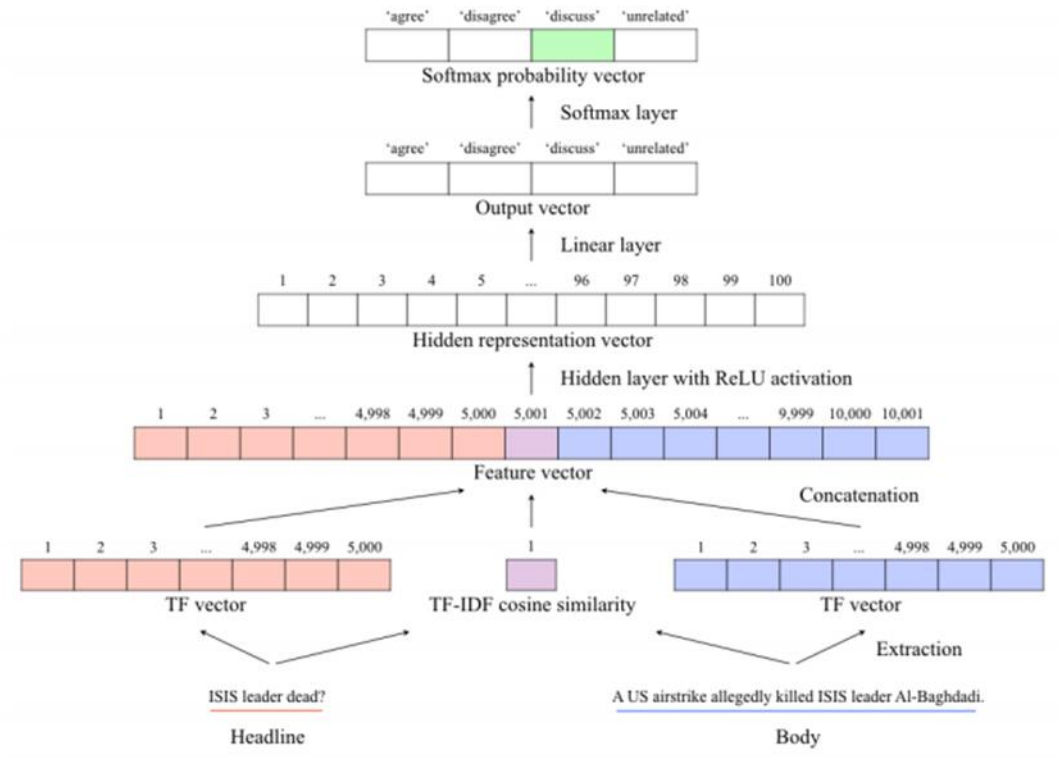
Το σύστημα λαμβάνει τον τίτλο και το κυρίως κείμενο/κορμό ενός άρθρου και παράγει μία ένδειξη εκ των: “συμφωνεί”, “διαφωνεί”, “συζητά” ή “άσχετο”. Αυτή η ένδειξη, παράγεται από ένα ενιαίο σύστημα ανίχνευσης από άκρο σε άκρο, το οποίο αποτελείται τόσο από λεξιλογικά χαρακτηριστικά, όσο και από χαρακτηριστικά ομοιότητας τα οποία διοχετεύονται σε ένα πολυεπίπεδο νευρώνα “Αντίληπτρο” (Multi-Layer Perceptron [MPL]) με ένα κρυφό επίπεδο.

Χρησιμοποιούνται δύο απλές αναπαραστάσεις-σάρωσης-λέξεων (bag-of-words) για τη παρουσίαση των εισερχόμενων κειμένων, καθώς και οι αριθμητικές στατιστικές από τη συχνότητα όρου (TF) και τη συχνότητα όρου-άνιση κατανομή του όρου (TF-IDF) στο κείμενο. (ελληνικές ορολογίες από Γλωσσική Τεχνολογία (upatras.gr)). Έτσι οι αναπαραστάσεις και τα χαρακτηριστικά που προέκυψαν καταλήγουν σε ένα διάνυσμα TF του τίτλου, σε ένα διάνυσμα TF του κορμού και στην ομοιότητα των συνημίτονων μεταξύ των δύο κανονικοποιημένων διανυσμάτων TF-IDF του κορμού και του τίτλου.

Αναγνωρίζονται (tokenization) οι λέξεις του τίτλου και του κορμού του κειμένου και παράγονται τα σχετικά διανύσματα μέσω του scikit-learn.

Χρησιμοποιούνται δύο διαφορετικά “λεξικά”. Για τα TF διανύσματα χρησιμοποιείται ένα “λεξικό” των 5.000 πιο συχνά χρησιμοποιημένων λέξεων για την εκπαίδευση του συστήματος, ενώ αφαιρούνται και οι κοινότυπες λέξεις (stop words). Για τα TF IDF-διανύσματα χρησιμοποιείται και πάλι ένα “λεξικό” των 5.000 πιο συχνά χρησιμοποιημένων λέξεων τόσο για την εκπαίδευση όσο και για την δοκιμή του συστήματος, ενώ αφαιρούνται ξανά οι ίδιες κοινότυπες λέξεις.

Τα TF διανύσματα, τα TF-IDF διανύσματα καθώς και η ομοιότητα των συνημίτονών τους συνδυάζονται σε ένα διάνυσμα χαρακτηριστικών συνολικού μεγέθους 10.001 και στη συνέχεια τροφοδοτούνται στον ταξινομητή.



Σχήμα 5: Σχηματικό Διάγραμμα του Συστήματος

Ο ταξινομητής είναι ένα MLP σύστημα με ένα κρυφό στρώμα 100 μονάδων και ένα μία συνάρτηση ενεργοποίησης softmax στην έξοδο του τελικού γραμμικού στρώματος. Χρησιμοποιήθηκε η συνάρτηση διορθωμένης γραμμικής μονάδας (ReLu) για το κρυφό στρώμα με τα αποτελέσματα που καταγράφονται να είναι αρκετά επιτυχημένα.

Σκοπός της εκπαίδευσης του συστήματος υπήρξε η ελαχιστοποίηση της εγκάρσιας εντροπίας μεταξύ των πιθανοτήτων της συνάρτησης softmax και των αληθινών ετικετών. Έτσι προστέθηκε κανονικοποίηση φθοράς βάρους L2 για τα βάρη του MLP και εφαρμόστηκε αποκλεισμός στην έξοδο και των δύο επιπέδων του αντίληπτρου κατά τη διάρκεια της εκπαίδευσης. Η εκπαίδευση έγινε σε μικρές παρτίδες για ολόκληρο το σύνολο των δεδομένων του εκπαιδευτικού πακέτου με τη χρήση της οπίσθιας διάδοσης (back-propagation), τη μέθοδο βελτιστοποίησης του Adam και τη τεχνική σταδιακής αποκοπής (gradient clipping) με μία καθολική αναλογία αποκοπής.

Παρακάτω παρουσιάζεται ο πίνακας των υπερπαραμέτρων που χρησιμοποιήθηκαν.

Πίνακας 5: Λεπτομέρειες σχετικά με τις υπερπαραμέτρους του συστήματος UCLMR.

Label	Description	Range	Optimised
`lim_unigram`	BOW vocabulary size	1,000 - 10,000	5,000
`hidden_size`	MLP hidden layer size	50 - 600	100
`train_keep_prob`	1 - dropout on layer outputs	0.5 - 1.0	0.6
`l2_alpha`	ℓ_2 regularisation strength	0.1 - 0.0000001	0.0001
`learn_rate`	Adam learning rate	0.1 - 0.001	0.01
`clip_ratio`	Global norm clip ratio	1 - 10	5
`batch_size`	Mini-batch size	250 - 1,000	500
`epochs`	Number of epochs	$\leq 1,000$	90

Τα αποτελέσματα του συστήματος κρίθηκαν βάσει των κανόνων του διαγωνισμού. Το συνολικό ποσοστό επιτυχίας ήταν 81.72% και τα επιμέρους αποτελέσματα παρουσιάζονται αναλυτικά παρακάτω.

Πίνακας 6: Πίνακας σύγχυσης της υποβολής FNC-1 του UCLMR

True \ Pred.	'agree'	'disagree'	'discuss'	'unrelated'	Overall	% Accuracy
'agree'	838	12	939	114	1,903	44.04
'disagree'	179	46	356	116	697	6.60
'discuss'	523	46	3,633	262	4,464	81.38
'unrelated'	53	3	330	17,963	18,349	97.90
Overall	1,593	107	5,258	18,455	25,413	88.46

Η τελική κατάταξη και τα αποτελέσματα του διαγωνισμού είναι τα εξής.

Πίνακας 7: Κορυφαία 10 κορυφαία πλακέτα FNC-1. Υποβολή UCLMR με έντονους χαρακτήρες

Team	% FNC-1 score
SOLAT in the SWEN	82.02
Athene	81.97
UCL Machine Reading	**81.72**
Chips Ahoy!	80.21
CLUlings	79.73
unconscious bias	79.69
OSU	79.65
MITBusters	79.58
DFKI LT	79.56
GTRI - ICL	79.33
Official baseline	75.20

3.5 Αυτόματος εντοπισμός εξαπάτησης: Μέθοδοι εύρεσης ψευδών ειδήσεων (Conroy et al. 2015 [59])

Η εργασία των Conroy et al. 2015[] προτείνει ένα υβριδικό μοντέλο ανίχνευσης ψευδών ειδήσεων, το οποίο περιέχει ποικίλες μεθόδους αξιολόγησης της ακρίβειας μέσω δύο μεγάλων κατηγοριών-προσεγγίσεων. 1) Τη γλωσσική προσέγγιση (linguistic approach), στην οποία το περιεχόμενο των μηνυμάτων εξάγεται και αναλύεται βάσει των γλωσσικών μοτίβων. 2)Την ανάλυση του δικτύου (Network Approach), στην οποία πληροφορίες δικτύου, όπως τα μεταδεδομένα (metadata) του μηνύματος ή τα ερωτήματα του δικτύου μπορούν να αξιοποιηθούν για την αντιμετώπιση της εξαπάτησης.

Στόχος της γλωσσικής προσέγγισης είναι να εντοπίσει “διαρροές” στον λόγο του “ψεύτη” εκδότη ή αλλιώς “προγνωστικά στοιχεία εξαπάτησης” στο περιεχόμενο του μηνύματος. Για την αναπαράσταση του κειμένου χρησιμοποιείται η αναπαράσταση-σάρωσης-λέξεων (bag-of-words) κατά την οποία κάθε λέξη αποτελεί μια οντότητα ίσης σημασίας με τις υπόλοιπες. Με την αναπαράσταση αυτή έχουμε μονάδες λέξεων ή “N-άδες” (n-grams). Βαθύτερη ανάλυση τη δομής του κειμένου επιτυγχάνεται μέσω της πιθανοτικής γραμματικής χωρίς πλαίσιο (Probabilistic context-free grammar [PCFG]). Οι προτάσεις μετατρέπονται σε ένα σύνολο κανόνων (συντακτικό δέντρο [parse tree]) για να περιγραφεί η συντακτική δομή. Τελικά, δημιουργείται ένα συντακτικό δέντρο με μία συγκεκριμένη πιθανότητα ανατεθειμένη [24]. Εργαλεία όπως, ο Stanford Parser (de Marneffe, MacCartney, Manning, 2006; Rahangdale & Agrawa, 2014 [11]) , AutoSlog-TS syntax analyzer (Oraby, Reed, Compton, Riloff, Walker, & Whittaker, 2015 [12]) και άλλα, βοηθούν στην αναπαράσταση.

Για επιπλέον αξιολόγηση, συγκρίνεται το περιεχόμενο του προφίλ του πομπού με τα ανάλογα δεδομένα. Για παράδειγμα υπάρχει η διαίσθηση ότι ένας πομπός ο οποίο δεν έχει εμπειρία για ένα συμβάν ή ένα αντικείμενο, θα συμπεριλάβει στο κείμενό του αντιφάσεις και παραλείψεις γεγονότων τα οποία είναι ήδη παρόντα και έχουν σχολιαστεί από άλλους πομπούς για την ιδια ή παρόμοια θεματολογία. Το εξαγόμενο περιεχόμενο από τις βασικές λέξεις αποτελείται από το ζεύγος “χαρακτηριστικό: περιγραφέας” (attribute: descriptor). Συγκρίνοντας διαφορετικά προφίλ και περιγραφές για τις προσωπικές εμπειρίες των πομπών, αξιολογούμε την εγκυρότητα με μία συνάρτηση των αποτελεσμάτων συμβατότητας. Οι δείκτες συμβατότητας μπορούν να περιγραφούν ως: 1) Συμβατότητα με την ύπαρξη μιας ξεχωριστής όψης (για παράδειγμα υπάρχει ένα μουσείο τέχνης κοντά στο ξενοδοχείο) και 2) Συμβατότητα με την περιγραφή κάποιων γενικών πτυχών, όπως η τοποθεσία ή η υπηρεσία. Τα αποτελέσματα αυτής της μεθόδου φαίνονται να έχουν επιτυχία της τάξεως του 91%. Αν και η παραπάνω τεχνική αποδείχθηκε χρήσιμη και αποτελεσματική στον τομέα των αξιολογήσεων (reviews), φαίνεται να είναι περιορισμένης χρησιμότητας. Υπάρχουν δύο βασικοί περιορισμοί: η δυνατότητα επιτυχούς σύγκρισης μεταξύ χαρακτηριστικών και περιγραφών απαιτεί σημαντική ποσότητα εξορυγμένου περιεχομένου από διάφορα προφίλ χρηστών, καθώς επίσης πρόκληση αποτελεί και η ορθή συσχέτιση των περιγραφών με τα εξαγόμενα χαρακτηριστικά.

Μία περιγραφή του λόγου μπορεί να επιτευχθεί μέσω του Αναλυτικού Πλαισίου της Θεωρίας (RST) της ρητορικής δομής, το οποίο εντοπίζει περιπτώσεις ρητορικών σχέσεων μεταξύ γλωσσικών στοιχείων. Συστηματικές διαφορές μεταξύ παραπλανητικών και αληθινών μηνυμάτων όσον αφορά τη συνοχή και τη δομή τους συνδυάζονται σε ένα Μοντέλο Διανυσματικού Χώρου (Vector Space Model [VSM]), το οποίο αξιολογεί τη θέση κάθε μηνύματος σε μία πολυεπίπεδη δομή ρητορικής (Rhetorical structure theory [RST]) σε σχέση με την απόσταση από την αλήθεια [25].

Τα σύνολα των λέξεων και οι κατηγορίες συχνοτήτων είναι χρήσιμα για την επακόλουθη αυτοματοποιημένη αριθμητική ανάλυση. Πιο συγκεκριμένα χρησιμοποιούνται για την εκπαίδευση των ταξινομητών (classifiers) όπως Support Vector Machines [SVM] και τα μοντέλα αλγορίθμων Naïve Bayes. Οι αλγόριθμοι Naïve Bayes κάνουν ταξινομήσεις με βάση συσσωρευμένες αποδείξεις της συσχέτισης μεταξύ μίας δεδομένης μεταβλητής (π.χ. σύνταξη) και τις άλλες μεταβλητές που υπάρχουν στο μοντέλο. [26]

Η ανάλυση και ταξινόμηση του συναισθήματος βασίζεται στην υποκείμενη διαίσθηση ότι οι ψευδείς πομποί άθελά τους χρησιμοποιούν συναισθηματικό λόγο. Για αυτό το λόγο, χρησιμοποιούνται ήδη γνωστά συντακτικά μοτίβα επιχειρηματολογίας για να εκπαιδεύσουν το μοντέλο να διακρίνει το συναίσθημα. Στη σύγκριση μεταξύ ανθρώπινης κρίσης και SVM οι ταξινομητές έδειξαν 86% ακρίβεια απόδοσης για αρνητικές/παραπλανητικές/spam δημοσιεύσεις. Εντούτοις μπορεί να έχουν περιορισμένη χρήση σε ειδήσεις πραγματικού χρόνου.

Η χρήση δικτύων γνώσης (knowledge networks) αποτελεί σημαντικό βήμα για την επεξεργασία των δεδομένων. Για ορισμένα δεδομένα, οι ψευδείς “πραγματικές δηλώσεις” μπορούν να αντιπροσωπεύουν μία μορφή εξαπάτησης αφού μπορούν να εξαχθούν και να εξεταστούν παράλληλα με εύχρηστες αδιαμφισβήτητες δηλώσεις για τον ήδη γνωστό κόσμο. Αυτή η προσέγγιση αξιοποιεί ένα υπάρχον σύνολο συλλογικής ανθρώπινης γνώσης για να αξιολογήσει την αλήθεια των νέων δηλώσεων. Οι γνώσεις αυτές προέρχονται είτε από ήδη υπάρχοντα δίκτυα γνώσης, είτε από δημόσιες βάσεις δεδομένων, όπως η οντότητα Dbpedia είτε μέσω του Google Relation Extraction Corpus (GREC). Ο έλεγχος γεγονότων μπορεί να μειωθεί αποτελεσματικά σε ένα απλό πρόβλημα ανάλυση δικτύου: τον υπολογισμό της πιο απλής και πιο σύντομης διαδρομής.

Στα ερωτήματα που βασίζονται σε εξαγόμενες δηλώσεις γεγονότων εκχωρείται σημασιολογική εγγύτητα ως συνάρτηση της μεταβατικής σχέσης μεταξύ του υποκειμένου και του κατηγορουμένου μέσω των κόμβων. Όσο πιο κοντά οι κόμβοι, τόσο πιο μεγάλη η πιθανότητα ότι η κατάσταση ενός συγκεκριμένου υποκειμένου-αντικειμένου-κατηγορουμένου να είναι αληθής.

Η εξακρίβωση της ταυτότητας ενός χρήστη στα μέσα κοινωνική δικτύωσης είναι συνυφασμένη με το πόσο έμπιστος είναι ο χρήστης αυτός. Στο Twitter για παράδειγμα, ορισμένα δεδομένα όπως οι υπερσυνδέσεις (hyperlinks) ή τα σχετικά μεταδεδομένα (metadata), μπορούν να επεξεργασθούν για τη διαπίστωση της εγκυρότητας. Η Ανάλυση Κεντρικής Απήχησης (Centering resonance analysis [CRA] [13]) είναι ένας τρόπος για ανάλυση κειμένου του δικτύου και αναπαριστά το περιεχόμενο των μεγάλων κειμένων εντοπίζοντας τις πιο σημαντικές λέξεις που συνδέονται με άλλες μέσα στο δίκτυο.

Εν κατακλείδι, οι γλωσσικές και βασισμένες στο δίκτυο προσεγγίσεις έχουν δείξει αποτελέσματα υψηλής ακρίβειας σε εργασίες ταξινόμησης εντός περιορισμένων τομέων, αλλά αποτελούν μια βασική τυπολογία των διαθέσιμων μεθόδων για περαιτέρω βελτίωση και αξιολόγηση στην προσπάθεια για εντοπισμό των ψευδών ειδήσεων.

3.6 Ένα εργαλείο για τον εντοπισμό ψευδών ειδήσεων (Al Asaad et al. 2016 [60])

Το συγκεκριμένο paper περιγράφει μία προσπάθεια επαλήθευσης της αξιοπιστίας ειδησεογραφικών άρθρων του διαδικτύου με βάση τα χαρακτηριστικά τους. Για την επίτευξη του στόχου αυτού εφαρμόστηκε ένας αλγόριθμος που συνδυάζει μηχανική μάθηση (machine learning) και διάφορες τεχνικές μεθόδους ταξινόμησης (classification methods) με μοντέλα κειμένου (text models). Στο εργαλείο που δημιουργήθηκε, εφαρμόστηκε εποπτευόμενη μάθηση (supervised learning) ως κύρια τεχνική μηχανικής μάθησης. Ένα μεγάλο σύνολο δεδομένων αληθών και ψευδών ειδήσεων χρησιμοποιήθηκε, επίσης, για την εκπαίδευση του μοντέλου (machine learning model) με τη βοήθεια της βιβλιοθήκης Scikit-learn [18] της Python. Η συγκεκριμένη βιβλιοθήκη είναι πολύ εξυπηρετική καθώς εμπεριέχει ενσωματωμένες μεθόδους που υλοποιούν διάφορες προσεγγίσεις ταξινόμησης που θα δούμε στη συνέχεια, και για το λόγο αυτό η Python αποτελεί την καλύτερη επιλογή για την υλοποίηση του εργαλείου.

3.6.1 Μεθοδολογία δημιουργίας εργαλείου & Τεχνικές που χρησιμοποιήθηκαν

3.6.1.1 Εποπτευόμενη Μάθηση (Supervised Learning)

Στην Εποπτευόμενη Μάθηση [19] παρέχεται ένα σύνολο δεδομένων που ονομάζεται «Δεδομένα Εκπαίδευσης» (training data) που αποτελείται από δεδομένα εισαγωγής (input data) και δεδομένα στόχου (target data), δηλαδή τις σωστές απαντήσεις που θα πρέπει ο αλγόριθμος κάθε φορά να παράγει. Τα βήματα που ακολουθούνται στη διαδικασίας της μηχανικής μάθησης είναι:

1. Συλλογή και προετοιμασία δεδομένων
2. Επιλογή και προσδιορισμός απαραίτητων χαρακτηριστικών για την εξέταση του προβλήματος
3. Επιλογή κατάλληλου αλγορίθμου
4. Επιλογή παραμέτρων και μοντέλων για την καλύτερη δυνατή απόδοση
5. Με βάση τα παραπάνω, η εκπαίδευση χρησιμοποιεί υπολογιστικούς πόρους ώστε να φτιάξει ένα μοντέλο που να προβλέπει τα αποτελέσματα (outputs) νέων δεδομένων
6. Αξιολόγηση ακρίβειας του μοντέλου

Πιθανοτική Ταξινόμηση (Naive Bayes)

Ο ταξινομητής Naive Bayes [20] πρόκειται για έναν πιθανοτικό ταξινομητή που εφαρμόζει το θεώρημα του Bayes στον κανόνα της απόφασης του, θεωρώντας πως τα χαρακτηριστικά έχουν ισχυρή ανεξαρτησία μεταξύ τους. Η ταξινόμηση αυτή είναι πολύ χρήσιμη για προβλήματα ταξινόμησης κειμένου, για το λόγο ότι τα μοντέλα που αναφέρονται στη συνέχεια μπορούν να εκπαιδευτούν με ακρίβεια στην εποπτευόμενη μαθησιακή υπόθεση.

Γραμμική Ταξινόμηση Support Vector Machine

Ο ταξινομητής Linear Support Vector Machine [21] εκτελεί την ταξινόμηση χρησιμοποιώντας την τιμή ενός γραμμικού συνδυασμού μεταξύ των χαρακτηριστικών μίας δεδομένης παρατήρησης. Γενικότερα, η γραμμική ταξινόμηση λειτουργεί καλά για προβλήματα με μεγάλο αριθμό χαρακτηριστικών.

3.6.1.2 Μοντέλα Ταξινόμησης

Τσάντα Λέξεων (Bag-of-words)

Το μοντέλο αυτό αναλύει το κείμενο όλων των εγγράφων και τα μετατρέπει σε μια μορφή Τσάντας Λέξεων [22]. Για παράδειγμα, για πάνω από ένα κείμενο εγγράφου (>1) θα έχουμε μία τσάντα λέξεων που θα περιέχει όλες τις ξεχωριστές λέξεις όλων των κειμένων μέσα σε μία «τσάντα», καθώς και θα αγνοεί τη σειρά των λέξεων και των γραμματικών.

Συχνότητα Όρων - Αντίστροφη Συχνότητα στο Έγγραφο (TF-IDF)[3]

Χρησιμοποιεί τη συχνότητα αριθμού εμφάνισης (count) ενός όρου μέσα στο έγγραφο, επιπλέον μαζί με τη συχνότητα αριθμού εμφάνισής του σε ένα σύνολο εγγράφων. Το μοντέλο συχνοτήτων υπολογίζει ουσιαστικά τη σημαντικότητα και την αξία κάθε όρου στο έγγραφο, ή με άλλα λόγια, υπολογίζει το βάρος (weight) κάθε όρου σε αυτό.

Συχνότητα Bigram

Το bigram [23] είναι μια ακολουθία δύο παρακείμενων στοιχείων από μια σειρά συμβόλων, τα οποία είναι συνήθως γράμματα, συλλαβές ή λέξεις. Η κατανομή συχνότητας κάθε bigram σε μια συμβολοσειρά χρησιμοποιείται συνήθως για απλή στατιστική ανάλυση του κειμένου σε πολλές εφαρμογές, συμπεριλαμβανομένου και του αλγορίθμου που περιγράφεται σε αυτό το paper.

3.6.1.3 Αλγόριθμοι

Στη συνέχεια θα δούμε τους αλγόριθμους που χρησιμοποιήθηκαν για την υλοποίηση του εργαλείου, με βασικό αλγόριθμο τον Αλγόριθμο 1, όπου για κάθε εισαγόμενο link, εμφανίζει πληροφορίες σχετικά με το περιεχόμενο, τον τίτλο και τον συντάκτη καλώντας άλλους υπο-αλγόριθμους.

Ανάλυση συνδέσμου (Web link parsing)

Για την ανάλυση (parsing) συνδέσμων, ο παρακάτω αλγόριθμος χρησιμοποιεί τη βιβλιοθήκη Beautiful Soup 4 για να εξάγει τον πηγαίο κώδικα του άρθρου και του περιεχομένου του.
Για την εύρεση της ημ/νίας δημοσίευσης [4] δε χρησιμοποιείται η παραπάνω βιβλιοθήκη, καθώς δεν υπάρχει συνήθης τρόπος αναπαράστασης ημ/νίας. Αντ'αυτού μεταφέρονται τα html bytes που εξήχθησαν με τη Beautiful Soup 4 σε μία συμβολοσειρά (string), όπου με τη βοήθεια ενός RegEx δύναται να βρεθεί η ζητούμενη ημ/νία.
Για την εύρεση του συντάκτη [5] γίνεται αναζήτηση στον πηγαίο κώδικα αφού έχει μετατραπεί σε string, όπου με τη βοήθεια του Google Cloud Natural Language Processing API εξάγονται όλες οντότητες τύπου *Person*. Μια απλή αναζήτηση της λέξης «author» θα μας δώσει το όνομα του συντάκτη, ενώ αν δεν υπάρχει, το όνομά του θα το πάρει το όνομα της ιστοσελίδας.

Αλγόριθμος 1 - Ανιχνευτής Ψευδών Ειδήσεων (Βασικός Αλγόριθμος)

Input: Link κάποιου ειδησεογραφικού άρθρου
Output: 1. Συντάκτης: Όνομα/Ιστότοπος
2. Τίτλος: Clickbait/Non-Clickbait
3. Ημ/νία: Τον παρομοιότερο τίτλο είδησης κοντά στην ημ/νία δημοσίευσης
4. Περιεχόμενο: Αληθές/Ψευδές

Βήμα 1. Επαλήθευση πως το link είναι έμπιστο
Βήμα 2. **Αν** το link είναι έμπιστο, **τότε** ΠΗΓΑΙΝΕ στο Βήμα 3, **αλλιώς** εμφάνισε σχετικό μήνυμα
Βήμα 3. Ανάλυση (Parsing) του πηγαίου κώδικα HTML χρησ. τον Αλγόριθμο 2 και εξαγωγή των ακόλουθων πληροφοριών:

- Συντάκτης, ημ/νία δημοσίευσης, τίτλος, περιεχόμενο (κείμενο)

Βήμα 4^{α}. Ανάλυση του ονόματος του συντάκτη. **Αν** το όνομα λείπει, **τότε** θεώρησε το όνομα του ιστότοπου ως το όνομα συντάκτη
Βήμα 4^{β}. Ανάλυση του τίτλου και του περιεχομένου κειμένου χρησ. τον Αλγόριθμο 3
Βήμα 4^{γ}. Ανάλυση της ημ/νίας χρησ. τον Αλγόριθμο 4

Αλγόριθμος 2 – Ανάλυση (Parsing) ιστοσελίδας

Input: Link κάποιου ειδησεογραφικού άρθρου
Output: 1. Συντάκτης: Όνομα/Ιστότοπος
2. Τίτλος: Clickbait/Non-Clickbait
3. Ημ/νία: Τον παρομοιότερο τίτλο είδησης κοντά στην ημ/νία δημοσίευσης
4. Περιεχόμενο: Αληθές/Ψευδές

Βήμα 1. Λήψη πηγαίου κώδικα σελίδας που φιλοξενεί το ειδησεογραφικό άρθρο
Βήμα 2. Εξαγωγή τίτλου
Βήμα 3. Εξαγωγή περιεχομένου (κειμένου)
Βήμα 4. Εξαγωγή ημ/νίας [4]

Βήμα 5. Εξαγωγή ονόματος εκδότη [5]. **Αν** υπάρχει, **τότε** εμφάνιση ονόματος, **αλλιώς** εμφάνιση ονόματος ιστοσελίδας

Μηχανική Μάθηση (Machine Learning)

Αφού γίνει η ανάλυση (parsing) του συνδέσμου που οδηγεί στο επιλεγμένο απ'τον χρήστη άρθρο, έρχεται η σειρά της μηχανική μάθησης, η οποία χρησιμοποιείται για να λύσει προβλήματα ταξινόμησης. Όπως αναφέρθηκε παραπάνω, γίνεται χρήση της «εποπτευόμενης μάθησης» ως κύρια τεχνική. Η διαδικασία μηχανικής μάθησης της εφαρμογής αποτελείται από τα παρακάτω βήματα:

1. Όσον αφορά τον εντοπισμό περιεχομένου, χρησιμοποιήθηκε το «fake real news.csv» από το σύνδεσμο https://github.com/GeorgeMcIntire/ το οποίο είναι «καθαρό» και έτοιμο να χρησιμοποιηθεί για εξαγωγή δεδομένων. Το παραπάνω αρχείο περιέχει 6.335 ειδησεογραφικά άρθρα, από τα οποία τα 3.164 είναι ψευδή και τα 3.171 είναι αληθή. Επίσης, ως δεύτερο σύνολο δεδομένων, χρησιμοποιήθηκε το «log 32k.csv» το οποίο περιέχει 32.000 τίτλους άρθρων, από τους οποίους οι 15.999 είναι clickbait, ενώ οι 16.001 δεν είναι clickbait.
2. Έγινε έλεγχος για το αν οι λέξεις (tokens) στα άρθρα και τους τίτλους έχουν σημαντική επίδραση σχετικά με το αν το άρθρο είναι αληθής ή ψευδής είδηση, καθώς και αν ο τίτλος είναι clickbait ή όχι.
3. Επιλέχτηκαν τα μοντέλα αναπαράστασης κειμένου που αναφέρονται στο [2].
4. Εφαρμόστηκαν τα παραπάνω μοντέλα με δύο κύριες προσεγγίσεις ταξινόμησης όπως αναφέρεται στο [1].
5. Αξιολογήθηκαν οι ταξινομητές χρησιμοποιώντας δεδομένα δοκιμής (test data) από τα εισαγόμενα (imported) datasets. Τα δεδομένα δοκιμής δε χρησιμοποιήθηκαν κατά την εκπαίδευση των ταξινομητών.

Ομοιότητα Συνημιτόνων (Cosine Similarity)

Χρησιμοποιήθηκε η ονομαζόμενη προσέγγιση για την επαλήθευση της ομοιότητας του επιλεγμένου άρθρου σε σχέση με μία λίστα από άρθρα. Η λίστα άρθρων εξήχθη χρησιμοποιώντας το API https://www.newsapi.org. Η λίστα των τίτλων βασίζεται στην ημερομηνία δημοσίευσης που εξήχθη από την ιστοσελίδα του ειδησεογραφικού άρθρου. Με άλλα λόγια, έχουμε μία λίστα τίτλων όλων των ειδήσεων που συνέβησαν την ίδια μέρα που το επιλεγμένο (input) άρθρο δημοσιεύτηκε.

Για να εκτελεστεί ο αλγόριθμος Ομοιότητας Συνημιτόνων στους τίτλους, θέσαμε τον τίτλο του άρθρου που έχουμε ως input, καθώς και όλους τους άλλους τίτλους, ως διανύσματα ΣΟΑΣΕ (TFIDF) [3]. Αυτό σημαίνει πως κάθε τίτλος παριστάνεται ως διάνυσμα βαρών (vector of weights) για κάθε λέξη, άρα δύναται να εκτελεστεί ο αλγόριθμος Ομοιότητας Συνημιτόνων μεταξύ του τίτλου του άρθρου που έχουμε ως input και κάθε άλλου τίτλου στη λίστα. Στο πέρας αυτού, θα καταγραφεί η υψηλότερη βαθμολογία ομοιότητας και θα εμφανιστεί στον χρήστη ο σχετικός τίτλος με αυτή.

Αλγόριθμος 3 – Ανάλυση Περιεχομένου/Τίτλου

Input: Περιεχόμενο & Τίτλος
Output: Ψευδές/Αληθές & Clickbait/Non-Clickbait

Βήμα 1. Διάβασμα δεδομένων με ψευδείς ή αληθείς ειδήσεις και clickbait ή non-clickbait τίτλων και χωρισμός σε σετ δοκιμής και σετ εκπαίδευσης.
Βήμα 2. Δημιουργία μοντέλου αναπαράστασης κειμένου [2] από το σετ εκπαίδευσης και δεδομένα δοκιμής (train set & test data).
Βήμα 3. Εισαγωγή σετ εκπαίδευσης στους ταξινομητές μηχανικής μάθησης [1].
Βήμα 4. Πρόβλεψη αποτελέσματος περιεχομένου (ψευδές ή αληθές) και τίτλου (clickbait ή όχι) χρησ. τους παραπάνω ταξινομητές.
Βήμα 5: Χρήση δεδομένων δοκιμής (test data) από το Βήμα 1 για τον υπολογισμό της βαθμολογίας ακρίβειας για τους ταξινομητές.

Αλγόριθμος 4 - Ανάλυση Ημ/νίας Δημοσίευσης

Input: Ημ/νία δημοσίεσυης
Output: Άρθρο με την υψηλότερη βαθμολογία ομοιότητας

Βήμα 1. **Αν** η ημ/νία είναι της μορφής ISO 8601 (yyyy-mm-dd or yyyy-mm-ddThh:mm:ss), **τότε** πήγαινε στο Βήμα 3, **αλλιώς**, πήγαινε στο Βήμα 2.
Βήμα 2. Μετατροπή την ημ/νίας στην παραπάνω μορφή.
Βήμα 3. Λήψη λίστας τίτλων ειδήσεων που συνέβησαν την αντίστοιχη ημ/νία στέλνοντας ένα web request στο https://www. newsapi.org.
Βήμα 4. Δημιουργία μοντέλου αναπαράστασης κειμένου (TFIDF) από τον τίτλο που εξήχθη και τη λίστα των ληφθέντων τίτλων από το newsapi.
Βήμα 5. Χρήση της Ομοιότητα Συνημιτόνων για να βρεθεί ο τίτλος που μοιάζει περισσότερο με τον τίτλο του άρθρου.
Βήμα 6. Εμφάνισε τον τίτλο που εξήχθη, τον πιο παρόμοιο τίτλο και τη βαθμολογία ομοιότητας.

3.6.2 Συμπεράσματα

Παρακάτω θα αναλύσουμε τα αποτελέσματα συνδυάζοντας κάθε ταξινομητή με κάθε μοντέλο αναπαράστασης κειμένου. Για τον υπολογισμό της ακρίβειας κάθε ταξινομητή, θα χρησιμοποιηθούν μετρικές της βιβλιοθήκης Scikit-learn και τα αποτελέσματα θα παρασταθούν σε μία χαρακτηριστική καμπύλη λειτουργίας δέκτη ή καμπύλη ROC.

Χρησιμοποιήθηκε ένα σύνολο δεδομένων με 6.335 άρθρα ειδήσεων και εκτελέστηκε τυχαίος διαχωρισμός σε αυτά τα δεδομένα σε δύο μέρη: δεδομένα εκπαίδευσης και δεδομένα δοκιμής. Τα δεδομένα εκπαίδευσης αποτελούνται από το 66,6% των δεδομένων και τα δεδομένα δοκιμών αποτελούνται από το 33,3%.

Παρακάτω χρησιμοποιούνται οι ακόλουθες συντομογραφίες:

1) Bag-of-Words: BoW

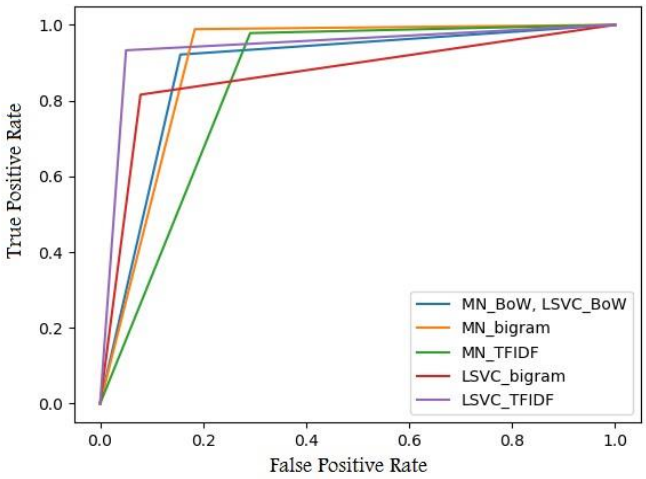
2) Bigram: bigram

3) Term Frequency-Inverse Document Frequency: TFIDF

4) Multinomial Naive Bayes: MN

5) Linear Support Vector Classifier: LSV

Αποτελέσματα εντοπισμού ψευδού περιεχομένου



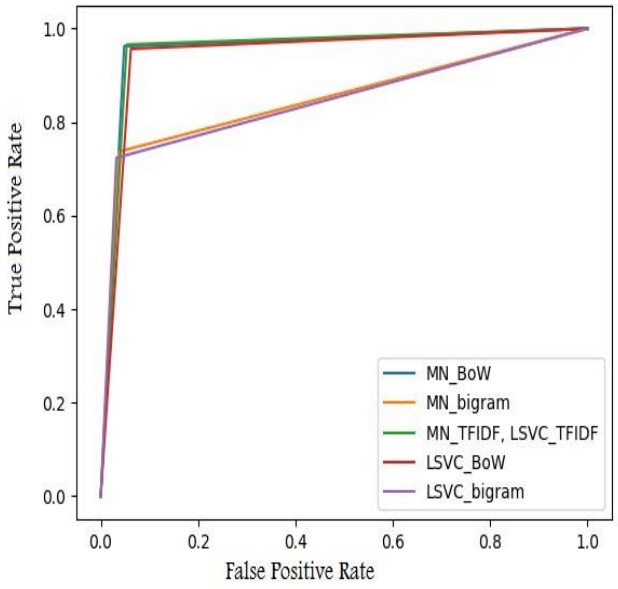
Σχήμα 6: ROC αναπαράσταση εντοπισμού ψευδούς περιεχομένου

Πίνακας 8: Βαθμολογία εντοπισμού περιεχομένου

	BoW	TFIDF	Bigram
MN	0.883	0.845	0.903
LSVC	0.883	0.941	0.868

Από τον παραπάνω πίνακα βλέπουμε πως ο LSVC ταξινομητής είχε την καλύτερη απόδοση μαζί με το μοντέλο TFIDF. Την ίδια στιγμή, τη χειρότερη απόδοση την είχε ο συνδυασμός του MN με το μοντέλο TFIDF.

Αποτελέσματα εντοπισμού clickbait τίτλων



Σχήμα 7: ROC αναπαράσταση εντοπισμού clickbait τίτλων

Πίνακας 9: Βαθμολογία εντοπισμού clickbait τίτλων

	BoW	TF-IDF	Bigram
MN	0.957	0.956	0.849
LSVC	0.947	0.956	0.845

Από τον πίνακα φαίνεται ότι οι βαθμολογίες ακρίβειας και των δύο ταξινομητών με τα μοντέλα BoW και TFIDF είναι οι υψηλότερες. Αυτό συμβαίνει καθώς το σύνολο δεδομένων που χρησιμοποιήθηκε έχει πολύ μεγάλο αριθμό τίτλων, συνεπώς αυτό συμβάλει στην καλύτερη εκπαίδευση και προβλεπτική απόδοση των μοντέλων.

3.7 Ο δρόμος προς τον αυτόματο έλεγχο γεγονότων: Εντοπίζοντας πραγματικούς ισχυρισμούς άξιους ελέγχου από το Claimbuster (Hassan et al. 2017 [61])

Το paper αυτό ασχολείται με τη δημιουργία ενός εργαλείου ονόματι «ClaimBuster», δηλαδή με την αυτόματη εξακρίβωση ισχυρισμών σε πολιτικές διαμάχες όσον αφορά την εγκυρότητά τους ή το αντίθετο. Πρόκειται για ένα πρότζεκτ που η ανάπτυξή του ξεκίνησε το 2014 και εξελίσσεται μέρα με τη μέρα. Το πρότζεκτ αυτό χρησιμοποιεί επεξεργασία φυσικής γλώσσας (NLP), την τεχνική της επιβλεπόμενης μάθησης (supervised learning) και τεχνικές ερωτημάτων σε βάση δεδομένων (database query techniques) για να πετύχει το στόχο δημιουργίας του. Η ιδέα πίσω από αυτό το εργαλείο είναι πως στις μέρες μας είναι αδύνατο να ανταπεξέλθουν ανθρώπινοι ελεγκτές γεγονότων με τα μεγάλα ποσοστά παραπληροφόρησης και ψευδών ειδήσεων, καθώς και με την ταχύτητα που εξαπλώνονται· Έτσι, δημιουργείται ένα παράθυρο ευκαιρίας για την ανάπτυξη ενός ολοκληρωμένου εργαλείου που να κάνει αυτόματο έλεγχο ισχυρισμών και γεγονότων, που όμοιό του δεν έχει υπάρξει μέχρι και τη στιγμή συγγραφής του εν λόγω paper.

Το εργαλείο δύναται να παρακολουθεί ζωντανές συζητήσεις (π.χ. συνεντεύξεις, ομιλίες και ντιμπέιτ), τα μέσα κοινωνικής δικτύωσης και τις ειδήσεις για να εντοπίσει ισχυρισμούς, όπου στη συνέχεια θα διασταυρώνει με τα στοιχεία ενός επιθεωρημένου συνόλου δεδομένων (curated repository) από επιβεβαιωμένους έγκυρους ισχυρισμούς που έχουν επιμεληθεί επαγγελματίες του χώρου. Κατόπιν, εμφανίζει τα αποτελέσματα κατευθείαν στον θεατή ή αναγνώστη (για παράδειγμα, εμφανίζει μία pop-up προειδοποίηση αν ο προεδρικός υποψήφιος των ΗΠΑ έχει πει κάτι ψευδές στη διάρκεια μίας ζωντανής μετάδοσης).

Στη συνέχεια, θα αναλύσουμε τη μεθοδολογία δημιουργίας του ClaimBuster, καθώς και τεχνικές που χρησιμοποιήθηκαν:

3.7.1 Ταξινόμηση και Κατάταξη

Το πρόβλημα της ανίχνευσης έγκυρων ισχυρισμών μοντελοποιήθηκε ως έργο ταξινόμησης και κατάταξης και χρησιμοποιήθηκε η τεχνική της επιτηρούμενης μάθησης (supervised learning) για την αντιμετώπισή του. Στο πλαίσιο αυτού, δημιουργήθηκε ένα ενδεικτικό σύνολο δεδομένων από προτάσεις που ειπώθηκαν από τους προεδρικούς υποψηφίους κατά τη διάρκεια των ντιμπέιτ των προηγούμενων εκλογών 2016 στις ΗΠΑ. Σε κάθε πρόταση δίνεται ένας εκ των τριών πιθανών χαρακτηρισμών:

- Μη πραγματικός ισχυρισμός (ΜΠΙ) - Non-Factual Sentence (NFS)

Οι υποκειμενικές προτάσεις (απόψεις, πεποιθήσεις, δηλώσεις) και πολλές ερωτήσεις εμπίπτουν σε αυτήν την κατηγορία. Αυτές οι προτάσεις δεν περιέχουν κανένα ισχυρισμό. Για παράδειγμα:

- Αλλά νομίζω ότι είναι καιρός να μιλήσουμε για το μέλλον.
- Θυμάσαι την τελευταία φορά που το είπες αυτό;

- Ασήμαντος ισχυρισμός (ΑΙ) - Unimportant Factual Sentence (UFS)

Αυτοί είναι πραγματικοί ισχυρισμοί, αλλά δεν είναι άξιοι ελέγχου εγκυρότητας. Ο περισσότερος κόσμος δεν ενδιαφέρεται να μάθει αν αυτές οι προτάσεις είναι αληθείς ή ψευδείς. Οι ελεγκτικοί παράγοντες δε θεωρούν αυτές τις προτάσεις τόσο σημαντικές. Μερικά παραδείγματα είναι:

- Η επόμενη Τρίτη είναι η ημέρα των εκλογών.
- Πριν από δύο μέρες φάγαμε μεσημεριανό σε εστιατόριο.

- Άξιος-ελέγχου ισχυρισμός (AEI) – Check-worthy Factual Sentence (CFS)

Περιέχουν ισχυρισμούς που το ευρύ κοινό ενδιαφέρεται να μάθει εάν είναι αληθείς. Οι δημοσιογράφοι αναζητούν αυτόν τον τύπο ισχυρισμών για έλεγχο γεγονότων. Μερικά παραδείγματα είναι:

- Ψήφισε κατά του πρώτου Πολέμου του Κόλπου.
- Πάνω από 1,25 εκατομμύρια Αμερικάνοι είναι θετικοί στον HIV.

Δεδομένης μια πρότασης, ο στόχος του ClaimBuster είναι να της **αποδώσει ένα σκορ** που να αντικατοπτρίζει τον βαθμό που η πρόταση ανήκει στο AEI (CFS). Οι προτάσεις που ανήκουν στην κατηγορία του CFS θεωρούνται ως θετικές περιπτώσεις, ενώ αυτές που ανήκουν στις NFS και UFS θεωρούνται ως αρνητικές. Για την ταξινόμηση μίας δεδομένης πρότασης x, υπολογίζουμε την εκ των υστέρων πιθανότητα (posterior probability) της CFS πρότασης χρησιμοποιώντας επιτηρούμενη μάθηση:

$$\textbf{score}\ (x) = P\ (\textbf{class} = \textbf{CFS}|x)$$

3.7.2 Χαρακτηρισμός Δεδομένων

Είναι αδήριτη ανάγκη να συγκεντρωθεί ένα σύνολο δεδομένων όπου για κάθε πρόταση που ειπώθηκε στα ντιμπέιτ των εκλογών του 2016 στις ΗΠΑ να καταγραφεί εάν επρόκειτο για NFS, UFS ή CFS. Για την υποβοήθηση αυτού, μαζεύτηκαν παροδικά όλα τα αντίγραφα όλων των αμερικανικών ντιμπέιτ που συνέβησαν μεταξύ 1960 – 2012, συνολικά δηλαδή από 30 ντιμπέιτ σε 11 χρονιές εκλογών. Υπάρχουν 28.029 προτάσεις στα αντίγραφα αυτά, για τις οποίες χρησιμοποιώντας κανόνες ανάλυσης (parsing rules) και ανθρώπινο σχολιασμό[1] (human annotation) εντοπίστηκε ο ομιλητής κάθε πρότασης. 23.075 είναι οι προτάσεις προεδρικών υποψηφίων και 4,815 αυτές των συντονιστών. Από όλες αυτές, συγκεντρώθηκαν 20.788 προτάσεις που ειπώθηκαν από τους υποψηφίους αποτελούμενες από τουλάχιστον 5 λέξεις. Στη συνέχεια, αναπτύχθηκε μία ιστοσελίδα (https://idir.uta.edu/classifyfact_survey) με στόχο τη συλλογή δεδομένων για το ground truth των χαρακτηρισμών των προτάσεων.

Με την ορολογία «ανθρώπινος σχολιασμός»[1] αναφερόμαστε στο σχολιασμό και χαρακτηρισμό προτάσεων από συμμετέχοντες του πρότζεκτ. Η πρόσληψη και η εκπαίδευσή τους έγινε ως εξής: Επιλέχτηκαν συμμετέχοντες επί πληρωμή (κυρίως φοιτητές, καθηγητές και δημοσιογράφοι που γνωρίζουν από αμερικανική πολιτική) και εκπαιδεύτηκαν με τη χρήση 30 προτάσεων. Αρχικά, ο κάθε συμμετέχοντας πρέπει να διαβάσει και να επισημάνει κάθε μία από αυτές τις 30 προτάσεις ως NFS, UFS ή CFS αντίστοιχα. Αφού χαρακτηρίσουν μία πρόταση, το site τους αποκαλύπτει το *ground truth* της και το επεξηγεί. Επιπλέον, διοργανώθηκαν πολλά επιτόπια εκπαιδευτικά εργαστήρια για τους συμμετέχοντες που ήταν διαθέσιμοι. Κατά τη διάρκεια κάθε εργαστηρίου, τουλάχιστον δύο εμπειρογνώμονες ήταν παρόντες για να ξεκαθαρίσουν

τυχόν αμφιβολίες που μπορεί να έχουν οι συμμετέχοντες σχετικά με τον ιστότοπο και τη διαδικασία συλλογής δεδομένων. Μέσα από συνεντεύξεις με τους συμμετέχοντες, παρατηρήθηκε ότι αυτά τα μέτρα κατάρτισης ήταν σημαντικά για να βοηθηθούν οι συμμετέχοντες ώστε να επιτύχουν υψηλή ποιότητα εργασίας.

3.7.3 Διασφάλιση Ποιότητας

Για τον εντοπισμό συμμετεχόντων χαμηλής ποιότητας επιλέχτηκαν 1.032 προτάσεις (731 NFS, 63 UFS, 238 CFS). Τρεις ειδικοί συμφώνησαν για τους χαρακτηρισμούς αυτών των προτάσεων ελέγχου. Κατά μέσο όρο, μία από τις δέκα προτάσεις που δόθηκαν σε έναν συμμετέχοντα (χωρίς να ενημερωθεί ο συμμετέχοντας) επιλέχθηκε τυχαία για να είναι μια πρόταση ελέγχου. Πρώτον, ένας τυχαίος αριθμός αποφασίζει τον τύπο (NFS, UFS, CFS) της πρότασης. Στη συνέχεια, η πρόταση επιλέγεται τυχαία από το σύνολο των προτάσεων αυτού του συγκεκριμένου τύπου. Ο βαθμός συμφωνίας για τον έλεγχο προτάσεων μεταξύ ενός συμμετέχοντα και των τριών ειδικών είναι ένας από τους παράγοντες για τη μέτρηση της ποιότητας του συμμετέχοντα· όταν ο χαρακτηρισμός ενός συμμετέχοντα ταιριάζει με τον χαρακτηρισμό των ειδικών, ανταμείβεται με μερικούς πόντους, ενώ αν δεν ταιριάζει, τιμωρείται. Δεν έχουν όλα τα λάθη ίδια βαρύτητα. Για παράδειγμα, χαρακτηρίζοντας μία NFS πρόταση ως CFS είναι μεγαλύτερο λάθος από τον χαρακτηρισμό μίας UFS πρότασης ως CFS. Επιπλέον, ορίστηκαν μερικά βάρη (weights) για διαφορετικά είδη λαθών και εντάχθηκαν στο μέτρο ποιότητας:

Έστω SS(p) ένα σετ από προτάσεις που εμφανίζονται χαρακτηρισμένες από τον συμμετέχοντα p, τότε η ποιότητα του χαρακτηρισμού p(LQp) είναι:

$$LQ_p = \frac{\sum_{s \in SS(p)} \gamma^{lt}}{|SS(p)|}$$

γ^{lt} είναι το βάρος (weight)
οι *p* χαρακτήρισαν την πρόταση *s* ως *l* και οι ειδικοί ως *t*.

Συνολικά, 374 συμμετέχοντας έχουν συνεισφέρει στη συλλογή δεδομένων μέχρι τη στιγμή συγγραφής του paper, εκ των οποίων οι 86 θεωρούνται υψηλής ποιότητας συμμετέχοντες.

Συλλέχτηκαν 76.552 χαρακτηρισμοί προτάσεων, από τους οποίους 52.533 (68%) προέρχονται από τους συμμετέχοντες υψηλής ποιότητας. Επίσης, υπάρχουν 20.617 προτάσεις που ικανοποιούν την κατάσταση διακοπής. Ως κατάσταση διακοπής θεωρείται μία πρόταση *s*, η οποία δε θα επιλεχθεί για περαιτέρω χαρακτηρισμό:

$$\mathbf{x \in \{NFS, UFS, CFS\},\ 2s_x > (s_{NFS} + s_{UFS} + s_{CFS})/2}$$

όπου s_x συμβολίζει τον αριθμό υψηλής ποιότητας χαρακτηρισμών τύπου X στην πρόταση s

Ο παρακάτω πίνακας δείχνει την κατανομή των ταξινομήσεων στις προτάσεις:

Πίνακας 10: Κατανομή ταξινομήσεων στις προτάσεις

	Πλήθος	Ποσοστό
NFS	13.671	66,31
UFS	2.097	10,17
CFS	4.849	23,52

3.7.4 Εξαγωγή χαρακτηριστικών από προτάσεις

«Όταν ο Πρόεδρος Μπους ανέλαβε την εξουσία, είχαμε πλεόνασμα προϋπολογισμού και το εθνικό χρέος ήταν λίγο πάνω από πέντε τρισεκατομμύρια.»

Με βάση την παραπάνω πρόταση θα γίνει επεξήγηση της εξαγωγής χαρακτηριστικών:

- Συναίσθημα: Έγινε χρήση του AlchemyAPI για να υπολογιστεί το σκορ συναισθήματος για κάθε πρόταση. Το σκορ έχει κυμαίνεται μεταξύ [-1,1], ενώ στην παραπάνω πρόταση [2] είναι -0,846376.
- Μήκος: Είναι το πλήθος λέξεων το οποίο παρσάρεται με το κιτ εργαλείων φυσικής γλώσσας ή NLTK.
- Λέξη: Οι λέξεις των προτάσεων χρησιμοποιήθηκαν για τη δημιουργία TFIDF χαρακτηριστικών (βλ. paper 2). Αφού απορρίφθηκαν όλες οι λέξεις διακοπής (stop words), υπήρχαν 6.549 μεμονωμένες λέξεις.
- Μαρκάρισμα ως μέρος του λόγου: Εφαρμόστηκε η επισήμανση POS (Part-of-Speech) του NTLK στις προτάσεις. Για κάθε πρόταση, ο αριθμός των λέξεων που ανήκει σε POS είναι η τιμή του αντίστοιχου χαρακτηριστικού. Για παράδειγμα, στην πρόταση [2] υπάρχουν 3 λέξεις (ανέλαβε, είχαμε, ήταν) με POS tag VBD (Verb Past Sentence – Ρήμα Παρελθοντικός Χρόνος) και 2 λέξεις (πέντε, τρισεκατομμύρια) με POS tag CD (Cardinal Number – Απόλυτος Αριθμός).
- Τύπος οντότητας: Χρησιμοποιήθηκε το AlchemyAPI για τη εξαγωγή οντοτήτων από προτάσεις. Υπάρχουν 2.727 οντότητες στις επισημασμένες προτάσεις που ανήκουν σε 26 τύπους. Η πρόταση[2] έχει μία οντότητα «Μπους» τύπου «Πρόσωπο». Για μια πρόταση, ο αριθμός των οντοτήτων ενός συγκεκριμένου τύπου είναι η τιμή του αντίστοιχου χαρακτηριστικού.
- Επιλογή χαρακτηριστικών: Υπάρχουν συνολικά 6.615 χαρακτηριστικά, γι'αυτό και η επιλογή χαρακτηριστικών είναι πολύ σημαντική διαδικασία. Για παράδειγμα, στην πρόταση [2] το πιο διακριτικό χαρακτηριστικό είναι το POS tag VBD, καθώς υποδηλώνει ρήματα σε παρελθοντικό χρόνο, που συνήθως χρησιμοποιείται για την περιγραφή κάποιου παρελθοντικού γεγονότος. Το δεύτερο πιο διακριτικό χαρακτηριστικό είναι το POS tag CD, αφού οι CFS ισχυρισμοί είναι πιθανότερο να περιέχουν αριθμητικές τιμές (45% του συνόλου δεδομένου του ClaimBuster), ενώ οι NFS ισχυρισμοί είναι το λιγότερο πιθανό να περιέχουν (6% του συνόλου δεδομένων).

3.7.5 Αξιολόγηση

Πραγματοποιήθηκε ταξινόμηση (classification) των NFS/UFS/CFS προτάσεων χρησιμοποιώντας τεχνική επιτηρούμενης μάθησης, όπως ο ταξινομητής Multinomial Naive Bayes (NBC), Support Vector Machine (SVC) και Random Forest Ταξινομητής (RFC). Οι μέθοδοι αυτές αξιολογήθηκαν με τετραπλή διασταύρωση. Επιπλέον, έγιναν πειράματα διάφορων συνδυασμών των εξαγόμενων χαρακτηριστικών των προτάσεων.

Στον παρακάτω πίνακα φαίνεται η σύγκριση των NBC, SVM, RFC σε συνδυασμό με διάφορα σύνολα χαρακτηριστικών, όσον αφορά την ακρίβεια p (precision), την ανάκληση r (recall) και F-measure f | wavg δηλώνει το σταθμισμένο μέσο όρο του αντίστοιχου μέτρου στις 3 κατηγορίες:

Πίνακας 11: Σύγκριση των NBC, SVM, RFC

algorithm	features	p_NFS	p_UFS	p_CFS	p_wavg	r_NFS	r_UFS	r_CFS	r_wavg	f_NFS	f_UFS	f_CFS	f_wavg
RFC	W	0.755	0.125	0.638	0.692	0.965	0.004	0.235	0.745	0.848	0.008	0.343	0.685
NBC	W	0.788	0	0.816	0.747	0.983	0	0.385	0.791	0.875	0	0.522	0.744
SVM	W	0.871	0.426	0.723	0.811	0.925	0.227	0.667	0.826	0.897	0.296	0.694	0.816
RFC	W_P	0.772	0.358	0.701	0.731	0.968	0.011	0.312	0.764	0.859	0.02	0.43	0.713
NBC	W_P	0.799	0	0.805	0.753	0.979	0	0.44	0.8	0.88	0	0.569	0.758
SVM	W_P	0.873	0.43	0.724	0.813	0.925	0.24	0.671	0.827	0.898	0.307	0.696	0.818
RFC	W_P_ET	0.77	0.238	0.665	0.715	0.964	0.008	0.298	0.758	0.856	0.016	0.411	0.706
NBC	W_P_ET	0.803	0	0.791	0.752	0.976	0	0.455	0.801	0.881	0	0.577	0.76
SVM	W_P_ET	0.873	0.427	0.723	0.813	0.925	0.24	0.67	0.827	0.898	0.307	0.695	0.817

Έγιναν πειράματα με διάφορους συνδυασμούς των εξαγόμενων χαρακτηριστικών των προτάσεων. Στον παραπάνω πίνακα βλέπουμε πως το SVM είχε την μεγαλύτερη ακρίβεια συνολικά. Όσον αφορά την ταξινόμηση των CFS, το SVM σε συνδυασμό με λέξεις, POS tags και τύπους οντοτήτων, πέτυχε μία **p** ακρίβεια 72% (δηλαδή 72% των φορών που δηλώνεται πως μια πρόταση είναι CFS, όντως είναι) και μία ανάκληση **r** 67% (δηλαδή το 67% των CFS ταξινομούνται όντως ως CFS).

Το ClaimBuster πέτυχε εξαιρετική απόδοση στην κατάταξη. Για παράδειγμα, για τις πρώτες 100 προτάσεις, η ακρίβειά του είναι 0,96. Αυτό δείχνει πως το ClaimBuster έχει μια ισχυρή συμφωνία με τους υψηλούς ποιότητας ανθρώπινους ελεγκτές ως προς την αξία ελέγχου εγκυρότητας των προτάσεων.

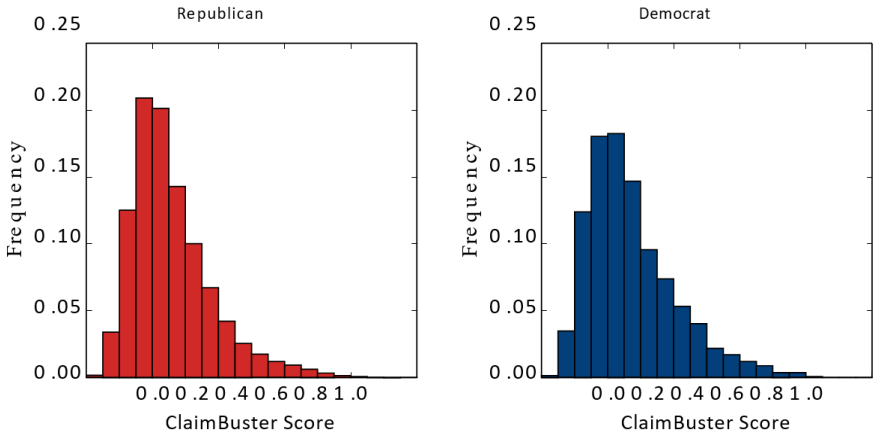
3.7.6 Ντιμπέιτ Αμερικανικών Εκλογών 2016: Μελέτη Περίπτωσης (Case Study)

Έγινε σύγκριση του ClaimBuster με τους ανθρώπινους ελεγκτές γεγονότων σε διάφορους Οργανισμούς που ασχολούνται με τη διασταύρωση γεγονότων και πληροφοριών. Θα έχει ενδιαφέρον αν στη συνέχεια προκύψει πως οι ισχυρισμοί που επιλέγονται από το ClaimBuster έχουν πολύ μεγαλύτερη πιθανότητα να είναι άξιοι ελέγχου, απ'ότι οι ισχυρισμοί που επιλέγονται από τους επαγγελματίες του χώρου. Αν αυτό προκύψει ως αληθές γεγονός, τότε το ClaimBuster θα μπορεί να χρησιμοποιηθεί, επίσης, ως βοήθημα των επαγγελματιών του χώρου ώστε να αυξήσουν την εργατική τους αποτελεσματικότητα.

Συλλέχθηκαν αντίγραφα όλων των συζητήσεων των ντιμπέιτ του 2016 από πολλούς ειδησεογραφικούς ιστοτόπους (Washington Post, CNN, Times κ.ο.κ.). Συνολικά, υπάρχουν 30.737 προτάσεις, οι οποίες προεπεξεργάστηκαν ώστε να προσδιοριστεί ο ομιλητής της κάθε πρότασης. Επίσης, προσδιορίστηκαν οι ρόλοι του κάθε ομιλητή,

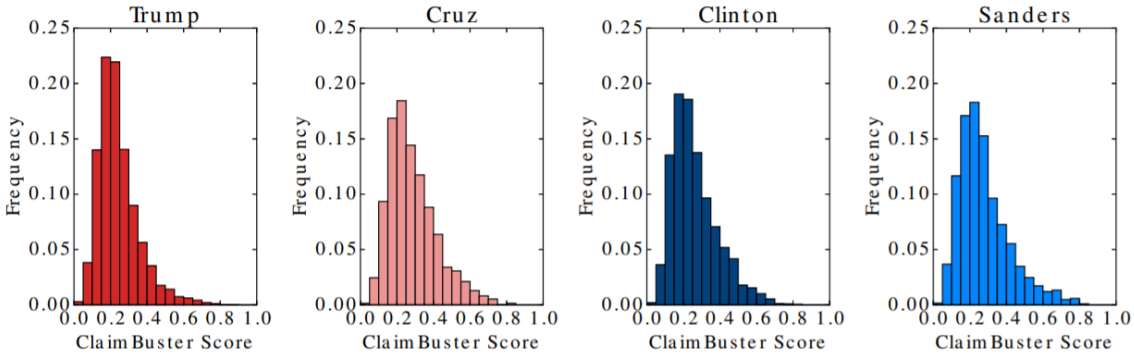
όπου όταν επρόκειτο για συντονιστή του ντιμπέιτ, οι προτάσεις αυτές αποκλείστηκαν από τη μελέτη.

Στους παρακάτω πίνακες φαίνονται οι κατανομές των σκορ του ClaimBuster σε όλες τις προτάσεις για τα δύο πολιτικά κόμματα. Οι κατανομές φαίνεται να μοιάζουν.



Σχήμα 8: Σκορ Claimbuster για Ρεπουμπλικανούς και Δημοκράτες

Προκύπτει πως υπάρχουν 776 ισχυρισμοί που είναι άξιοι ελέγχου (CFS) που έχουν ομιληθεί από Ρεπουμπλικάνους με το σκορ του ClaimBuster να είναι πάνω από 0,5. Αυτό αποτελεί το 5,06% όλων των προτάσεων που ομιλήθηκαν από Ρεπουμπλικάνους υποψηφίους. Όσο αφορά τους Δημοκρατικούς, έχουμε 484 (6,73%) ισχυρισμούς με σκορ στο ClaimBuster πάνω από 0,5.



Σχήμα 9: Σκορ Claimbuster για Trump, Cruz, Clinton, Sanders

Οι παραπάνω κατανομές δείχνουν το σκορ αξίας ελέγχου για τους κύριους και πιο σημαντικούς υποψήφιους (νικητές υποψηφιότητας και επιλαχόντες) και από τις δυο πλευρές. Μεταξύ αυτών των 4 υποψηφίων, ο Ντόναλντ Τραμπ φαίνεται να έχει παρουσιάσει λιγότερους ισχυρισμούς CFS (ClaimBuster σκορ 0,5) σε σχέση με τους υπόλοιπους 3. Επιπλέον, έχει κυρίως χρησιμοποιήσει προτάσεις NFS (ClaimBuster σκορ 0,3) σε σχέση με τους υπόλοιπους.

Από κάθε ένα από τα 21 ντιμπέιτ, επιλέχθηκαν τα 20 με υψηλότερο σκορ και χειροκίνητα τοποθετήθηκαν σε θεματικές κατηγορίες. Κατόπιν, δημιουργήθηκε ένα πρόγραμμα εντοπισμού θέματος συζήτησης το οποίο, δεδομένης μιας πρότασης, υπολογίζει ένα σκορ για κάθε θέμα από τη λίστα θεμάτων με βάση τις λέξεις-κλειδιά σε

κάθε πρόταση. Το σκορ απαρτίζεται από το συνολικό αριθμό εμφανίσεων τέτοιων λέξεων-κλειδιών. Έτσι, η πρόταση αντιστοιχίζεται με το θέμα με το υψηλότερο σκορ.

Προκειμένου να αξιολογηθεί αυτή η προσέγγιση εντοπισμού θεμάτων, δημιουργήθηκαν δεδομένα ground-truth για ένα Ρεπουμπλικανό ντιμπέιτ και ένα Δημοκρατικό. Χρησιμοποιήθηκαν προτάσεις με τουλάχιστον σκορ 0,5 στο ClaimBuster. Στα δεδομένα ground-truth για το Δημοκρατικό ντιμπέιτ υπάρχουν 52 προτάσεις με 39 από αυτές να έχουν αντιστοιχιστεί με θέμα. Το πρόγραμμα εντοπίζει θέματα στις 27 από τις 39 προτάσεις και μόνο 1 πρόταση αντιστοιχήθηκε με λάθος θέμα. Για τα ground-truth του Ρεπουμπλικανού ντιμπέιτ, υπάρχουν 62 προτάσεις, όπου οι 44 έχουν αντιστοιχισθεί με θέμα. Το πρόγραμμα βρήκε θέματα για τις 30 από τις 44 προτάσεις και 5 από αυτές ήταν με λάθος θέμα. Στη συνέχεια, εφαρμόστηκε το πρόγραμμα εντοπισμού θέματος σε όλες τις απομένουσες προτάσεις των ντιμπέιτ.

3.7.7 Αποτελέσματα

Το CNN και το PolitiFact αποτελούν μέσα στα τα οποία μπορούν να συγκριθούν τα αποτελέσματα του ClaimBuster. Για κάθε ένα εκ των 21 ντιμπέιτ, το CNN και το PolitiFact προετοίμασαν μία σύνοψη των ισχυρισμών που επέλεξαν να ελέγξουν και ετοίμασαν την ετυμηγορία τους. Όλες αυτές οι ετυμηγορίες συγκεντρώθηκαν ώστε να συγκριθούν με τα αποτελέσματα του ClaimBuster. Ο παρακάτω πίνακας δείχνει τα σκορ που δόθηκαν από το ClaimBuster στους ισχυρισμούς που ελέγχθηκαν για την εγκυρότητά τους από το CNN και το PolitiFact.

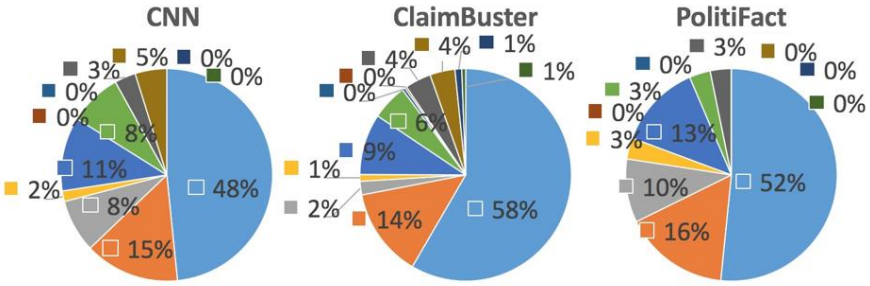
Πίνακας 12: Μέσος όρος σκορ ClaimBuster

Platforms	avg(YES)	avg(NO)	t-value	p-value
CNN	0.433	0.258	21.137	1.815E-098
PolitiFact	0.438	0.258	16.362	6.303E-060

Ο μέσος όρος σκορ του ClaimBuster για τις προτάσεις που ελέγθηκαν από το CNN είναι 0,433 σε σύγκριση με το 0,258 για αυτές που δεν ελέγχθηκαν, μία στατιστικά σημαντική διαφορά. Ομοίως, ο μέσος όρος σκορ για τις προτάσεις που ελέγχθηκαν από το PolitiFact είναι 0,438 σε σχέση με το 0,258 για αυτές που δεν ελέχθηκαν, επίσης σημαντική διαφορά.

Τα αποτελέσματα αυτής της σύγκρισης δείχνουν τη χρησιμότητα του ClaimBuster στον εντοπισμό προτάσεων που να περιέχουν αξιόλογους ισχυρισμούς.

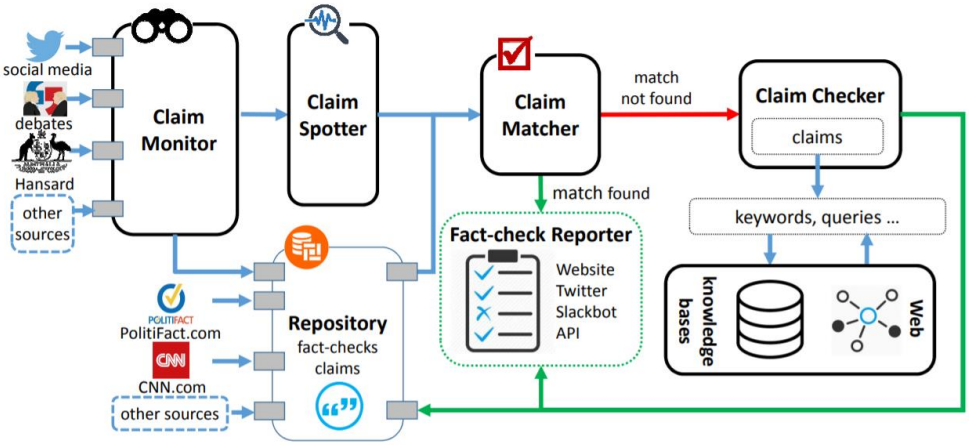
Στο παρακάτω διάγραμμα μπορούμε να παρατηρήσουμε τις κατανομές θεμάτων των προτάσεων των CNN και PolitiFact, καθώς και των προτάσεων CFS με ClaimBuster σκορ > 0,5. Από το διάγραμμα αυτό δύναται να βγει το συμπέρασμα πως υπάρχουν ισχυρές ομοιότητες μεταξύ του ClaimBuster και των οργανισμών ελέγχου ισχυρισμών και γεγονότων. Επίσης, το ClaimBuster τείνει να δίνει υψηλά σκορ στα θέματα που το CNN και το PolitiFact τείνουν να επιλέγουν για αξιολόγηση ως προς την εγκυρότητά τους.



Σχήμα 10: Διαγράμματα CNN, Claimbuster, PolitiFact

3.7.8 Τρέχουσα Κατάσταση ClaimBuster

Το ClaimBuster αναπτύσσεται και εξελίσσεται σε καθημερινή βάση. Το εργαλείο φιλοξενείται στο σύνδεσμο https://idir.uta.edu/claimbuster/ και τα χαρακτηριστικά του διευρύνονται συνεχώς. Παρακάτω μπορούμε να δούμε ένα σχεδιάγραμμα της αρχιτεκτονικής συστήματος του εργαλείου:



Σχήμα 11: Σχεδιάγραμμα της αρχιτεκτονικής συστήματος ClaimBuster

- Claim Monitor: Παρακολουθεί συνεχώς και λαμβάνει κείμενο από διάφορες πηγές.
- Claim Matcher: Δεδομένου ενός σημαντικού πραγματικού ισχυρισμού που αναγνωρίστηκε από το Claim Spotter, το Claim Matcher ψάχνει σε ένα σύνολο δεδομένων (repository) και αξιολογεί αν η πρόταση είναι άξια ελέγχου.
- Claim Checker: Συλλέγει αποδεικτικά στοιχεία από knowledge bases (π.χ. Wolfram Alpha, Google Answer Boxes). Στο μεταξύ, ο εκάστοτε ισχυρισμός στέλνεται στο Google ως search query και κατόπιν ο Claim Checker παρσάρει τα αποτελέσματα αναζήτησης και κατεβάζει την πρώτη σελίδα των αποτελεσμάτων. Οι προτάσεις που ταιριάζουν και μερικές προτάσεις πριν/μετά από αυτές ομαδοποιούνται μαζί σε ένα πλαίσιο, στο οποίο θα προστεθούν και τα αποδεικτικά στοιχεία που πάρθηκαν από τις knowledge bases.
- Fact-check Reporter: Συνθέτει ένα report συνδυάζοντας τα αποδεικτικά στοιχεία που υπάρχουν και το επιστρέφει στο χρήστη μέσω της ιστοσελίδας που αναφέρθηκε.

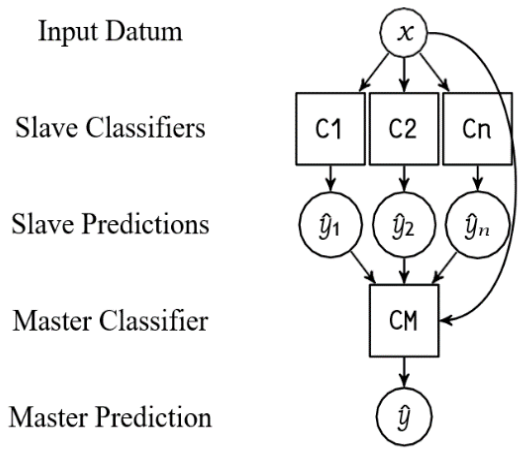
3.8 Ανίχνευση ψευδών ειδήσεων χρησιμοποιώντας συσσωρευμένο σύνολο ταξινομητών (Thorne et al. 2015 [62])

Το paper αυτό ασχολείται με το «Fake News Challenge 2017» και παρουσιάζει ένα σύστημα ταξινομητών που αναπτύχθηκε από κάποιους φοιτητές στο πλαίσιο της ενότητας επεξεργασίας φυσικής γλώσσας (NLP) στο Πανεπιστήμιο του Σέφιλντ.

Ουσιαστικά το «Fake News Challenge» είναι μία εργασία (task) ταξινόμησης κειμένου, όπου δεδομένου ενός τίτλου και σώματος ενός άρθρου, ο ταξινομητής πρέπει πρώτα να προβλέψει αν τα δύο αυτά σχετίζονται και αν ναι, τότε να εκχωρήσει μία ετικέτα τοποθέτησης (stance label) όσον αφορά το αν το τίτλος συμφωνεί, διαφωνεί ή συζητείται στο άρθρο.

Η αξιολόγηση του FNC είναι η εξής: Για κάθε τοποθέτηση, 0.25 πόντοι είναι διαθέσιμοι αν ταξινομηθεί σωστά το άρθρο με τον τίτλο του, ενώ περεταίρω 0.75 πόντοι είναι διαθέσιμοι αν χαρακτηριστεί σωστά η σχέση μεταξύ ενός ζεύγους τίτλου-άρθρου.

Το σύνολο δεδομένων που χρησιμοποιήθηκε είχε 49.972 ετικέτες τοποθέτησης (ζεύγη τίτλων και σωμάτων άρθρων) που πάρθηκαν από 2.582 άρθρα, καθώς και ένα σετ δοκιμών με 25.413 ετικέτες τοποθέτησης από 904 άρθρα που χρησιμοποιήθηκαν για την αξιολόγησης της τελικής λύσης.



Σχήμα 12: Αρχιτεκτονική προτεινόμενου συστήματος

Η λύση που προτείνεται από τους φοιτητές, αποτελεί ένα συσσωρευμένο σύνολο πέντε ανεξάρτητων ταξινομητών, που είναι αρχιτεκτονικής δύο επιπέδων και αξιοποιεί τις προβλέψεις των εξαρτώμενων (slave) ταξινομητών ως χαρακτηριστικά ενός ισχυρότερου κύριου (master) ταξινομητή. Στο διπλανό πίνακα δύναται να παρατηρηθεί η εν λόγω αρχιτεκτονική, όπου οι κυκλικοί κόμβοι συμβολίζουν τα δεδομένα, τα τετράγωνα συμβολίζουν τους ταξινομητές και τα βελάκια της ροή των δεδομένων. Στη συνέχεια θα γίνει μία επισκόπηση των πέντε ταξινομητών (C1-C5), καθώς και του κύριου ταξινομητή (CM):

- C1: Συνδυάζει τον μέσο όρο των word2vec διανυσμάτων για τον τίτλο και σώμα του άρθρου, την ομοιότητα συνημίτονου μεταξύ των TF-IDF διανυσμάτων τίτλου και σώματος και μετράει τον αριθμό των αντικρουόμενων λέξεων. Αποτελεί ταξινόμηση τετραπλής διασταύρωσης χρησιμοποιώντας (300,8) multi-layer perceptron (MLP) μαζί με ReLU συχνότητα ενεργοποίησης.
- C2: Παίρνει τον μέσο όρο από τις word2vec ενσωματώσεις για τις λέξεις των τίτλων και τις λέξεις των σωμάτων εξαιρώντας τις stop words, για σημεία στίξης, την επικάλυψη λέξεων, καθώς και μετράει τις αντικρουόμενες λέξεις. Αποτελεί ταξινόμηση τετραπλής διασταύρωσης χρησιμοποιώντας (1010,6) MLP μαζί με ReLU συχνότητα ενεργοποίησης.
- C3: Ταξινόμηση τετραπλής διασταύρωσης χρησιμοποιώντας ένα-ενταντίων-όλων λογική παλινδρόμηση με κανονικοποίηση L2 σε TF-IDF διανύσματα unigram και bigram.

- C4: Συνδυάζει τις word2vec ενσωματώσεις για λέξεις τίτλου και άρθρου και γίνεται ταξινόμηση τετραπλής διασταύρωσης χρησιμοποιώντας (256,128,128) MLP με πιθανότητες εγκατάλειψης (0,5,0,3,0,1) μεταξύ των επιπέδων και της συχνότητας ενεργοποίησης ReLU.
- C5: Επίσημος baseline ταξινομητής του FNC.
- CM: Ταξινομητής ενισχυμένου δέντρου αποφάσεων που χρησιμοποιεί ως χαρακτηριστικά του τις τιμές που προκύπτουν από τα C1-C5.

Ο κύριος ταξινομητής CM εκπαιδεύεται χρησιμοποιώντας διπλή διασταυρούμενη επικύρωση (2 fold cross validation) ως εξής: Το σύνολο δεδομένων χωρίζεται τυχαία σε δύο μεγέθη. Δύο περιπτώσεις (instances) των C1-C5 εκπαιδεύονται ανεξάρτητα σε κάθε πτυχή δεδομένων. Οι προβλέψεις συνδυάζονται με τα αρχικά δεδομένα εισόδου για να σχηματίσουν ένα σύνολο δεδομένων, όπου και θα είναι τα κύρια δεδομένα που θα εκπαιδεύσουν τον CM. Νέες περιπτώσεις των C1-C5 εκπαιδεύονται σε ολόκληρο το αρχικό σύνολο δεδομένων εκπαίδευσης και χρησιμοποιούνται για να παρέχουν είσοδο στο CM κατά τη διάρκεια της τελικής δοκιμής.

Πίνακας 13: Αποτελέσματα εκπαιδευμένων ταξινομητών

System	Dev %	Test %
Official Baseline[3]	79.53	75.20
SOLAT in the SWEN[4]	-	82.02
Athene[5]	-	81.97
UCL Machine Reading[6]	-	81.72
C1	88.09	75.77
C2	86.68	75.08
C3	87.48	77.99
C4	87.36	58.69
C5	79.25	75.22
Our Ensemble (CM)	90.05	78.04
CM Upper Limit	*97.25*	*90.89*

[3](Galbraith et al., 2017) [4](Bird et al., 2017)
[5](Hanselowski et al., 2017) [6](Riedel et al., 2017)

Στο διπλανό πίνακα παρουσιάζονται τα αποτελέσματα των ταξινομητών που εκπαιδεύτηκαν και αξιολογήθηκαν πάνω στα baseline δεδομένα του «Fake News Challenge» (Dev) και του τελικού σετ δοκιμής (Test). Διαπιστώθηκε πως η πολλαπλή επικύρωση πάνω στο Dev απέδωσε ταξινομητές που δεν μπόρεσαν να λειτουργήσουν κατάλληλα πάνω στα μη ορατά ακόμα άρθρα του σετ δοκιμής, βλάπτοντας έτσι την ακρίβεια ταξινόμησης. Στο Dev σύνολο, το σύστημα ταξινομητών των φοιτητών απέδωσε μία απόλυτη βελτίωση κατά τουλάχιστον 1,6% σε σχέση με οποιονδήποτε από τους μεμονωμένους εξαρτώμενους ταξινομητές.

Ο CM Upper Limit χρησιμοποιεί ένα είδος βαθμολογίας που απονέμει σκορ FNC αν τουλάχιστον ένας από τους εξαρτώμενους ταξινομητές επισημαίνει με σωστή ετικέτα την τοποθέτηση της εισόδου (input stance). Αυτό λειτουργεί ως ένα μέτρο που περιγράφει το μέγιστο πιθανό σκορ που θα μπορούσε να δώσει ο CM, υποθέτοντας πως πάντα επέλεγε μία σωστή ετικέτα από έναν από τους εξαρτώμενους ταξινομητές. Στην περίπτωση αυτή, το ανώτατο όριο ήταν 90,89%, που υπερβαίνει το ανώτατο μέχρι τώρα σύστημα. Αν και το αποτέλεσμα αυτό είναι ενθαρρυντικό, υπογραμμίζεται η ανάγκη να δημιουργηθεί ένας ισχυρότερος κύριος ταξινομητής που να είναι πιο ανθεκτικός στις θορυβώδεις προβλέψεις των εξαρτώμενων ταξινομητών.

Η απόδοση των εξαρτώμενων ταξινομητών των φοιτητών (C1-C4) είναι μεταβλητή και εξαρτάται σε μεγάλο βαθμό από την τοπολογία του δικτύου, την επιλογή των χαρακτηριστικών και το σύνολο δεδομένων. Ο πιο ανθεκτικός ταξινομητής, C5, χρησιμοποιεί εντελώς μη λεξικά χαρακτηριστικά, ενώ ο C4, που χρησιμοποιεί μόνο τον μέσο όρο διανυσμάτων λέξεων και μία μεγάλη τοπολογία δικτύου, υπέστη τη μεγαλύτερη απώλεια απόδοσης.

3.9 Μπορούν οι μηχανές να μάθουν να εντοπίζουν ψευδείς ειδήσεις; Μια έρευνα εστιασμένη στα μέσα κοινωνικής δικτύωσης (Silva et al. 2019 [63])

Το paper αυτό αποτελεί μία μελέτη διάφορων papers των τελευταίων 5 χρόνων (2015-2019) που αφορούν τεχνικές εντοπισμού ψευδών ειδήσεων και ασχολείται με μία γενική ανάλυση αυτών, εστιάζοντας στα χαρακτηριστικά των διαφορετικών προσεγγίσεων, τα εννοιολογικά μοντέλα, καθώς και τον ρόλο των bots (γνωστικών παραγόντων) σε αυτό το πλαίσιο αφού έχουν αποκτήσει μεγάλη δημοτικότητα τα τελευταία χρόνια.

Για την συστηματική ανασκόπηση της βιβλιογραφίας χρησιμοποιήθηκε η μέθοδος SLR, όπως περιγράφεται στα [1] και [2]. Για την αυτοματοποίηση αυτής της διαδικασίας έγινε χρήση του εργαλείου «Parsifal», το οποίο είναι ένα διαδικτυακό εργαλείο SLR που επέτρεψε τον καθορισμό ενός συνόλου λέξεων κλειδιών και βασικών ερευνητικών ερωτήσεων, καθώς και ενός συνόλου πηγών αναζήτησης.

Πίνακας 14: Λέξεις κλειδιά που χρησιμοποιήθηκαν κατά την αναζήτηση

Keywords	Synonyms	Related To
Detection	Stance, Tracking, Veracity	Intervention
Fake News	Automated Fact Checking, Disinformation, Hoax, Misbehavior, Misinformation, Rumor	Outcome
Machine Learning	Artificial Intelligence, ML, Natural Language, Processing, NLP	Comparison
Social Media	Facebook News, Newspaper, Twiter	Population

Το ερώτημα (query) που δημιουργήθηκε από τα επιλεγμένα keywords είναι το παρακάτω, το οποίο και έφερε 1.093 άρθρα ως αποτελέσματα αναζήτησης:

(”Detection” OR ”Stance” OR ”Tracking” OR ”Veracity”) AND (”machine learning” OR ”Artificial Intelligence” OR ”ML” OR ”Natural Language Processing” OR ”NLP”) AND (”Fake news” OR ”Automated Fact checking” OR ”disinformation” OR ”Hoax” OR ”misbehavior” OR ”misinformation” OR ”Rumor”)”

Από αυτά, όπως αναφέρθηκε, επιλέχτηκαν εκείνα που δημοσιεύτηκαν το 2015 και μετά, καθώς και προτιμήθηκαν όσα έγγραφα ήταν πειραματικά με πραγματικά δεδομένα και αποτελέσματα που χρησιμοποιούσαν οποιονδήποτε αλγόριθμο μηχανικής μάθησης, τεχνητής νοημοσύνης ή αυτοματοποιημένου αλγορίθμου λήψης αποφάσεων. Επιπλέον, προτιμήθηκαν και επιλέχτηκαν papers που σχετίζονται με πολιτικά, εκ των οποίων εκτελέστηκε περαιτέρω διαχωρισμός αυτών προκειμένου να γίνει μία επιθεώρηση των τεχνικών, των ορισμών, του θεωρητικού υποβάθρου, τον τύπο της κάθε μελέτης και των αποτελεσμάτων της.

Ακόμα, τόσο η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) όσο και οι τεχνικές μηχανικής μάθησης (machine learning) ήταν το κύριο επίκεντρο ενδιαφέροντος.

3.9.1 Θεωρητική Αναφορά

Υπάρχουν πολλοί ορισμοί για τις ψευδείς ειδήσεις στη βιβλιογραφία [26]. Επίσης, τα μέσα ενημέρωσης έχουν χρησιμοποιήσει υπερβολικά τον παραπάνω όρο σε πολλά διαφορετικά πλαίσια και με διαφορετικές προθέσεις, κάτι που επιδεινώνει το πρόβλημα της κατανόησης ως προς το -τι- χαρακτηρίζει μια δεδομένη ιστορία ως ψευδή είδηση.

3.9.2 Εκδότης

Η οντότητα που παρέχει την ιστορία σε ένα κοινό. Για παράδειγμα, ο εκδότης μπορεί να είναι χρήστης μιας υπηρεσίας micro-blogging όπως το Twitter, δημοσιογράφος σε μια διαδικτυακή εφημερίδα ή ένας οργανισμός στον δικό του ιστότοπο. Σημειώνεται πως ο εκδότης μπορεί να είναι ο συγγραφέας της ιστορίας ή και όχι.

3.9.3 Περιεχόμενο

Οι κύριες πληροφορίες που παρέχονται από τον εκδότη στην ιστορία. Τη στιγμή της δημοσίευσης, η εγκυρότητα αυτών των πληροφοριών μπορεί να είναι αληθής, ψευδής ή άγνωστη. Εάν η εγκυρότητα είναι άγνωστη, τότε η εν λόγω ιστορία μπορεί να χαρακτηριστεί ως φήμη. Οι πληροφορίες μπορούν ακόμα να ταξινομηθούν ως πραγματικοί ισχυρισμοί, γνώμες ή μίξη αυτών.

3.9.4 Clickbaiting

Εκτός του περιεχομένου, η εκάστοτε ιστορία μπορεί να περιλαμβάνει ορισμένα μέσα όπως εικόνα, βίντεο, ήχο. Η χρήση μέσων που δε σχετίζονται με το περιεχόμενο και έχουν στόχο την αύξηση της θέλησης του αναγνώστη να διαβάσει το περιεχόμενο, είναι γνωστό ως Clickbaiting.

Τα διάφορα papers χρησιμοποιούν διαφορετικούς ορισμούς, όπως «ψευδείς ειδήσεις», «φήμη», «απάτη», όλα για να περιγράψουν την ίδια σημασιολογική έννοια, δηλαδή πληροφορίες που δεν έχουν επαληθευτεί, με την πρόθεση είτε να μπλοκάρουν τη διάδοση γνώσης (με τη διάδοση άσχετων ή λανθασμένων πληροφοριών), είτε για να χειραγωγήσουν τη γνώμη των αναγνωστών. [30] [31] [32] [33] [34] [35] [36]

Συνεχίζοντας, λόγω της διάδοσης της τεχνητής νοημοσύνης και των συναφών τομέων των γνωστικών υπολογιστικών συστημάτων (cognitive computing), ο αριθμός των bots είναι πλέον τεράστιος σε κάθε γωνιά του διαδικτύου. Μερικοί συγγραφείς υποστηρίζουν ότι με τη δημιουργία bots, οι γνωστικοί παράγοντες θα ήταν πιο επιβλαβείς για τη διαδικασία ανάκτησης πληροφοριών, λόγω του γεγονότος ότι θα εντείνουν τη διάδοση παραπληροφόρησης και των ανεπιθύμητων μηνυμάτων. [37] Ωστόσο, όπως ανακαλύφθηκε μέσω πειραμάτων, στην πραγματικότητα τα bots θα αυξάνουν πράγματι

τη διάδοση παραπληροφόρησης, αλλά, επίσης, δύναται να αυξάνουν και τη διάδοση αληθινών πληροφοριών [38]. Το συμπέρασμα λοιπόν είναι πως τα bots δεν αποτελούν αρωγοί παραπληροφόρησης, αλλά είναι απλώς οι διανομείς κάθε είδους πληροφορίας ανεξάρτητα για το αν επρόκειτο για ψευδείς ή αληθείς.

3.9.5 Μέσα Κοινωνικής Δικτύωσης και Μηχανική Μάθηση

Λόγω της τεράστιας αύξησης χρήσης των κοινωνικών δικτύων τα τελευταία χρόνια, διαπιστώθηκε πως τα περισσότερα papers χρησιμοποιούν αυτά ως κύρια πηγή ανάλυσης πληροφοριών και ειδήσεων. Τα μέσα κοινωνικής δικτύωσης συνήθως παρέχουν ένα API (Application Programming Interface) για την αναζήτηση και παροχή των δεδομένων. Συνήθως, το εκάστοτε API κάποιας κοινωνικής πλατφόρμας παρέχει το περιεχόμενό του δομημένο σε απλό κείμενο, γεγονός που μειώνει σημαντικά τον χρόνο της προεπεξεργασίας. [39] [40] [41]

Ένας ακόμα λόγος που γίνεται χρήση των μέσων κοινωνικής δικτύωσης αντί, για παράδειγμα, πιο παραδοσιακών μέσων όπως οι εφημερίδες, είναι γιατί οι περισσότερες εφημερίδες εκφράζουν συνήθως μία γενικότερη πολιτική γνώμη σε σύγκριση με τα social media όπου εκφράζουν μεμονωμένες απόψεις πολλών διαφορετικών χρηστών με διαφορετικές πεποιθήσεις.

Όσον αφορά τον αυτόματο έλεγχο γεγονότων, υπάρχουν πολλοί ιστότοποι που διεκπεραιώνουν αυτή τη λειτουργία. Δύο από τα πιο δημοφιλή είναι το snopes.com και το factcheck.org. Επίσης, υπάρχουν εξειδικευμένοι ιστότοποι για εξειδικευμένους τομείς όπως η πολιτική, σαν το politifact.com. Αντίθετα, υπάρχουν, επιπλέον, πολλοί ιστότοποι, όπως το theonion.com, που δημοσιεύουν ειδήσεις που δηλώνονται ρητά ψευδείς. Πολλοί από αυτούς τους ιστότοπους δημοσιεύουν αυτές τις ψευδείς ειδήσεις ως σατιρικό ή χιουμοριστικό περιεχόμενο.

Όσον αφορά το βήμα της προεπεξεργασίας πληροφοριών, τα περισσότερα papers τη χρησιμοποιούν για να έχουν ταχύτερη συνολική επεξεργασία. [42] [43] [44]
Υπάρχουν έρευνες που εστιάζουν στην αυτόματη ανίχνευση της αφετηρίας ή της αρχικής πηγής της ροής των ψευδών ειδήσεων, μέσω τοπολογικής εξερεύνησης. Οι συγγραφείς του [45] πρότειναν έναν αλγόριθμο για να το κάνουν και έλαβαν πολύ καλά αποτελέσματα (σε σύγκριση με τους άλλους που δοκιμάστηκαν) βρίσκοντας την προέλευση των ψευδών πληροφοριών.

Επιπλέον, σε πολλά papers χρησιμοποιήθηκε η ανάλυση συναισθημάτων για να ταξινομηθεί η πολικότητα μιας είδησης [46] [47] [48] [49] [43]. Μερικά χρησιμοποίησαν διάφορα λεξικά συναισθημάτων, τα οποία απαιτούν αρκετή ανθρώπινη προσπάθεια για να δημιουργηθούν και να διατηρηθούν σε καλό επίπεδο, μέσω τον οποίων δημιουργήθηκε ένας ταξινομητής βασισμένος στην επιτηρούμενη μάθηση. Ορισμένα papers που χρησιμοποιούν μια τέτοια προσέγγιση ανάλυσης συναισθημάτων ως χαρακτηριστικό για τους τελικούς ταξινομητές, χρησιμοποιούν μοντέλα αλυσίδων όπως Hidden Markov Models ή Artificial Neural Network για να συναγάγουν συναισθήματα. Τέλος, τα πιο συνήθη χρησιμοποιούμενα λεξικά είναι τα WordNet και το Linguist Inquiry and Word Count (LIWC).

Εν συνεχεία, ομαδοποιήθηκαν τα χαρακτηριστικά των ταξινομητών των συνόλων δεδομένων με βάση την πηγή τους. Το πρώτο σύνολο περιέχει τα χαρακτηριστικά που βασίζονται σε χαρακτηριστικά κοινωνικών μέσων (#likes, #retweets, #friends). Το

δεύτερο σύνολο περιέχει τα χαρακτηριστικά που βασίζονται στο περιεχόμενο των ειδήσεων (σημεία στίξης, ενσωματώσεις λέξεων, πολικότητα συναισθημάτων λέξεων). Όπως μπορούμε να δούμε στο [50], υπάρχει σαφής προτίμηση των πιο κλασσικών αλγορίθμων ταξινόμησης, ωστόσο παρατηρείται αυξημένη χρήση νέων μεθόδων που δύναται να δώσουν καλύτερα αποτελέσματα, όπως Μοντέλα Ανάλυσης Τοπολογίας Δικτύου (Network Topology Analysis Models) και Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks). Ειδικότερα, παρατηρήθηκε αυξημένη χρήση των Τεχνητών Νευρωνικών Δικτύων, που έχουν αποκτήσει μεγάλη δημοτικότητα τελευταία, σε σχέση με τους κλασσικά μοντέλα, όπως Naive Bayes, Support Vector Machine κλπ.

3.9.6 Μελλοντικά Σχέδια και Συμπεράσματα

Η κύρια ανοιχτή πρόκληση που βρέθηκε ήταν πως εξακολουθεί να υπάρχει αβεβαιότητα σχετικά με τις πραγματικές προθέσεις ενός tweet. Λόγω γλωσσικών πόρων, όπως μεταφορές, σχήματα λόγου και σαρκασμός, ενώ ένα post σε κάποιο από τα social media μπορεί να είναι απόλυτα κατανοητό από ανθρώπινο αναγνώστη, μία μηχανή δε δύναται ακόμα να εκπαιδευτεί κατάλληλα ώστε να κατανοεί απόλυτα το νόημα μιας καθομιλούμενης φράσης, αγνοώντας τους γλωσσικούς αυτούς πόρους.

Αν και επικρατεί η γνώμη των περισσότερων ερευνητών που υποστηρίζουν ότι βασικό χαρακτηριστικό της πρόβλεψης των αποτελεσμάτων των εκλογών είναι οι μετρικές που λαμβάνονται από τα social media, μερικοί υποστηρίζουν αντίθετα πως η προσέγγιση αυτή είναι πολύ απλοϊκή λόγω της αβεβαιότητας σχετικά με τον πραγματικό στόχο μιας πολιτικής συζήτησης, αφού πολλοί διάλογοι είναι σατιρικοί και όχι πραγματικά σοβαροί. Επίσης, η έλλειψη αλγοριθμικών και λογικών φορμαλισμών και προκαταρκτικών ορισμών συνάδουν στο ότι η καλή απόδοση / βαθμολογία των υποψηφίων των εκλογών στα social media δεν είναι αρκετή για να δημιουργήσει μία σχέση αιτιότητας και να δείξει μία σίγουρη νίκη. [50]

Τέλος, η τρέχουσα κατάσταση της αυτόματης ανίχνευσης ψευδών ειδήσεων με τη χρήση σύνθετων προσεγγίσεων ανάλυσης δικτύου και μηχανικής μάθησης, οδηγεί στο συμπέρασμα πως θα μπορούσε να οριστεί μία πιο γενική έννοια των ψευδών ειδήσεων, ώστε τελικά να καταστεί δυνατή η παραγωγή καλύτερων αποτελεσμάτων όσον αφορά την ανίχνευση ψευδούς πληροφόρησης.

3.10 Ο εντοπισμός ψευδών ειδήσεων μέσω επεξεργασίας φυσικής γλώσσας είναι ευάλωτος σε επιθέσεις (Zhou et al. 2019 [64])

Οι ψευδείς ειδήσεις πληθαίνουν ολοένα μέρα με τη μέρα και κατακλύζουν το διαδίκτυο, ιδίως όσο αφορά τον πολιτικό τομέα. Για την αντιμετώπιση αυτού του προβλήματος, οι ερευνητές κι οι ειδικοί έχουν αναπτύξει τεχνικές ανίχνευσης ψευδών ειδήσεων που υιοθετούν την επεξεργασία φυσικής γλώσσας (NLP) για την επεξεργασία και ανάλυση των δοθέντων πληροφοριών. Αν και οι ανιχνευτές αυτοί σημειώνουν αρκετά καλή ακρίβεια σε ήδη υπάρχοντα παραδείγματα χειραγωγημένων ειδήσεων, η ανάλυση που γίνεται είναι συνήθως αρκετά επιφανειακή. Παρατηρείται πως τα μοντέλα που χρησιμοποιούνται [51] μπορούν να εντοπίσουν ψευδείς ειδήσεις μόνο όταν έχουν καλυφθεί, για παράδειγμα όταν το περιεχόμενο δε σχετίζεται καθόλου με τον τίτλο (το λεγόμενο «clickbait») ή όταν το άρθρο περιέχει λέξεις που θεωρούνται προκατειλημμένες και προκλητικές. Μερικοί πιο εξελιγμένοι φορείς διάδοσης φημών και ψευδών πληροφοριών μπορούν να δημιουργήσουν πιο διακριτικές επιθέσεις, όπως για παράδειγμα να λάβουν ένα καλογραμμένο πραγματικό ειδησεογραφικό άρθρο και να το παραλλάξουν ελάχιστα με στοχευμένο τρόπο. Έτσι, διατηρώντας παρόμοιο θέμα με το πρωτότυπο άρθρο και συνδέοντας το περιεχόμενο με τον τίτλο του, ένα *ανταγωνιστικό* άρθρο μπορεί εύκολα να αποφύγει τον εντοπισμό.

Για την επίδειξη αυτού του είδους της επίθεσης, δύναται να αξιολογηθεί ένα σύγχρονο μοντέλο ονόματι «Fakebox». Παρατηρούνται 3 μορφών επιθέσεις:

3.10.1 Μορφές Επιθέσεων

- Παραμόρφωση γεγονότων: Τροποποίηση ορισμένων λέξεων, όπως χαρακτήρας, χρόνος, τοποθεσία, σχέση ή οποιοδήποτε άλλο χαρακτηριστικό μπορεί να παραμορφωθεί.
- Ανταλλαγή υποκειμένου-αντικειμένου: Με αυτήν την επίθεση οι αναγνώστες θα μπερδευτούν ως προς το ποιος είναι ο ερμηνευτής και ποιος ο αποδέκτης μιας δράσης. Μπορεί να πραγματοποιηθεί σε επίπεδο πρότασης.
- Στοχευμένη χειραγώγηση: Μπορεί να επιτευχθεί είτε χτίζοντας μία ανύπαρκτη αιτιώδη σχέση μεταξύ δύο ανεξάρτητων γεγονότων, είτε κόβοντας μερικά μέρη μιας ιστορίας.

Στον παρακάτω πίνακα φαίνονται τέτοιου είδους επιθέσεις. Επαναλαμβάνοντας τις τροποποιήσεις δύναται να αλλάξει σημαντικά το σημασιολογικό περιεχόμενο ενός ειδησεογραφικού άρθρου χωρίς να διαστρεβλώνεται το στυλ γραφής του πρωτότυπου, και έτσι το τροποποιημένο άρθρο εξακολουθεί να παρουσιάζεται με φαινομενικά λογικό τρόπο:

Πίνακας 15: Παραδείγματα ειδών επιθέσεων

Attack type	Original	Adversarial
Παραμόρφωση γεγονότων	12 people were injured in the shooting.	24 people were killed in the shooting.
Ανταλλαγή υποκειμένου-αντικειμένου	A gangster was shot by the police.	A policeman was shot by the gangster.
Στοχευμένη χειραγώγηση	The condom policy originated in 1992 . . . The Boy Scouts have decided to accept people who identify as gay and lesbian. (unrelated events)	The inclusion of gays, lesbians and girls in the Boy Scouts led to the condom policy.

3.10.2 Ανάλυση αξιοπιστίας και ακρίβειας

Στη συνέχεια θα πραγματοποιηθεί ανάλυση ως προς την αξιοπιστία και την ακρίβεια του Fakebox σε τέτοιου είδους επιθέσεις· ως *«ανταγωνιστικές επιθέσεις»* εννοούμε την εμφάνιση «*ανταγωνιστικών άρθρων»*, δηλαδή ελαφρώς διαφοροποιημένων από τα πρωτότυπά τους.

Το Fakebox ουσιαστικά αναλύει τα γλωσσικά χαρακτηριστικά των ειδησεογραφικών άρθρων ώστε να εκτιμηθεί αν είναι πιθανό να είναι πραγματικά νέα ή όχι. Κοιτάζοντας διαφορετικά χαρακτηριστικά ενός άρθρου (τίτλος, περιεχόμενο και url), χρησιμοποιώντας NLP μοντέλα και εκπαιδεύοντας πάνω σε μία επιμελημένη βάση δεδομένων, το Fakebox μπορεί να αναγνωρίσει επιτυχώς ψευδείς ειδήσεις.

Αν ένα άρθρο έχει γραφεί σαν πραγματικό, το Fakebox το χαρακτηρίζει ως αμερόληπτο και δίνει σκορ μεταξύ 60 και 100. Αν ένα άρθρο δε γράφεται σαν πραγματικό, τότε το χαρακτηρίζει ως μεροληπτικό και δίνει σκορ μεταξύ 0 και 40. Διαφορετικά το χαρακτηρίζει ως αβέβαιο με σκορ μεταξύ 40 και 60. Επισημαίνει και εκχωρεί ποσοτικά αποτελέσματα για τίτλους, περιεχόμενα και domains αντίστοιχα. Τα άρθρα με υψηλότερο σκορ είναι πιθανό να είναι περισσότερο αξιόπιστα. Το Fakebox επικεντρώνεται στα γλωσσικά χαρακτηριστικά ενός άρθρου (linguistic characteristics) και δεν κάνει έλεγχο γεγονότων, το οποίο πιθανώς το καθιστά ευάλωτο όταν αντιμετωπίζει ψευδή άρθρα που έχουν γραφτεί σε παρόμοιο στυλ με πραγματικές ειδήσεις. Για την πρακτική δοκιμή αυτής της υπόθεσης, πραγματοποιήθηκε πειραματική αξιολόγηση του Fakebox με μη τροποποιημένα παραδείγματα από το σύνολο δεδομένων του McIntire (2018) και κατόπιν εφαρμόστηκαν οι αναφερθέντες ανταγωνιστικές επιθέσεις.

3.10.2.1 Αρχική έρευνα

Αρχικά, εξερευνήθηκε η βασική απόδοση του Fakebox με το McIntire dataset. Χρησιμοποιείται:

True Positive (TP) – Σωστά ταξινομημένες ψευδείς ειδήσεις
True Negative (TN) – Σωστά ταξινομημένες αληθείς ειδήσεις
False Positive (FP) – Λάθος ταξινομημένες ψευδείς ειδήσεις
False Negative (FN) – Λάθος ταξινομημένες αληθείς ειδήσεις

Τα false positive rate (FPR) και false negative rate (FPN) ορίζονται ως εξής:

$$FPR = FP/(FP + TN)$$
$$FNR = FN/(FN + TP)$$

Δοθέντων 6.355 τίτλων και άρθρων, μπορούμε να δούμε το output στον παρακάτω πίνακα:

Πίνακας 16: Fakebox output

Labels	Impartial	Biased	Unsure
Real news	1159	1477	535
Fake news	537	2184	443

Η ακρίβεια του Fakebox στο McIntire dataset είναι 52,77%, ενώ false rate σε 31,79% και για το υπόλοιπο 15,44% δεν είναι σίγουρο για την εγκυρότητά τους. Έχει καλή απόδοση όταν ασχολείται με ψευδείς ειδήσεις όπου FNR = 19,74%. Από την άλλη, χαρακτηρίζει τις αληθείς ειδήσεις περισσότερο ως μεροληπτικές παρά ως αμερόληπτες. Ποσοτικά το FPR είναι 56,03% και η συνολική του ακρίβεια εξαιρώντας αβέβαιων περιπτώσεων είναι 62,40%. Το αποτέλεσμα φαίνεται στον παρακάτω πίνακα:

Πίνακας 17: Ακρίβεια του Fakebox στο McIntire dataset

News type	Number of articles	Correctly classified	Classification accuracy
Real	2636	1159	43.97%
Fake	2721	2184	80.26%
Total	5357	3343	62.40%

3.10.2.2 Απόδοση σε ανταγωνιστικές επιθέσεις

Στη συνέχεια, θα δούμε την απόδοση του σε ανταγωνιστικές επιθέσεις. Δημιουργούνται χειροκίνητα παραδείγματα από αληθείς ειδήσεις που επισημάνθηκαν ως αμερόληπτες από το Fakebox.

Για την παραμόρφωση γεγονότων, αντικαταστάθηκαν απλώς οι άνθρωποι, τα μέρη και οι δράσεις. Για παράδειγμα, στο άρθρο με τίτλο «Is the GOP losing Walmart?», αντικαταστάθηκε κάθε λέξη «Walmart” στο περιεχόμενο με τη λέξη «Apple». Το σκορ εγκυρότητας που δίνει το Fakebox πέφτει μόλις 0,0073 το οποίο θεωρείται αμελητέο. Η παρεμβολή σε άλλα άρθρα επίσης δεν προκαλεί πτώση του σκορ εγκυρότητας. Αυτό το είδος παραμόρφωσης μπορεί να έχει τεράστιο αντίκτυπο – έστω ότι μια εταιρία Α εμπλέκεται σε σκάνδαλο παραβίασης πληροφοριών, αλλά η εταιρία Β θεωρείται υπεύθυνη από μαζικές ψευδείς ειδήσεις.

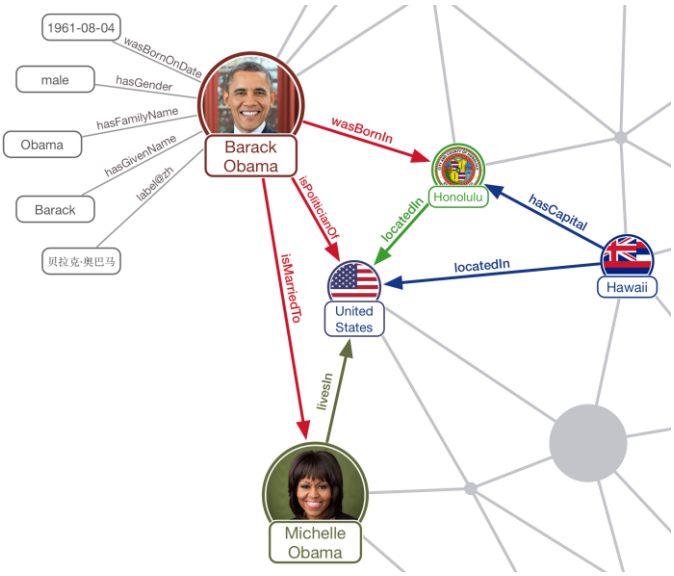
Για την ανταλλαγή υποκειμένου-αντικειμένου, το σκορ εγκυρότητας παραμένει αμετάβλητο, καθώς η συχνότητα όρου (term frequency) παραμένει η ίδια. Αυτό μπορεί να είναι σε μεγάλο βαθμό παραπλανητικό. «Ένας γκάνγκστερ πυροβολήθηκε από αστυνομικό» και «ένας αστυνομικός πυροβολήθηκε από γκάνγκστερ» είναι εντελώς διαφορετικές σημασιολογικές προτάσεις, όπου το τελευταίο δύναται να προκαλέσει πανικό στο κοινό της είδησης.

Η στοχευμένη χειραγώγηση είναι ίσως το πιο ευάλωτο τμήμα ανιχνευτών που βασίζονται σε NLP. Για παράδειγμα, υπάρχουν δύο αληθή και άσχετα μεταξύ τους άρθρα που χαρακτηρίζονται ως αμερόληπτα από το Fakebox, ένα για το Walmart με σκορ εγκυρότητας 0,7151 και το άλλο για τοπικά πολιτικά ζητήματα στο Κλίβελαντ με σκορ 0,7652. Όταν, λοιπόν, αυτά τα άρθρα μπλέξουν μεταξύ τους και ενωθούν σε ένα ενιαίο, το άρθρο που θα παραχθεί θα εξακολουθεί να χαρακτηρίζεται ως αμερόληπτο και θα έχει ακόμα και υψηλότερο σκορ 0,8585. Όσο κάποιος διατηρεί τη γραφή των άρθρων με κλασσικό τρόπο, μπορεί να αναμιγνύει εντελώς άσχετά μεταξύ τους γεγονότα, να χτίζει ανύπαρκτες αιτιώδεις σχέσεις και να αποφεύγει επιτυχώς τον εντοπισμό.

3.10.3 Crowdsourcing

Συνεχίζοντας, γνωρίζουμε πως τα media και οι ειδικοί έχουν αξιόλογες δυνατότητες αλλά περιορισμένο χρόνο και ενέργεια για να συλλέξουν γεγονότα και πληροφορίες από πολλές διαφορετικές πηγές. Οι ψευδείς ειδήσεις συνήθως κυκλοφορούν πολύ σύντομα αφού κάποια γεγονότα έχουν συμβεί, συνεπώς απαιτείται έγκαιρος εντοπισμός, επιδεινώνοντας την όλη κατάσταση.

Ένα γράφημα γνώσης (knowledge graph) είναι ένα γράφημα με οντότητες διαφορετικών τύπων ως κόμβους και διάφορες σχέσεις μεταξύ τους ως ακμές [52]. Ένα γράφημα γνώσης χρησιμοποιείται από το Google για να βελτιώσει τα αποτελέσματα της μηχανής αναζήτησης με πληροφορίες που συλλέγονται από διάφορες πηγές. Οι πληροφορίες παρουσιάζονται στους χρήστες σε ένα infobox δίπλα από τα αποτελέσματα αναζήτησης. Ένα παράδειγμα γραφήματος γνώσεων Google φαίνεται στο παρακάτω σχήμα:



Σχήμα 13: Παράδειγμα γραφήματος γνώσης

Το Crowdsourcing είναι ένα κατανεμημένο μοντέλο επίλυσης προβλημάτων, στο οποίο ένα πλήθος απροσδιόριστου μεγέθους καλείται να λύσει ένα περίπλοκο πρόβλημα μέσω ανοιχτών κλήσεων (open calls) [53]. Διαιρεί την εργασία (task) μεταξύ των συμμετεχόντων για να επιτύχει ένα αθροιστικό αποτέλεσμα. Είναι πιθανό ένα μεγάλο πλήθος μη ειδικών να μπορεί να συνεργαστεί καλά σε μία εργασία που διαφορετικά θα απαιτούσε εκτεταμένες προσπάθειες μιας μικρής ομάδας επιστημόνων [54].

Ένα γράφημα γνώσης από crowdsourcing μπορεί να είναι πολύ αποτελεσματικό και έγκαιρο στο πλαίσιο της διάδοσης ειδήσεων. Ενώ οι ψευδείς ειδήσεις πλημυρίζουν συνήθως στο αρχικό στάδιο μετά από ένα συμβάν, οι ντόπιοι ή οι καλά-ενημερωμένοι άνθρωποι γνωρίζουν καλύτερα από το καθένα την ακρίβεια των γεγονότων. Εάν μπορούσε να δημιουργηθεί μία δομημένη οπτικοποιημένη διεπαφή για τη δημιουργία και επεξεργασία γραφημάτων γνώσης, οι χρήστες θα μπορούσαν να συμπληρώσουν εύκολα και γρήγορα οντότητες όπως θέμα, δράση, αντικείμενο, ώρα και θέση. Ο σχεδιασμός θα μπορούσε να είναι παρόμοιος με το γράφημα γνώσης Google, φιλικό προς τον χρήστη και να ενημερώνεται δυναμικά. Το μόνο μειονέκτημα αυτής της

λογικής, είναι πως όσοι θέλουν να διαμοιράσουν ψευδείς ειδήσεις, θα έχουν ίση πρόσβαση στο εργαλείο αυτό.

4. ΥΛΟΠΟΙΗΣΗ ΤΕΧΝΙΚΩΝ NLP ΚΑΙ ΣΥΓΚΡΙΣΗ

4.1 Περιγραφή συστήματος:

Μελετώντας το state-of-the-art των ερευνητικών εργασιών για εντοπισμό ψευδών ειδήσεων με χρήση τεχνικών επεξεργασίας φυσικής γλώσσας, εύλογα καταλήγει κανείς ότι αυτή η ερευνητική περιοχή ενσωματώνει ποικίλες προσεγγίσεις και διαφορετικές λύσεις που μελετώνται ως προς την αποτελεσματικότητά τους, τα πλεονεκτήματα και τις προκλήσεις που αντιμετωπίζουν.

Βασικό μας ερώτημα καθ' όλη τη διάρκεια αυτής της μελέτης ήταν ποιες από αυτές τις τεχνικές και αλγορίθμους τους οποίους μελετούσαμε, θα είχαν τα καλύτερα αποτελέσματα υπό τις ίδιες συνθήκες (κοινή βάση δεδομένων). Έτσι λοιπόν αποφασίσαμε να συνδυάσουμε μερικές από αυτές τις προσεγγίσεις σε ένα εποπτευόμενο σύστημα εντοπισμού ψευδών ειδήσεων και να καταγράψουμε τα αποτελέσματα. Το σύστημα αυτό περιγράφεται αναλυτικά στις επόμενες υποενότητες.

Τα δεδομένα προέρχονται από πραγματικές δημοσιεύσεις (tweets), που προέρχονται από το σύστημα κοινωνικής δικτύωσης Twitter [65], τα οποία είναι επεξεργασμένα, ενώ τους έχει τοποθετηθεί ετικέτα σχετικά με την εγκυρότητά τους, όπως περιγράφεται στην ενότητα 4.4. Αποθηκεύονται σε αρχεία τύπου .tsv τα οποία διαφέρουν μεταξύ τους ως προς τα tweets που περιέχουν. Η διαφοροποίηση αυτή έγκειται σε δύο παράγοντες: α) Την ύπαρξη μοναδικών ή Ν-ότυπων tweets και β) το πλήθος των ετικετών, δηλαδή αν θα είναι δύο (αληθές, ψευδές) ή τρείς (αληθές, ψευδές, απροσδιόριστο). Έτσι, καταλήγουμε σε τέσσερα διαφορετικά αρχεία.

Η επεξεργασία των δεδομένων ξεκινά με την επεξεργασία του συναισθήματος κάθε tweet του εκάστοτε αρχείου. Η επεξεργασία συναισθήματος συμπεραίνει αν το εκάστοτε tweet έχει «θετικό» ή «αρνητικό» συναίσθημα.

Εφόσον η επεξεργασία συναισθήματος των tweets τελείωσε, διαχωρίζονται τα δεδομένα, με αναλογία 70%/30%, στα δεδομένα εκπαίδευσης του συστήματος (training set) και στα δεδομένα που θα χρησιμοποιηθούν για τον έλεγχο της απόδοσης του συστήματος (test set).

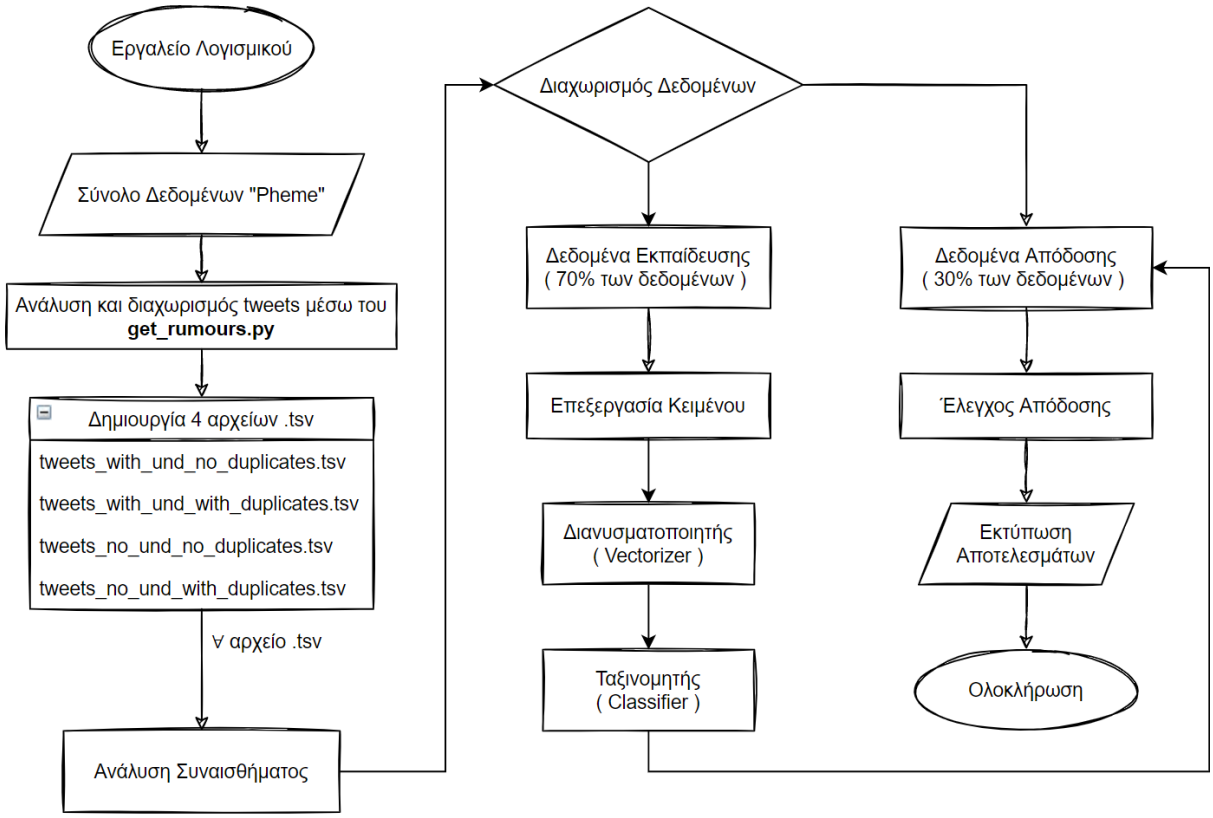
Κατά την εκπαίδευση του συστήματος με το training set γίνεται χρήση διοχέτευσης (pipelining), όπου αρχικά τα δεδομένα δέχονται και γλωσσική επεξεργασία. Πιο συγκεκριμένα γίνονται αναγνώριση λεξικών μονάδων (tokenization) και λημματοποίηση (lemmatization) σε αναπαράσταση σάρωσης λέξεων (bag of words). Δεύτερον, ακολουθεί ο διανυσματοποιητής (vectorizer) και τρίτον ο ταξινομητής (classifier).

Σε αυτό το σημείο γίνεται και η διαφοροποίηση των εκάστοτε συστημάτων εκπαίδευσης. Χρησιμοποιούνται συνδυασμοί δύο διαφορετικών vectorizers και δύο διαφορετικών classifiers, καταλήγοντας έτσι, σε τέσσερις διαφορετικούς συνδυασμούς, καθένας από τους οποίους χρησιμοποιείται αντίστοιχα σε κάθε σύστημα, ώστε να παραχθούν εν τέλει οι διαφορετικές μετρήσεις.

Εφόσον εκπαιδευτεί το σύστημα, χρησιμοποιούμε τα tweets του test set, ώστε να ελέγξουμε την απόδοσή του, με τα 16 διαφορετικά αποτελέσματα να εκτυπώνονται στην οθόνη μας.

Παρακάτω φαίνεται η αρχιτεκτονική του συστήματος και ακολουθεί η λεπτομερής περιγραφή του:

4.2 Αρχιτεκτονική Συστήματος:



Σχήμα 14: Αρχιτεκτονική Συστήματος

4.3 Επιλογή και διαχωρισμός tweets:

Αρχικά, επιλέξαμε να χρησιμοποιήσουμε το PHEME Dataset [15], το οποίο περιέχει μερικές χιλιάδες tweets που έχουν εξαχθεί από το Twitter μέσω του TwitterAPI. Επιλέξαμε το συγκεκριμένο dataset διότι αποτελείται από χιλιάδες πραγματικά tweets στα οποία έχουν ήδη τοποθετηθεί ετικέτες. Είναι διαχωρισμένα σε δύο κατηγορίες οι οποίες ονομάζονται: 1) Φήμες (rumours) και 2) Μη-φήμες (non-rumours). Εμάς μας ενδιαφέρουν οι φήμες (πάνω από 2.400 tweets), όπου για την οργάνωση και κατάταξή τους δημιουργήσαμε το πρόγραμμα “get_rumours.py” σε γλώσσα Python.

Το πρόγραμμα αρχικά παρσάρει όλους του υποφακέλους του συνόλου δεδομένων (γραμμή 64), και αρχικοποιεί ένα-ένα όλα τα προεπεξεργασμένα tweets που βρίσκονται σε μορφή .json στη μεταβλητή data (γραμμή 69).

```
57  # Read directories of incidents
58  root, dirs, files = os.walk("./").next()
59
60  # Save tweets in .tsv file
61  ftsv = open("tweets_all.tsv", 'w')
62  ftsv.write("label\tmessage\tlength\tpunct\n")
63
64  for i in dirs:
65      root_, tweets, files_ = os.walk(i+"/rumours/").next()
66
67      for tweet in tweets:
68          with open(i+"/rumours/"+tweet+"/annotation.json") as f:
69              data = json.load(f)
70          f.close()
71
72          text = data.get('category')
73          label = get_tweet_veracity(data)
74
75          if (label == "true" or label == "false" or label =="und"):
76              ftsv.write(label+"\t"+text+"\t"+str(len(text))+"\t"+str(get_punct_count(text))+"\n")
77
78  ftsv.close()
```

Σχήμα 15: Μέρος κώδικα get_rumours.py

Κατόπιν, αφού έχει δημιουργηθεί το αρχείο “tweets_all.tsv” (γραμμή 61), προσθέτουμε κάθε tweet ανά γραμμή αρχείου (γραμμή 76) με συγκεκριμένη μορφή. Τα tweets του αρχείου είναι χωρισμένα με tab (\t) ως εξής:

```
4   def get_tweet_veracity(data, string = True):
5       if 'misinformation' in data.keys() and 'true'in data.keys():
6           if int(data['misinformation'])==0 and int(data['true'])==0:
7               if string:
8                   label = "und"
9               else:
10                  label = 2
11          elif int(data['misinformation'])==0 and int(data['true'])==1 :
12              if string:
13                  label = "true"
14              else:
15                  label = 1
16          elif int(data['misinformation'])==1 and int(data['true'])==0 :
17              if string:
18                  label = "false"
19              else:
20                  label = 0
21          elif int(data['misinformation'])==1 and int(data['true'])==1:
22              #print ("Wow! They both are 1!")
23              #print(data['misinformation'])
24              #print(data['true'])
25              label = None
26
27      elif 'misinformation' in data.keys() and 'true' not in data.keys():
28          # all instances have misinfo label but don't have true label
29          if int(data['misinformation'])==0:
30              if string:
31                  label = "und"
32              else:
33                  label = 2
34          elif int(data['misinformation'])==1:
35              if string:
36                  label = "false"
37              else:
38                  label = 0
39
40      elif 'true' in data.keys() and 'misinformation' not in data.keys():
41          print ('Has true not misinformation')
42          label = None
43      else:
44          print('No annotations')
45          label = None
46
47      return label
```

Σχήμα 16: Συνάρτηση get_tweet_veracity()

label: true or false or und (undefined) (γραμμή 73)
text: το κείμενο του tweet (γραμμή 72)
length: len(text) (γραμμή 76)
punct: τα σημεία στίξης και σύμβολα του κειμένου (γραμμή 76)

Για να πάρουμε την ετικέτα (label) των tweets, τρέχουμε τη συνάρτηση get_tweet_veracity(), όπου γίνεται διαχωρισμός των δεδομένων που βρίσκονται μέσα στο κάθε .json αρχείο (που περιέχει το προεπεξεργασμένο tweet), και επιστρέφεται η ανάλογη τιμή στη μεταβλητή label (γραμμή 47).

Οι τιμές που μπορεί να λάβει είναι true, false ή und.

Για να βρούμε τον αριθμό των σημείων στίξης που θα μας χρειαστεί για τη συνέχεια, τρέχουμε τη συνάρτηση get_punct_count() όπως φαίνεται παρακάτω:

```
49  def get_punct_count(str):
50      count = 0
51      for i in range (0, len (str)):
52          # Checks whether given character is a punctuation mark
53          if str[i] in ('!', "," ,"\'" ,";", ":" ,"\"", "-", "+", "-" ,"?", "`", "<", ">", "&", "#", "%", "(", ")", "/", "=", "@", "`", "{", "}", "|", "~"):
54              count = count + 1
55      return count
```

Σχήμα 17: Συνάρτηση get_punct_count()

Τέλος, όλα τα tweets που μας ενδιαφέρουν έχουν αποθηκευτεί στο “tweets_all.tsv”.

```
80    print "Tweets from PHEME dataset where extracted and labeled successully and inserted in tweets_all.tsv"
```

Σχήμα 18: Final print of python program

Μετά το τέλος του προγράμματος, χωρίζουμε το αρχείο “tweets_all.tsv” σε 4 επιπλέον αρχεία με τη βοήθεια του εργαλείου “Sublime Text 3”.

Έτσι, δημιουργούνται τα παρακάτω αρχεία, τα οποία θα αναλυθούν και χρησιμοποιηθούν στη συνέχεια:

- tweets_no_und_no_duplicates.tsv (tweets χωρίς απροσδιοριστία, χωρίς διπλότυπα)
- tweets_no_und_with_duplicates.tsv (tweets χωρίς απροσδιοριστία, με διπλότυπα)
- tweets_with_und_no_duplicates.tsv (tweets με απροσδιοριστία, χωρίς διπλότυπα)
- tweets_with_und_with_duplicates.tsv (tweets με απροσδιοριστία, με διπλότυπα)

Επιλογή αρχείου:

Επιλέγουμε με τη σειρά τα 4 αρχεία που δημιουργήθηκαν με σκοπό τον έλεγχο της συμπεριφοράς του συστήματος με input dataset διαφορετικών χαρακτηριστικών. Όπως αναφέραμε υπάρχουν τέσσερα διαφορετικά αρχεία. Όπως αναφέρθηκε, οι διαφορές μεταξύ τους σχετίζονται με:

α) την ύπαρξη ή μη, διπλότυπων tweets. Δηλαδή αν αποδεχόμαστε το ότι το ίδιο tweet μπορεί να το έχουν δημοσιεύσει παραπάνω από ένας χρήστες.

β) την ύπαρξη ή μη, ετικετών της κατηγορίας undefined. Δηλαδή αν θα υπάρχουν μόνο οι ετικέτες True ή Fake που καθορίζουν αληθή ή ψευδή tweets ή και τα απροσδιόριστης εμπιστοσύνης tweets με ετικέτα Undefined.

4.4 Προτεινόμενη Υλοποίηση Συστήματος για Εντοπισμό Ψευδών Ειδήσεων με ΧρήσηΤεχνικών Επεξεργασίας Φυσικής Γλώσσας και Σύγκριση

Επεξεργασία συναισθήματος:

Εφόσον επιλεχθεί το αρχείο εφαρμόζουμε στα tweets sentiment analysis με τον αλγόριθμο SentimentIntensityAnalyzer [66]. Το polarity_scores [67], μας επιστρέφει μία λίστα για κάθε tweet, με 4 καταχωρημένες αξίες: 'neg', "pos', "neu', 'compound'. Οι τρείς πρώτες κατηγορίες αντιστοιχούν στις λέξεις negative, positive, neutral, όπου ορίζουν το ποσοστό που προκύπτει ως προς το πόσο αρνητική, θετική ή ουδέτερη ήταν αντίστοιχα, η πρόταση/κείμενο που εισήχθη. Ουσιαστικά, υπολογίζεται η βαθμολογία **compound** ως το άθροισμα όλων των αξιολογήσεων όλων των λέξεων που περιέχονται στο λεξικό της αγγλικής γλώσσας που έχουν κανονικοποιηθεί μεταξύ -1 (ακραίο αρνητικό) και +1 (ακραίο θετικό). Όλες αυτές οι τιμές για τα δικά μας dataset φαίνονται στις στήλες **scores** (όπου περιλαμβάνονται και οι τέσσερις τιμές) και **compound** (όπου περιλαμβάνεται μόνο η τιμή του compound).

Εμείς χρησιμοποιούμε αυτές τις τιμές ώστε να αποφανθούμε για το αν το εκάστοτε tweet είναι pos ή neg με την τοποθέτηση της ανάλογης ετικέτας σε μία νέα στήλη στο dataset με όνομα **results**. Πώς γίνεται αυτή η απόφαση;

Δημιουργούμε μία συνθήκη κατά την οποία αν η τιμή του compound είναι μεγαλύτερη ή ίση με 0, τότε το tweet θεωρείται pos, αλλιώς είναι αρνητικό neg.

label	message	length	punct	scores	compound	results
und	Multiple gunmen were involved in the Charlie Hebdo attack	61	2	{'neg': 0.279, 'neu': 0.721, 'pos': 0.0, 'compound': -0.4767}	-0.4767	neg
spam	(Alleged) ISIS militants are behind the hostage-taking in the Sydney cafe	73	3	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0.0000	pos
ham	(At least) 10 people are dead at Charlie Hebdo offices	54	2	{'neg': 0.323, 'neu': 0.677, 'pos': 0.0, 'compound': -0.6486}	-0.6486	neg
ham	(Up to) 150 people perished in the crash (144 passengers, 6 crew)	65	5	{'neg': 0.213, 'neu': 0.787, 'pos': 0.0, 'compound': -0.4019}	-0.4019	neg
ham	11 people died during the Charlie Hebdo attack	46	0	{'neg': 0.528, 'neu': 0.472, 'pos': 0.0, 'compound': -0.7717}	-0.7717	neg
ham	12 people died in connection with the Charlie Hebdo attack	58	0	{'neg': 0.456, 'neu': 0.544, 'pos': 0.0, 'compound': -0.7717}	-0.7717	neg
ham	2 police officers died during the Charlie Hebdo attack	54	0	{'neg': 0.528, 'neu': 0.472, 'pos': 0.0, 'compound': -0.7717}	-0.7717	neg
und	200 police officers at Ferguson protests on August 10	53	0	{'neg': 0.192, 'neu': 0.808, 'pos': 0.0, 'compound': -0.2263}	-0.2263	neg
und	30 shots were fired inside/on Parliament Hill	45	1	{'neg': 0.375, 'neu': 0.625, 'pos': 0.0, 'compound': -0.5574}	-0.5574	neg
spam	40-50 hostages are being held at cafe in Sydney (according to Lindt CEO)	72	3	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0.0000	pos

Σχήμα 19: Απεικόνιση πρώτων 10 σειρών για κατανόηση των περιεχομένων

Διαχωρισμός dataset:

Η διαδικασία έχει ως βασικό σκοπό της επιλογή των μεγεθών του συνόλου δεδομένων εκπαίδευσης του συστήματος (trainins set) και του συνόλου δεδομένων ελέγχου απόδοσης του συστήματος (test set). Το ποσοστό συνήθως αναγράφεται ως μία τιμή μεταξύ του μηδενός και της μονάδας. Για παράδειγμα, ένα σύνολο εκπαίδευσης μεγέθους 0,67 (67%) σημαίνει ότι το υπολειπόμενο ποσοστό 0.33 (33%) αποτελεί το σύνολο ελέγχου απόδοσης.

Για την επιλογή του ιδανικού ποσοστού πρέπει να ληφθούν υπόψιν:

- Το υπολογιστικό κόστος του συνόλου εκπαίδευσης

- Το υπολογιστικό κόστος του συνόλου ελέγχου απόδοσης
- Με μικρότερο σύνολο εκπαίδευσης, οι εκτιμήσεις των παραμέτρων έχουν μεγαλύτερη διακύμανση.
- Με μικρότερο σύνολο ελέγχου απόδοσης, οι εκτιμήσεις της απόδοσης θα έχουν μεγαλύτερη διακύμανση.
- Ο όγκος του συνόλου των δεδομένων. Εάν ο όγκος των δεδομένα είναι πολύ μικρός, τότε κανένας διαχωρισμός δεν θα δώσει ικανοποιητική διακύμανση, αν ο όγκος των δεδομένων είναι πολύ μεγάλος, δεν έχει σημασία αν επιλέξετε διαχωρισμό 80/20 ή διαχωρισμό 90/10.

Ωστόσο, οι πιο συνηθισμένοι διαχωρισμοί των συνόλων είναι:

- Σύνολο εκπαίδευσης: 80%, Σύνολο ελέγχου απόδοσης: 20%.
- Σύνολο εκπαίδευσης: 67%, Σύνολο ελέγχου απόδοσης: 33%.
- Σύνολο εκπαίδευσης: 50%, Σύνολο ελέγχου απόδοσης: 50%.

Στην περίπτωσή μας επιλέξαμε να διαχωρίσουμε τα δεδομένα μας σε ποσοστό 0.7 (70%) για το σύνολο δεδομένων εκπαίδευσης και 0.3 (30%) για το σύνολο δεδομένων ελέγχου απόδοσης του συστήματος, καθότι το μέγεθος του συνόλου των δεδομένων μας δεν είναι πολύ μεγάλο και ο διαχωρισμός αυτός αποτελεί την πιο διαδεδομένη τεχνική.

Pipeline:

Το pipeline είναι το κύριο κομμάτι, στο οποίο διοχετεύονται οι τρεις κύριες διαδικασίες της επεξεργασίας κειμένου, της διανυσματοποίησης και της ταξινόμησης. Πιο συγκεκριμένα:

Α) Στην επεξεργασία κειμένου: Χρησιμοποιούμε tokenization και lemmatization σε bag of words (βλ. παρακάτω). Η επεξεργασία αυτή είναι το πρώτο σημαντικό βήμα κατά τη μοντελοποίηση δεδομένων κειμένου. Κατά την περιγραφή των όρων, μέσω κατάλληλων διαδικασιών, προετοιμάζεται το «λεξιλόγιο», το οποίο αναφέρεται στο σύνολο των μοναδικών λέξεων και συμβόλων, που θα χρησιμοποιηθούν στη συνέχεια της διαδικασίας.

Β) Διανυσματοποίηση: Πρόκειται για τη διαδικασία μετατροπής του κειμένου σε αριθμητική αναπαράσταση. Χρησιμοποιούμε δύο διαφορετικούς vectorizers. Τους count-vectorizer και tfidf-vectorizer, καθότι είναι πολύ δημοφιλείς. [27] [28]

Γ) Ταξινόμηση: Ένας καλός ταξινομητής είναι σε θέση να εντοπίσει μοτίβα στις κατανομές λέξεων και να μάθει να προβλέπει το συναίσθημα ενός κειμένου με βάση το ποιες λέξεις εμφανίζονται και πόσες φορές το κάνουν. Χρησιμοποιούμε δύο διαφορετικούς classifiers. Τους support vector classifier (SVC) και Multinomial Naïve Bayes καθότι και αυτοί είναι πολύ δημοφιλείς και αποτελεσματικοί. [29]

Περιγραφή όρων:

Sentiment Analysis:
Η ανάλυση συναισθημάτων είναι η διαδικασία προσδιορισμού του εάν ένα κομμάτι της γραφής είναι θετικό, αρνητικό ή ουδέτερο. Ένα σύστημα ανάλυσης συναισθημάτων για ανάλυση κειμένου συνδυάζει τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και μηχανικής μάθησης για να αποδώσει σταθμισμένες βαθμολογίες συναισθημάτων στις οντότητες, τους τίτλους, τα θέματα και τις κατηγορίες μέσα σε μια πρόταση ή φράση.

Η βασική ανάλυση συναισθημάτων των εγγράφων κειμένου ακολουθεί μια απλή διαδικασία:

- Διαχωρίζουμε κάθε έγγραφο κειμένου σε συστατικά μέρη του (προτάσεις, φράσεις, λέξεις και σημεία στίξης).
- Προσδιορίζουμε κάθε φράση και συστατικό που φέρνει συναίσθημα.
- Αντιστοιχούμε μια βαθμολογία συναισθημάτων σε κάθε φράση και συστατικό (-1 έως +1).
- Προαιρετικά: Συνδυάζουμε τις βαθμολογίες για πολυεπίπεδη ανάλυση συναισθημάτων.

Όπως θα δείτε, η βασική τεχνολογία είναι πολύ περίπλοκη. Αλλά για μια απλή εξήγηση της ανάλυσης συναισθημάτων, σκεφτείτε αυτές τις προτάσεις:

i. Terrible pitching and awful hitting led to another crushing loss.
ii. Bad pitching and mediocre hitting cost us another close game.

Και οι δύο προτάσεις συζητούν ένα παρόμοιο θέμα, την απώλεια ενός παιχνιδιού μπέιζμπολ. Αλλά εσείς, ως άνθρωπος που τα διαβάζετε, μπορείτε να δείτε καθαρά ότι ο τόνος της πρώτης πρότασης είναι πολύ πιο αρνητικός.

Ο εγκέφαλός σας το καταλαβαίνει αναζητώντας και ερμηνεύοντας φράσεις που φέρνουν συναισθήματα-δηλαδή λέξεις και φράσεις που φέρουν έναν τόνο ή μία συγκεκριμένη άποψη. Αυτά συνήθως εμφανίζονται ως συνδυασμοί επιθέτων-ουσιαστικών. Στα παραπάνω παραδείγματα, οι φράσεις που φέρνουν συναίσθημα είναι:

```
Terrible pitching|awful hitting|crushing loss

Bad pitching|mediocre hitting|close game
```

Έχετε συναντήσει λέξεις όπως αυτές πολλές χιλιάδες φορές κατά τη διάρκεια της ζωής σας σε μια ποικιλία πλαισίων. Και από αυτές τις εμπειρίες, έχετε μάθει να καταλαβαίνετε τη δύναμη κάθε επιθέτου, λαμβάνοντας πληροφορίες και σχόλια στην από την οικογένεια, τους δασκάλους και τους συνομήλικούς σας.

Όταν διαβάζετε τις παραπάνω προτάσεις, ο εγκέφαλός σας αντλεί από τις συσσωρευμένες γνώσεις σας για να προσδιορίσει κάθε φράση που φέρνει συναίσθημα και να ερμηνεύσει την αρνητικότητα ή τη θετικότητά τους. Συνήθως αυτό συμβαίνει υποσυνείδητα. Για παράδειγμα, γνωρίζετε ενστικτωδώς ότι ένα παιχνίδι που καταλήγει

σε "συντριβή" έχει υψηλότερη διαφορά στο σκορ από το "κλειστό παιχνίδι", επειδή καταλαβαίνετε ότι η "σύνθλιψη" έχει μεγαλύτερη βαρύτητα και “ισχύ”.

Η ανάλυση συναισθημάτων στον υπολογιστή λειτουργεί (σχεδόν) με τον ίδιο τρόπο. Σε μεγάλο βαθμό με τον τρόπο που ο εγκέφαλός σας θυμάται τις λέξεις που συναντάτε κατά τη διάρκεια της ζωής σας και το σχετικό «βάρος συναισθήματος», ένα βασικό σύστημα ανάλυσης συναισθημάτων βασίζεται σε μια βιβλιοθήκη συναισθημάτων για να κατανοήσει τις φράσεις που προκαλούν συναισθήματα.

Οι βιβλιοθήκες συναισθημάτων είναι πολύ μεγάλες συλλογές επιθέτων (καλός, υπέροχος, φοβερός, φρικτός) και φράσεις (καλό παιχνίδι, υπέροχη ιστορία, φοβερή απόδοση, φρικτή παράσταση) που έχουν βαθμολογηθεί από ανθρώπους. Αυτή η χειροκίνητη βαθμολόγηση συναισθημάτων είναι μια δύσκολη διαδικασία, επειδή όλοι οι εμπλεκόμενοι πρέπει να καταλήξουν σε κάποια συμφωνία σχετικά με το πόσο ισχυρή ή αδύναμη πρέπει να είναι κάθε βαθμολογία σε σχέση με τις άλλες βαθμολογίες. Εάν ένα άτομο δίνει "κακό" βαθμολογία συναισθήματος -0,5, αλλά ένα άλλο άτομο δίνει "απαίσιο" την ίδια βαθμολογία, το σύστημα ανάλυσης συναισθημάτων θα καταλήξει στο συμπέρασμα ότι και οι δύο λέξεις είναι εξίσου αρνητικές.

Μόλις προετοιμαστούν οι βιβλιοθήκες συναισθημάτων, οι μηχανικοί λογισμικού γράφουν μια σειρά κατευθυντήριων γραμμών («κανόνες») για να βοηθήσουν τον υπολογιστή να αξιολογήσει το συναίσθημα που εκφράζεται για μια συγκεκριμένη οντότητα (ουσιαστικό ή αντωνυμία) με βάση την βαθμολογία συναισθήματός της.

Εντοπίζοντας στο κείμενο τη βαθμολογία συναισθήματος μπορούμε να εντοπίσουμε “διαρροές” στο λόγο ενός χρήστη που πιθανώς θέλει να παραπληροφορήσει. Τέτοιες “διαρροές” συναισθημάτων είναι σύνηθεις στα κείμενα που θέλουν να παραπληροφορήσουν καθώς γίνεται χρήση λέξεων με έντονη συναισθηματικό “βάρος”, ώστε να τραβήξουν τη προσοχή του αναγνώστη και να τον παρασύρουν, ή πολλών σημείων στίξης (όπως πολλά θαυμαστικά). Ωστόσο αυτό δεν αποτελεί πάντα τρόπο για να βγάλουμε ασφαλή συμπεράσματα.

Tokenization: Είναι μια θεμελιώδης έννοια που ασχολείται με τον διαχωρισμό του κειμένου σε φράσεις / προτάσεις ή λέξεις, όπου έπειτα τα αποθηκεύει σε μια λίστα. Για παράδειγμα:

Έστω ότι το κείμενό μας είναι:

«Εδώ αναλύουμε τους παραπάνω όρους. Ας ξεκινήσουμε με το tokenization.»

Ο διαχωρισμός γίνεται ως εξής:

Πρόταση 1: Εδώ αναλύουμε τους παραπάνω όρους.
Πρόταση 2: Ας ξεκινήσουμε με το tokenization.

Lemmatization: Πρόκειται για τη διαδικασία ομαδοποίησης των διαφορετικών μορφών μίας λέξης, ώστε να μπορούν να αναλυθούν ως ενιαίο στοιχείο. Για παράδειγμα, η λέξη «υπολογιστές» είναι μία παραμορφωμένη μορφή της λέξης «υπολογιστής».

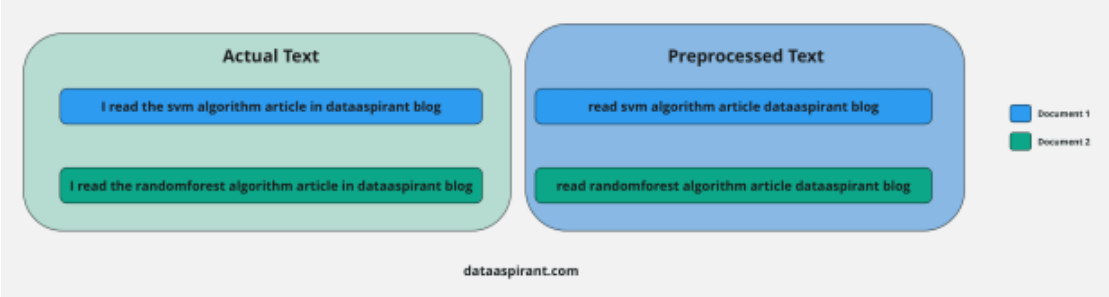
Είναι ουσιαστικά η επιστροφή μίας λέξης στη ρίζα από όπου προέρχεται γλωσσολογικά και εννοιολογικά. Ωστόσο, αν και με μια πρώτη ματιά μπορεί να φαίνεται σχετικά απλό,

παρατηρούμε ότι σε λέξεις όπως «ομιλητής» και «μιλάω» δεν μπορεί να εφαρμοστεί lemmatization ώστε από τον «ομιλητή» να φτάσουμε στο «μιλάω». Αυτό συμβαίνει καθώς εννοιολογικά οι δύο λέξεις διαφέρουν, αφού ο «ομιλητής» (ουσ.) είναι κάποιος που βγάζει λόγο, ενώ η λέξη «μιλάω» (ρήμα) είναι η πράξη της ομιλίας.

Επομένως, αυτό που κρατάμε είναι πως με τη διαδικασία του lemmatization, διατηρείται το νόημα της λέξης καθ'όλες τις μορφές της.

Bag-of-words: Όπως υποδηλώνει και το όνομα, το κόνσεπτ είναι η δημιουργία μίας τσάντας λέξεων οι οποίες θα παρθούν από προτάσεις ενός κειμένου.

Για παράδειγμα:



Σχήμα 20: Παράδειγμα για bag of words

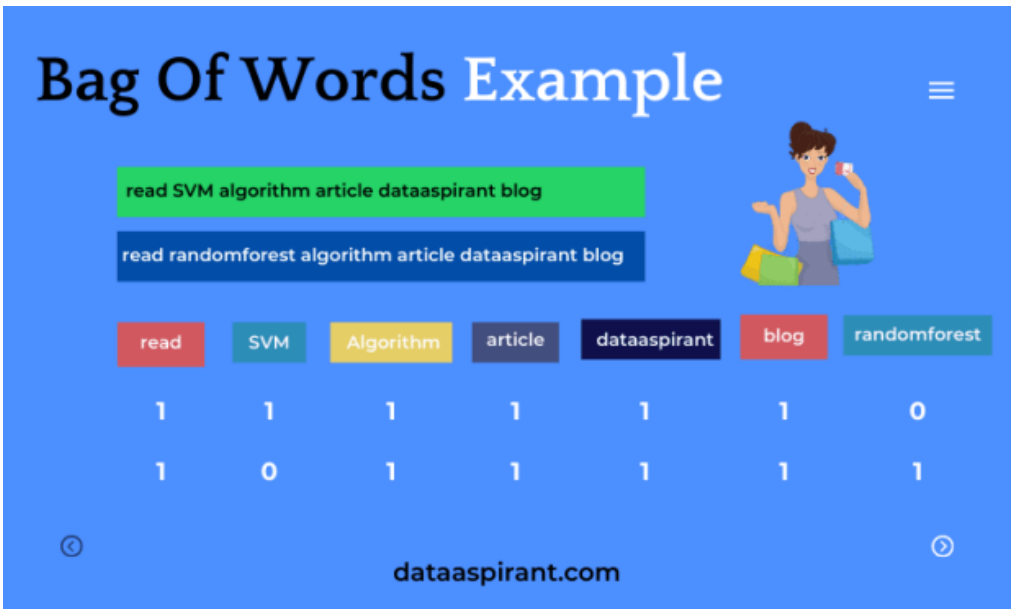
Document 1: I read the SVM algorithm article in dataaspirant blog
Document 2: I read the randomforest algorithm article in dataaspirant blog

Ύστερα από την προεπεξεργασία, θα καταλήξουμε στις παρακάτω προτάσεις:

Document 1: read SVM algorithm article dataaspirant blog
Document 2: read randomforest algorithm article dataaspirant blog

Εδώ θα δημιουργηθεί ένα λεξιλόγιο που θα αποτελείται από όλες τις λέξεις των παραπάνω προτάσεων.

Αποτελεί δηλαδή ένα μέρος όπου διατηρείται ένα αρχείο για την εμφάνιση / παρουσία μίας λέξης σε μία συγκεκριμένη πρόταση. Παρακάτω φαίνεται πώς γίνεται η μετατροπή των λέξεων σε αριθμούς:



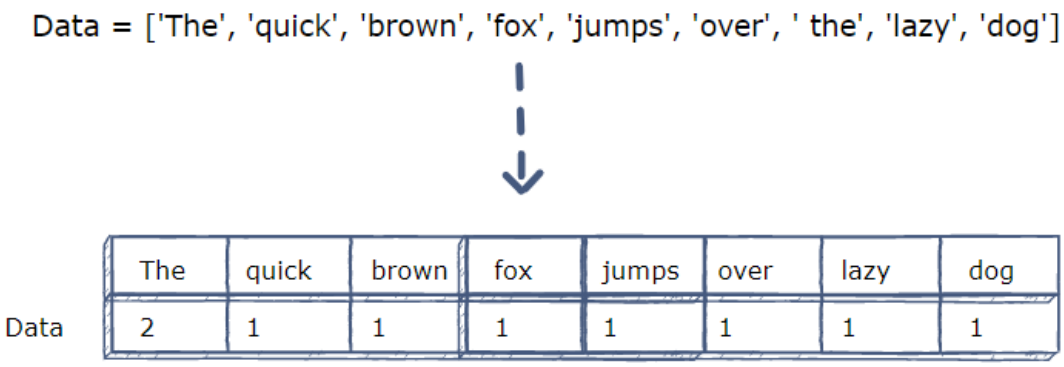
Σχήμα 21: Μετατροπή λέξεων σε αριθμούς σε bag of words

Είναι η απλούστερη μορφή αναπαράστασης λέξεων με τη μορφή αριθμών. Μετατρέπουμε τις λέξεις σε ψηφία επειδή το σύστημα χρειάζεται τις πληροφορίες με τη μορφή αριθμών, αλλιώς δεν θα μπορεί να επεξεργαστεί τα δεδομένα. Ένας αριθμός δηλώνεται ως κωδικοποιημένη τιμή έναντι της λέξης. Αυτός είναι ο αριθμός των φορών που η λέξη έχει εκφραστεί στην πρόταση.

Count-Vectorizer: Για να χρησιμοποιηθούν δεδομένα κειμένου για προγνωστική μοντελοποίηση, πρέπει να γίνει ανάλυση (parsing) του κειμένου ώστε να αφαιρεθούν συγκεκριμένες λέξεις. Όπως αναφέραμε παραπάνω, αυτή είναι η διαδικασία του tokenization. Οι λέξεις αυτές στη συνέχεια κωδικοποιούνται ως ακέραιοι για να χρησιμοποιηθούν ως είσοδοι σε αλγόριθμους μηχανικής μάθησης. Η διαδικασία αυτή ονομάζεται Εξαγωγή Χαρακτηριστικών ή Διανυσματοποίηση (Vectorization).

Ο Count-Vectorizer (μέρος της βιβλιοθήκης Scikit-learn) χρησιμοποιείται για τη μετατροπή μίας συλλογής εγγράφων κειμένου σε ένα διάνυσμα αριθμού των όρων / tokens. Επιτρέπει επίσης την προεπεξεργασία δεδομένων κειμένου πριν από τη δημιουργία του vector representation. Αυτή η λειτουργικότητα το καθιστά ένα εξαιρετικά ευέλικτο χαρακτηριστικό αναπαράστασης χαρακτηριστικών για ένα κείμενο.

Για παράδειγμα:



Σχήμα 22: Παράδειγμα Count Vectorizer

TF-IDF: Πρόκειται για ένα στατιστικό μέτρο που αξιολογεί πόσο σχετική είναι μια λέξη με ένα έγγραφο σε μια συλλογή εγγράφων. Αυτό γίνεται πολλαπλασιάζοντας δύο μετρήσεις: πόσες φορές μια λέξη εμφανίζεται σε ένα έγγραφο και τη αντίστροφη συχνότητα του κειμένου της λέξης σε ένα σύνολο εγγράφων. Έχει πολλές χρήσεις, αυτή που μας ενδιαφέρει και γι' αυτό χρησιμοποιούμε αυτόν τον vectorizer είναι η αυτοματοποιημένη ανάλυση κειμένου, καθώς και είναι πολύ χρήσιμο για τη βαθμολόγηση λέξεων σε αλγόριθμους μηχανικής εκμάθησης για την επεξεργασία φυσικής γλώσσας (NLP).

Λειτουργεί αυξάνοντας αναλογικά τον αριθμό των φορών που μια λέξη εμφανίζεται σε ένα έγγραφο, αλλά αντισταθμίζεται από τον αριθμό των εγγράφων που περιέχουν τη λέξη. Έτσι, λέξεις που είναι κοινές σε κάθε έγγραφο, κατατάσσονται χαμηλά, παρόλο που μπορεί να εμφανίζονται πολλές φορές, δεδομένου ότι δεν έχουν ιδιαίτερη σημασία για αυτό το έγγραφο.

Το TF-IDF μια λέξης σε ένα έγγραφο υπολογίζεται πολλαπλασιάζοντας δύο διαφορετικές μετρήσεις:

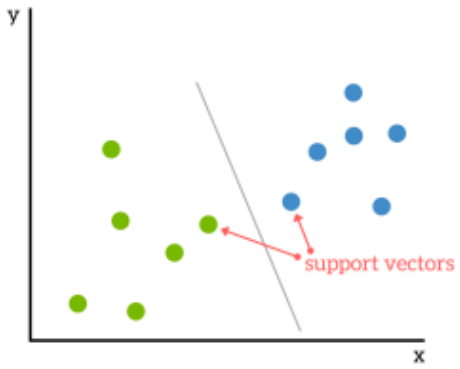
- Term Frequency μιας λέξης σε ένα έγγραφο: Υπάρχουν διάφοροι τρόποι υπολογισμού αυτής της συχνότητας, με τον απλούστερο να είναι απλά ο αριθμός εμφάνισης της λέξης στο έγγραφο. Στη συνέχεια, υπάρχουν τρόποι για να ρυθμιστεί η συχνότητα, κατά μήκος ενός εγγράφου ή από την αρχική συχνότητα της πιο συχνής λέξης σε ένα έγγραφο.
- Inverse Document Frequency λέξης σε ένα σύνολο εγγράφων: Εδώ θα βρεθεί πόσο κοινή ή σπάνια είναι μια λέξη σε ολόκληρο το σύνολο εγγράφων. Όσο πιο κοντά είναι στο 0, τόσο πιο κοινή είναι η λέξη. Αυτή η μέτρηση μπορεί να υπολογιστεί λαμβάνοντας τον συνολικό αριθμό εγγράφων, διαιρώντας τον με τον αριθμό των εγγράφων που περιέχουν μια λέξη και υπολογίζοντας τον λογάριθμο.
- Έτσι, εάν η λέξη είναι πολύ κοινή και εμφανίζεται σε πολλά έγγραφα, αυτός ο αριθμός θα πλησιάσει το 0. Διαφορετικά, θα πλησιάσει το 1.

Ο πολλαπλασιασμός αυτών των δύο αριθμών έχει ως αποτέλεσμα τη βαθμολογία TF-IDF μιας λέξης σε ένα έγγραφο. Όσο υψηλότερη είναι η βαθμολογία, τόσο πιο σχετική είναι η λέξη στο συγκεκριμένο έγγραφο.

Text Classification: Η ταξινόμηση κειμένου είναι μια αυτοματοποιημένη διαδικασία ταξινόμησης κειμένου σε προκαθορισμένες κατηγορίες. Μπορούμε να ταξινομήσουμε e-mails σε ανεπιθύμητα ή μη ανεπιθύμητα, τα άρθρα ειδήσεων διαφορετικών κατηγοριών όπως πολιτική, χρηματιστήριο, αθλητικά κ.λπ.

Αυτό μπορεί να γίνει με τη βοήθεια της Επεξεργασίας Φυσικής Γλώσσας και διαφορετικών Αλγορίθμων Ταξινόμησης, όπως Naive Bayes, Support Vector Machine μέσω Python.

Support Vector Machine: Το SVM είναι ένας εποπτευόμενος αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για σκοπούς ταξινόμησης. Βασίζεται στην ιδέα εύρεσης ενός υπερπλάνου (hyperplane) που διαιρεί καλύτερα ένα σύνολο δεδομένων σε δύο κατηγορίες, όπως φαίνεται στην παρακάτω εικόνα:



Σχήμα 23: Παράδειγμα Support Vector Machine

Naive Bayes: Είναι πιθανολογικός αλγόριθμος που βασίζεται στο θεώρημα του Bayes για να υπολογίσει την ετικέτα (tag) ενός κειμένου (αν είναι για παράδειγμα ένα μέρος ειδησεογραφικού άρθρου, ή μία κριτική σε μαγαζί). Με τον όρο «πιθανολογικός» εννοείται πως υπολογίζει την πιθανότητα κάθε ετικέτας του κειμένου, και στο τέλος κρατάει και εμφανίζει την υψηλότερη.

4.5 Προβλέψεις και αποτελέσματα:

Στη Μηχανική Μάθηση γίνεται μια προσπάθεια δημιουργίας ενός μοντέλου για την πρόβλεψη των τελικών test_data. Έτσι, χρησιμοποιούνται τα training_data (70% του input dataset) για την εκπαίδευση του μοντέλου, το οποίο εν τέλει θα προβλέψει αν τα

tweets που περιέχονται στα test_data είναι αληθή, ψευδή ή απροσδιόριστα – στα datasets που περιέχουν την ετικέτα “und”.

Κατά την εκτέλεση του προγράμματός μας, γίνεται η πρόβλεψη με το testing set και εκτυπώνονται τα ποσοστά επιτυχίας. Όπως αναφέρθηκε, έχουν δημιουργηθεί τα παρακάτω 4 datasets:

- tweets_with_und_no_duplicates.tsv
- tweets_with_und_with_duplicates.tsv
- tweets_no_und_no_duplicates.tsv
- tweets_no_und_no_duplicates.tsv

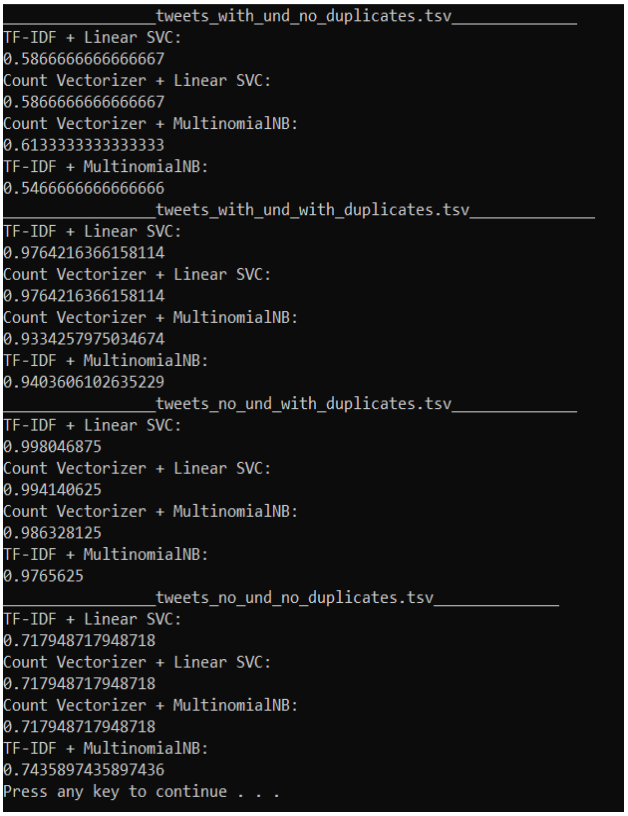
Για κάθε ένα από τα παραπάνω αρχεία, εκτελούνται pipelines από συνδυασμό κάθε φορά ενός vectorizer και ενός classifier. Έτσι προκύπτουν 4 συνδυασμοί:

- TF-IDF + Linear SVC
- Count Vectorizer + Linear SVC
- Count Vectorizer + Multinomial Naive Bayes
- TF-IDF + Multinomial Naive Bayes

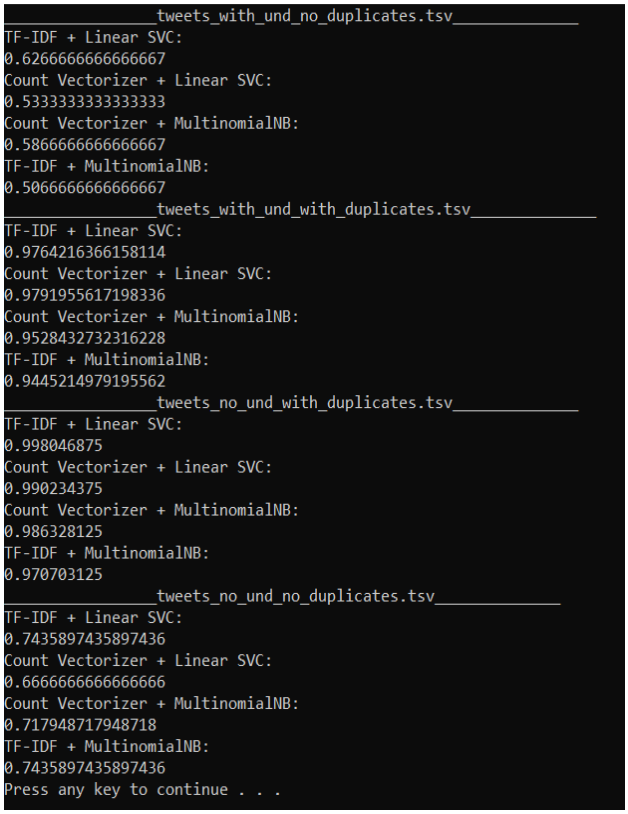
Για τους ίδιους ακριβώς συνδυασμούς έχει εκτελεστεί και sentiment analysis. Το αποτέλεσμα του sentiment analysis, δηλαδή η ετικέτα “pos” ή “neg” ή οποία χαρακτηρίζει το εκτιμόμενο συναίσθημα της πρότασης, συνυπολογίζεται μαζί με την επεξεργασία των λέξεων των tweets στο training set. Παρατηρείται ότι σε κάποιες περιπτώσεις η διαφορά είναι αξιοσημείωτα βελτιωμένη, ωστόσο σε κάποιες άλλες έχουμε τα αντίθετα αποτελέσματα. Αυτό είναι εύλογο καθώς υπάρχουν δύο σενάρια:

- Το θετικό ή αρνητικό συναίσθημα δεν μπορεί να ταυτιστεί πλήρως με το αν ταυτόχρονα το κείμενο περιέχει real ή fake news. Δηλαδή, αν το αποτέλεσμα της ανάλυσης συναισθήματος μας υποδείξει ότι είναι positive δεν είναι αυτομάτως βέβαιη ένδειξη ότι το περιεχόμενο είναι fake ή real.
- Τα αποτελέσματα της ανάλυσης συναισθήματος δεν είναι εγγυημένα σωστά. Δεν υπάρχει 100% επιτυχία στον NLTK Sentiment Analyzer, οπότε και αυτό δημιουργεί «θόρυβο» στα δεδομένα μας.

Παρακάτω μπορείτε να παρατηρήσετε τα αποτελέσματα της εκτέλεσης του προγράμματός μας με τα παραπάνω pipelines, με ή χωρίς sentiment analysis:



Σχήμα 24: Αποτελέσματα με sentiment analysis



Σχήμα 25: Αποτελέσματα χωρίς sentiment analysis

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Αναγιγνώσκοντας τα αποτελέσματα του εργαλείου που αναπτύξαμε, πρωτίστως αντιλαμβανόμαστε πως με τη χρήση των .tsv αρχείων με διπλότυπα tweets ως προεπιλεγμένο dataset καταλήγουμε σε πολύ καλύτερα ποσοστά ορθότητας στην επιλογή της ετικέτας (label).

Ας χωρίσουμε και ας σχολιάσουμε αρχικά τα αποτελέσματα σε αυτά με sentiment analysis και εκείνα χωρίς:

Αποτελέσματα με Sentiment Analysis:

tweets_with_und_no_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	58.7	54.7
Count	58.7	61.3

tweets_with_und_with_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	97.6	94
Count	97.6	93.3

Συγκρίνοντας τα αποτελέσματα στα datasets με undefined ετικέτες, παρατηρούμε ότι η διαφορά στην απόδοση των αρχείων με διπλότυπα ή χωρίς κυμαίνεται μεταξύ 32% και 42.9%.

tweets_no_und_with_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	99.8	97.7
Count	99.4	98.6

tweets_no_und_no_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	71.8	74.3
Count	71.8	71.8

Στα αποτελέσματα στα datasets χωρίς undefined ετικέτες, παρατηρούμε ότι η διαφορά στην απόδοση των αρχείων με διπλότυπα ή χωρίς κυμαίνεται μεταξύ 23.4% και 27.6%.

Αποτελέσματα χωρίς Sentiment Analysis:

tweets_with_und_no_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	62.7	50.6
Count	53.3	58.6

tweets_with_und_with_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	97.6	94.4
Count	97.9	95.3

Συγκρίνοντας τα αποτελέσματα στα datasets με undefined ετικέτες, παρατηρούμε ότι η διαφορά στην απόδοση των αρχείων με διπλότυπα ή χωρίς κυμαίνεται μεταξύ 31.7% και 47.3%.

tweets_no_und_with_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	99.8	97
Count	99	98.6

tweets_no_und_no_duplicates.tsv		
Vectorizer/Classifier	Linear SVC	MultinomialNB
TF-IDF	74.3	74.3
Count	66.7	71.8

Στα αποτελέσματα στα datasets χωρίς undefined ετικέτες, παρατηρούμε ότι η διαφορά στην απόδοση των αρχείων με διπλότυπα ή χωρίς κυμαίνεται μεταξύ 22.7% και 33.1%.

Από τα παραπάνω, δύναται να συμπεράνουμε ότι η χρήση των αρχείων με διπλότυπα tweets αυξάνει κατά πολύ την αποδοτικότητα του συστήματός μας, εφόσον η επανάληψη ίδιων tweets κατά την εκπαίδευση μοιραία το προετοιμάζει καλύτερα σε περίπτωση που ελέγξει ίδιο ή παρόμοιο tweet στο test set. Ωστόσο, αυτό αποτελεί το ρεαλιστικό σενάριο καθώς σε μία ροή ειδήσεων στο Twitter, είναι φυσιολογικό η ίδια είδηση, δηλαδή το ίδιο tweet, να έχει κοινοποιηθεί και προωθηθεί από πολλαπλούς χρήστες.

Βλέπουμε, επίσης, ότι η χρήση αρχείων με την ετικέτα “undefined” επηρεάζει αρνητικά την απόδοση του συστήματος, κάτι το οποίο ωστόσο είναι εύλογο καθώς το σύστημα πλέον έχει τρεις επιλογές ως output και όχι δύο, και έτσι αυξάνεται το περιθώριο λάθους.

Τέλος, παρατηρούμε πως η χρήση του sentiment analysis έχει διττά αποτελέσματα στην απόδοση του συστήματός μας. Αυτό οφείλεται στο γεγονός ότι η απόδοση του Sentiment Analyzer δεν έχει 100% ποσοστό επιτυχίας στα tweets με ορθώς τοποθετημένη ετικέτα. Επιπλέον, ακόμα και ορθώς να είναι τοποθετημένη η ετικέτα, δεν αποτελεί καθοριστική συνθήκη ως προς την αξιοπιστία του tweet. Η ανάλυση συναισθήματος θα μπορούσε δηλαδή να είναι πολύ πιο αποδοτική κατά την επεξεργασία κριτικών και όχι ειδήσεων.

Συνοψίζοντας, καταλήγουμε ότι δεν υφίσταται σταθερά βέλτιστη προκαθορισμένη επιλογή συστήματος, αλλά αυτή καθορίζεται και επηρεάζεται από τα χαρακτηριστικά και το πλαίσιο της δημιουργίας του συστήματος αυτού. Ένα εποπτευόμενο σύστημα με προεπεξεργασμένα tweets με την χρήση του διανυσματοποιητή TF-IDF και του

ταξινομητή Linear SVC, θα έχει εξαιρετικά ακριβή αποτελέσματα ακόμα και χωρίς τη χρήση επεξεργασίας συναισθήματος. Εντούτοις, υπάρχει κάποια αρνητικά τα οποία φανερώνουν ότι ένα τέτοιο σύστημα αν και εξαιρετικά αποδοτικό δεν αποτελεί πανάκεια στην αντιμετώπιση του προβλήματος των ψευδών ειδήσεων.

Ένα τέτοιο σύστημα απαιτεί συνεχή ενημέρωση του συνόλου δεδομένων του. Παράλληλα απαιτεί εργασία προσωπικού, το οποίο θα είναι υπεύθυνο, για κάθε εξορυγμένο tweet, να κάνει τη σχετική έρευνα αξιοπιστίας και στη συνέχεια να του αποδώσει τη κατάλληλη ετικέτα (αληθές ή ψευδές). Με τόσες χιλιάδες tweets που κοινοποιούνται καθημερινά, κάτι τέτοιο απαιτεί χιλιάδες αν όχι και εκατομμύρια άτομα προσωπικού τα οποία θα είναι υπεύθυνα για τη συνεχή ενημέρωση και συντήρηση αυτού του εργαλείου λογισμικού.

Το εργαλείο λογισμικού μας λοιπόν, μας φανερώνει τους βέλτιστους συνδυασμούς για ένα εποπτευόμενο σύστημα αναλόγως τα χαρακτηριστικά του συνόλου δεδομένων που διαθέτουμε. Ωστόσο, φανερώνει και τα μειονεκτήματα ενός εποπτευόμενου συστήματος. Ο τεράστιος όγκος πληροφοριών και δημοσιεύσεων στη σύγχρονη εποχή γεννά την ανάγκη για τη δημιουργία εξελιγμένων μη εποπτευόμενων συστημάτων εντοπισμού ψευδών ειδήσεων. Η απασχόληση τόσου προσωπικού καθίσταται αδύνατη και ένα μη εποπτευόμενο σύστημα το οποίο θα είναι θα επεξεργάζεται και θα καταλήγει σε ετικέτες αξιοπιστίας χωρίς την ανάγκη παρέμβασης ανθρώπου είναι μονόδρομος.

Μελλοντική χρήση:

Αυτό που θέλαμε να επιτύχουμε είναι να συλλέξουμε τις πιο δημοφιλείς τεχνικές για ένα εποπτευόμενο σύστημα επεξεργασίας φυσικής γλώσσας για τον εντοπισμό ψευδών ειδήσεων, να τις συνδυάσουμε και να συγκρίνουμε τα αποτελέσματα ώστε οποιοσδήποτε θελήσει να αναπτύξει ένα πιο εξελιγμένο εργαλείο λογισμικού, να έχει στη διάθεσή του τη μελέτη μας. Σε συνέχεια της δικής μας εργασίας, θα μπορούσαν να προστεθούν και περισσότερες παράμετροι ως inputs στο σύστημα, όπως για παράδειγμα το προφίλ του χρήστη που κοινοποίησε το εκάστοτε tweet, η δημιουργία βαθμολογίας αξιοπιστίας του χρήστη, καθώς και η διαδρομή που ακολούθησε το tweet μέσα στο δίκτυο του Twitter με σκοπό τον εντοπισμό αναξιόπιστων “κόμβων”/χρηστών.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Analyzing Module	Μονάδα Ανάλυσης
Back-Propagation	Οπίσθια Διάδοση
Bag-Of-Words	Αναπαράσταση-Σάρωσης-Λέξεων
Bots	Εικονικοί/Ψεύτικοι Χρήστες
Classifier	Ταξινομητής
Clustering Analysis	Ανάλυση Ομαδοποίησης
Count Vectorizer	Απαριθμητής Διανυσμάτων
Credibility Networks	Δίκτυα Αξιοπιστίας
Database	Βάση Δεδομένων
Decision Trees	Δέντρα Απόφασης
Diffusion Networks	Δίκτυα Διάχυσης
Friendship Networks	Φιλικά Δίκτυα
Gradient Clipping	Σταδιακή Αποκοπής
Interaction Networks	Δίκτυα Αλληλεπίδρασης
K-Nearest-Neighbourhood	Κ-Κοντινότεροι Γείτονες
Knowledge Networks	Δίκτυα Γνώσης
Lemmatization	Λημματοποίηση
Linguistic Features	Γλωσσικά Χαρακτηριστικά
Link	Σύνδεσμος
Logistic Regression	Λογιστική Παλινδρόμηση
Multinomial Naive Bayes	Πολυωνυμικός Απλοϊκός Ταξινομητής Bayes
Multi-Layer Perceptron [Mpl]	Πολυεπίπεδος Νευρώνας “Αντίληπτρο”
N-Grams	Ν-Άδες
Named Entity Recognition	Αναγνώριση Ονοματικών Οντοτήτων
Normalization	Κανονικοποίηση
Processing Module	Μονάδα Επεξεργασίας
Random Forest Classifier	Ταξινομητής Τυχαίων Δασών
Recurrent Neural Network	Επαναλαμβανόμενο Νευρωνικό Δίκτυο
Regression	Παλινδρόμηση
Relu	Συνάρτηση Διορθωμένης Γραμμικής Μονάδας
Semantic Features	Σημασιολογικά Χαρακτηριστικά
Sentiment Features	Χαρακτηριστικά Συναισθημάτων
Service-Oriented Architecture	Υπηρεσιοστραφής Αρχιτεκτονική
Stance Networks	Δίκτυα Στάσεων
Stemming	Αποκοπή
Stop Words	Κοινότυπες Λέξεις
Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
Tf	Συχνότητα Όρου (Σε ένα κείμενο)
Tf-Idf	Συχνότητα Όρου - Άνιση Κατανομή Του Όρου (Σε Ένα Κείμενο)
Tokenization	Αναγνώριση Λεξικών Μονάδων
Twitter Crawler	Twitter Ανιχνευτής
Unsupervised News Embedding	Μη Εποπτευόμενη Ενσωμάτωση Ειδήσεων
Vectorization	Διανυσματοποίηση

Vector Space Model	Μοντέλο Διανυσματικού Χώρου
Visual Based Features	Χαρακτηριστικά Οπτικού Υλικού
Writing Style Features	Χαρακτηριστικά Τρόπου Γραφής

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

NLP	Natural Language Processing
NER	Named Entity Recognition
NBC	Naive Bayes Classifier
SVM	Support Vector Machine
KNN	K-Nearest-Neighbors
LR	Logistic Regression
RFC	Random Forest Classifier
POS	Part-of-Speech
RCFG	Probabilistic Context Free Grammars
LIWC	Linguistic Enquiry Word Count
VSM	Vector Space Model
RST	Rhetorical Structure Theory
MPL	Multi-Layer Perceptron
CRA	Centering resonance analysis
TF-IDF	Term Frequency–Inverse Document Frequency
TF	Term Frequency
ReLu	Rectified Linear Unit

ΑΝΑΦΟΡΕΣ

[1] . Open Calais: Thomson Reuters Open Calais REST API | ProgrammableWeb accessed last time on March, 2021.

[2] . Sentiment140: API - Sentiment140 - A Twitter Sentiment Analysis Tool accessed last time on March, 2021.

[3] G. C. Santia and J. R. Williams, “Buzzface: A news veracity dataset with facebook user commentary and egos,” in Twelfth International AAAI Conference on Web and Social Media, 2018.

[4] T. Mitra and E. Gilbert, “Credbank: A large-scale social media corpus with associated credibility annotations,” in Ninth International AAAI Conference on Web and Social Media, 2015.

[5] W. Y. Wang, “” liar, liar pants on fire”: A new benchmark dataset for fake news detection,” arXiv preprint arXiv:1705.00648, 2017.

[6] A. Kirilin and M. Strube, “Exploiting a speakers credibility to detect fake news,” in In Proceedings of Data Science, Journalism and Media workshop at KDD (DSJM18), 2018.

[7] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, “Fake news detection through multi-perspective speaker profiles,” in Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp. 252– 256, 2017.

[8] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, “Multimodal fusion with recurrent neural networks for rumor detection on microblogs,” in Proceedings of the 25th ACM international conference on Multimedia, pp. 795–816, ACM, 2017.

[9] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, “Eann: Event adversarial neural networks for multi-modal fake news detection,” in Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp. 849–857, ACM, 2018.

[10] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, “Mvae: Multimodal variational autoencoder for fake news detection,” in The World Wide Web Conference, pp. 2915–2921, ACM, 2019.

[11] de Marneffe, M., MacCartney, B. & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)R. Bartle, “Early MUD History» Nov. 1990; www.ludd.luth.se/mud/aber/mud-history.html [Προσπελάστηκε 15/1/08]

[12] *Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M. & Whittaker, S. (2015). And That’s A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue.*

[13] *Papacharissi, Z. & Oliveira, M. (2012).The Rhythms of News Storytelling on #Egypt. Journal of Communication. 62. pp. 266–282.*.

[14] Motivations, Methods and Metrics of Misinformation Detection: a NLP perspective. Qi Su, Mingyu Wan1,2, Xiaoqian Liu, and Chu-Ren Huang, Peking University, Beijing, China The Hong Kong Polytechnic University, Hong Kong, China.

[15] PHEME dataset for Rumour Detection and Veracity Classification.

[16] Twitter API Documentation | Docs | Twitter Developer Platform.

[17] SENTIMENTAL ANALYSIS USING VADER. interpretation and classification of… | by Aditya Beri | Towards Data Science

[18] https://scikit-learn.org/, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[19] https://scikit-learn.org/stable/supervised_learning.html, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[20] https://scikit-learn.org/stable/modules/naive_bayes.html, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[21] https://scikit-learn.org/stable/modules/svm.html, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[22] https://medium.com/swlh/text-classification-using-the-bag-of-words-approach-with-nltk-and-scikit-learn-9a731e5c4e2f, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[23] https://medium.com/swlh/text-classification-using-the-bag-of-words-approach-with-nltk-and-scikit-learn-9a731e5c4e2f, τελευταία επίσκεψη 14 Σεπτεμβρίου, 2021.

[24] Feng, S., Banerjee, R. & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 171–175.

[25] Rubin, V. & Lukoianova, T. (2014). Truth and deception at the rhetorical structure level. Journal of the American Society for Information Science and Technology, 66(5).DOI: 10.1002/asi. 23216 ·

[26] Mihalcea, R. & Strapparava, C. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. Proceedings of the ACL-IJCNLP Conference Short Papers, pp. 309–312.

[26],https://el.wikipedia.org/wiki/%CE%A8%CE%B5%CF%85%CE%B4%CE%B5%CE%AF%CF%82_%CE%B5%CE%B9%CE%B4%CE%AE%CF%83%CE%B5%CE%B9%CF%82 , τελευταία επίσκεψη 11 Ιουλίου, 2021.

[27]https://www.linkedin.com/pulse/count-vectorizers-vs-tfidf-natural-language-processing-sheel-saket/ , 2021

[28] https://medium.com/@wenxuan0923/feature-extraction-from-text-using-countvectorizer-tfidfvectorizer-9f74f38f86cc , τελευταία επίσκεψη 19 Αυγούστου, 2021.

[29] https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568 , τελευταία επίσκεψη 19 Αυγούστου, 2021.

[30] E. C. Tandoc Jr, Z. W. Lim, and R. Ling, “Defining fake news a typology of scholarly definitions,” Digital Journalism, pp. 1–17, 2017.

[31] L. Zheng and C. W. Tan, “A probabilistic characterization of the rumor graph boundary in rumor source detection,” pp. 765–769, IEEE, July 2015.

[32] J. A. Ceron-Guzman and E. Leon-Guzman, “A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election,” pp. 250–257, IEEE, Oct. 2016.

[33] J. Radianti, S. R. Hiltz, and L. Labaka, “An Overview of Public Concerns During the Recovery Period after a Major Earthquake: Nepal Twitter Analysis,” pp. 136–145, IEEE, Jan. 2016.

[34] S. Ahmed, R. Monzur, and R. Palit, “Development of a Rumor and Spam Reporting and Removal Tool for Social Media,” pp. 157–163, IEEE, Dec. 2016.

[35] M. Rajdev and K. Le, “Fake and Spam Messages: Detecting Misinformation During Natural Disasters on Social Media,” pp. 17–20, IEEE, Dec. 2015.

[36] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, “Study of hoax news detection using nave bayes classifier in Indonesian language,” pp. 73–78, IEEE, Oct. 2017.

[37] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, “The spread of fake news by social bots,” arXiv preprint arXiv:1707.07592, 2017.

[38] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[39] C. Buntain and J. Golbeck, “Automatically Identifying Fake News in Popular Twitter Threads,” pp. 208–215, IEEE, Nov. 2017.

[40] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, “The DARPA Twitter Bot Challenge,” Computer, vol. 49, pp. 38–46, June 2016.

[41] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, “Predicting the Political Alignment of Twitter Users,” pp. 192–199, IEEE, Oct. 2011.

[42] A. Boutet, D. Frey, R. Guerraoui, A. Jegou, and A.-M. Kermarrec, “WHATSUP: A Decentralized Instant News Recommender,” pp. 741–752, IEEE, May 2013.

[43] J. A. Ceron-Guzman and E. Leon-Guzman, “A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election,” pp. 250–257, IEEE, Oct. 2016.

[44] S. Fong, S. Deb, I.-W. Chan, and P. Vijayakumar, “An event driven neural network system for evaluating public moods from online users’ comments,” pp. 239–243, IEEE, Feb. 2014.

[45] Sahana V P, A. R. Pias, R. Shastri, and S. Mandloi, “Automatic detection of rumoured tweets and finding its origin,” pp. 607–612, IEEE, Dec. 2015.

[46] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” pp. 795–816, ACM Press, 2017.

[47] N. Hassan, F. Arslan, C. Li, and M. Tremayne, “Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster,” pp. 1803–1812, ACM Press, 2017.

[48] S. Vosoughi, M. . Mohsenvand, and D. Roy, “Rumor Gauge: Predicting the Veracity of Rumors on Twitter,” ACM Transactions on Knowledge Discovery from Data, vol. 11, pp. 1–36, July 2017.

[49] J. Ross and K. Thirunarayan, “Features for Ranking Tweets Based on Credibility and Newsworthiness,” pp. 18–25, IEEE, Oct. 2016.

[50] N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” in Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, ASIST ’15, (Silver Springs, MD, USA), pp. 82:1–82:4, American Society for Information Science, 2015.

[51] Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017): Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.

[52] Jia, Y., Wang, Y., Lin, H., Jin, X., and Cheng, X. (2016). Locally adaptive translation for knowledge graph embedding. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.

[53] Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., and ZeinalipourYazti, D. (2012). Crowdsourcing with smartphones. In IEEE Internet Comp., 36-44.

[54] Howe, J. (2006). The rise of crowdsourcing. In Wired magazine, 14: 1-4.

[55] Costel-Sergiu Atodiresei, Alexandru Tănăselea, Adrian Iftene, Identifying Fake News and Fake Users on Twitter. International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia. p. 451-461.

[56] Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang, Motivations, Methods and Metrics of Misinformation Detection: a NLP perspective. Peking University, Beijing, China, The Hong Kong Polytechnic University, Hong Kong, China, June 2020, p. 1-26.

[57] Adnan Hussein, Farzana Kabir Ahmad, Siti Sakira Kamaruddin, Content-Social Based Features for Fake News Detection Model from Twitter. International Journal of Advanced Trends in Computer Science and Engineering, December 2019, p. 2806-2810.

[58] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, Sebastian Riedel, A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. arXiv:1707.03264v2 [cs.CL], 21 May 2018, p. 1-6.

[59] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen, Automatic Deception Detection: Methods for Finding Fake News. ASIST 2015, November 6-10, 2015, St. Louis, MO, USA, p. 1-4.

[60] Bashar Al Asaad, Madalina Erascu, A Tool for Fake News Detection. West University of Timisoara, Romania 2016, p. 1-8.

[61] Naeemul Hassan, Fatma Arslan, Chengkai Li, Mark Tremayne, Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster. KDD ’17, August 13-17, 2017, Halifax, NS, Canada, p. 1-11.

[62] James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, Andreas Vlachos, Fake News Detection using Stacked Ensemble of Classifiers. EMNLP Workshop on Natural Language Processing meets Journalism, Copenhagen, Denmark, September 7, 2017, pages 80–83.

[63] Fernando Cardoso Durier da Silva, Rafael Vieira da Costa Alves, Ana Cristina Bicharra Garcia, Can Machines Learn to Detect Fake News? A Survey Focused on Social Media. 52nd Hawaii International Conference on System Sciences, 2019, pages 2764–2770.

[64] Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, Justin Hsu, Fake News Detection via NLP is Vulnerable to Adversarial Attacks. Department of Computer Science, University of Wisconsin-Madison, Madison, U.S.A., 2019, pages 1–8.

[65] https://twitter.com/ , τελευταία επίσκεψη 4 Οκτωβρίου, 2021.

[66] https://www.nltk.org/api/nltk.sentiment.html **,** τελευταία επίσκεψη 27 Σεπτεμβρίου, 2021.

[67]https://www.nltk.org/_modules/nltk/sentiment/vader.html#SentimentIntensityAnalyzer.polarity_scores , τελευταία επίσκεψη 27 Σεπτεμβρίου, 2021.

[68] Μετάφραση Google , τελευταία επίσκεψη 30 Σεπτεμβρίου, 2021.

[69] Chatbots: Τι είναι και πώς λειτουργούν; (developgreece.com), τελευταία επίσκεψη 30 Σεπτεμβρίου, 2021.