ΕΡΓΑΣΙΑ στο μάθημα «Προηγμένα Θέματα Ανάλυσης δεδομένων»

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΑΚΑΔ. ΕΤΟΣ 2024-2025

Ανίχνευση Ψευδών Ειδήσεων από Κείμενο

Μυλωνάς Αθανάσιος e21118

1. Εισαγωγή

Η ψηφιακή εποχή έχει επιφέρει πρωτοφανή αύξηση στη διακίνηση της πληροφορίας, με ειδήσεις και άρθρα να διαδίδονται μέσα σε λίγα λεπτά μέσω διαδικτύου και κοινωνικών μέσων. Ωστόσο, αυτή η ευκολία πρόσβασης έχει συνοδευτεί από την εξάπλωση ψευδών ειδήσεων (fake news). Τέτοιου είδους ειδήσεις δεν προκαλούν απλώς παραπληροφόρηση, αλλά δύνανται να επηρεάσουν κοινωνικά και πολιτικά ζητήματα, να εντείνουν την πόλωση και να υπονομεύσουν την εμπιστοσύνη στην ενημέρωση. Στόχος αυτής της εργασίας είναι να αναπτύξει και να αξιολογήσει μια προσέγγιση για την αυτόματη ανίχνευση ψευδών ειδήσεων, συνδυάζοντας σύγχρονες μεθόδους μηχανικής μάθησης και αποτελεσματική ανάλυση δεδομένων. Μέσα από αυτήν την προσπάθεια, επιδιώκουμε να συνεισφέρουμε στην ενίσχυση της αξιοπιστίας της διαδικτυακής πληροφορίας, συμβάλλοντας έτσι στην καταπολέμηση της παραπληροφόρησης.

2. Ορισμός προβλήματος

Η ανίχνευση ψευδών ειδήσεων είναι μια πρόκληση που επικεντρώνεται στο να διακρίνουμε τις πραγματικές από τις ψευδείς ειδήσεις, με στόχο την καλύτερη πληροφόρηση των χρηστών. Συγκεκριμένα, το πρόβλημα που εξετάζουμε μπορεί να διατυπωθεί ως εξής: όταν δοθεί ένας τίτλος ειδήσεων, πρέπει να προβλεφθεί αν ο τίτλος αυτός αντιστοιχεί σε αληθινή είδηση ή αν πρόκειται για ψεύτικη.

Το εγχείρημα αυτό δεν είναι απλό. Οι ψευδείς ειδήσεις συχνά εμφανίζονται εξίσου επαγγελματικά γραμμένες με τις πραγματικές. Χρησιμοποιούν παρόμοια

γλώσσα, παραπέμπουν σε τρέχοντα θέματα και προέρχονται από πηγές που μοιάζουν έγκυρες. Αυτή η αοριστία κάνει την ανίχνευσή τους ιδιαίτερα δύσκολη, ιδίως σε περιπτώσεις όπου τα δεδομένα είναι ελλιπή ή περιέχουν θόρυβο.

Για να αντιμετωπιστούν αυτές οι προκλήσεις, το πρόβλημα διατυπώνεται ως μια επιβλεπόμενη διαδικασία μάθησης, όπου οι αλγόριθμοι ταξινόμησης καλούνται να διακρίνουν ψευδείς από πραγματικές ειδήσεις. Η προσέγγιση αυτή περιλαμβάνει:

- Τον καθαρισμό του θορύβου και των μη απαραίτητων πληροφοριών, ώστε να παραμείνουν μόνο τα στοιχεία που συμβάλλουν στη διάκριση.
- Την εξαγωγή σχετικών χαρακτηριστικών από τους τίτλους, όπως το περιεχόμενο των λέξεων και οι δομές των φράσεων.
- Τη χρήση ισχυρών αλγορίθμων μηχανικής μάθησης που μπορούν να διαχειριστούν πολύπλοκα μοτίβα και συσχετίσεις.

Η προσέγγιση που προτείνουμε δεν στοχεύει μόνο να απαντήσει στην ερώτηση αν μια είδηση είναι ψευδής, αλλά και να παρέχει μια αξιόπιστη και αναπαραγώγιμη διαδικασία που θα ενισχύσει την εμπιστοσύνη στην πληροφορία που καταναλώνουμε.

3. Παρουσίαση προσέγγισης/μοντέλου

Η μέθοδος που ακολουθήθηκε στηρίζεται στην ενσωμάτωση διαφορετικών τεχνικών και εργαλείων για την προετοιμασία δεδομένων, τη δημιουργία χαρακτηριστικών, και την επιλογή κατάλληλων μοντέλων μηχανικής μάθησης. Κάθε βήμα της προσέγγισης σχεδιάστηκε με γνώμονα την ενίσχυση της απόδοσης του συστήματος και τη μεγιστοποίηση της ακρίβειας στις προβλέψεις.

1. Προετοιμασία και Καθαρισμός Δεδομένων
Για να καταστούν τα δεδομένα κατάλληλα για μοντελοποίηση, εφαρμόστηκαν τεχνικές καθαρισμού και προεπεξεργασίας. Αυτές περιλάμβαναν:

- **Αφαίρεση ειδικών χαρακτήρων και stop words:** Έτσι μειώθηκε ο θόρυβος στο κείμενο, κάνοντας τα δεδομένα πιο συνεκτικά και κατανοητά για τους αλγορίθμους.
- **Μετατροπή σε πεζά γράμματα:** Διασφαλίσαμε ότι όλες οι λέξεις είχαν κοινή μορφή, αποτρέποντας τη δημιουργία περιττών χαρακτηριστικών λόγω διαφορετικής γραφής.

- **Δημιουργία numeric χαρακτηριστικών από URL:** Η εξαγωγή χαρακτηριστικών όπως το μήκος του URL, η παρουσία της λέξης “fake” και το domain επιτρέπει στο μοντέλο να εντοπίζει μοτίβα που δεν είναι εμφανή μόνο από το κείμενο του τίτλου.

Η προσεκτική καθαριότητα των δεδομένων είναι κρίσιμη, καθώς τα θορυβώδη δεδομένα μπορούν να μειώσουν την ακρίβεια και να αυξήσουν την πιθανότητα overfitting.

2. Δημιουργία Χαρακτηριστικών με TF-IDF Vectorizer
Για να κατανοήσει το μοντέλο τη σημασία των λέξεων στο κείμενο, χρησιμοποιήθηκε ο TF-IDF Vectorizer. Αυτός ο μετασχηματισμός αποτυπώνει τη σχετική συχνότητα μιας λέξης στο σύνολο δεδομένων και την συσχέτισή της με τον τίτλο.

- **Γιατί TF-IDF:** Η συχνότητα των λέξεων από μόνη της δεν είναι πάντα ενδεικτική της σημασίας. Για παράδειγμα, λέξεις που εμφανίζονται παντού μπορεί να μην έχουν ουσιαστική πληροφορία. Ο TF-IDF εξισορροπεί αυτή τη συχνότητα με το πόσο μοναδική είναι μια λέξη στο σύνολο δεδομένων, παρέχοντας καλύτερη πληροφόρηση για την κατάταξη.

3. Εξαγωγή και Κανονικοποίηση Numeric Χαρακτηριστικών
Εκτός από το κείμενο, προσθέσαμε χαρακτηριστικά που βασίζονται σε numeric τιμές, όπως το μήκος του URL. Αυτά τα χαρακτηριστικά κανονικοποιήθηκαν με τη χρήση του StandardScaler, ώστε να έχουν συγκρίσιμη κλίμακα και να μην επηρεάζουν δυσανάλογα την εκπαίδευση του μοντέλου.

- **Γιατί να κανονικοποιήσουμε:** Οι διαφορές στην κλίμακα των numeric χαρακτηριστικών (π.χ., το μήκος του URL μπορεί να είναι από λίγους χαρακτήρες έως εκατοντάδες) θα μπορούσαν να κάνουν το μοντέλο να δίνει υπερβολική έμφαση στα χαρακτηριστικά με τις μεγαλύτερες τιμές. Η κανονικοποίηση εξασφαλίζει ότι κάθε χαρακτηριστικό συμβάλλει ισοδύναμα.

4. Επιλογή Μοντέλων Μηχανικής Μάθησης
Στην επόμενη φάση εξετάστηκαν διάφοροι αλγόριθμοι μηχανικής μάθησης, ο καθένας για διαφορετικό λόγο:

- **Support Vector Machine (SVM):**
  Επελέγη επειδή είναι αποτελεσματικό σε υψηλής διάστασης δεδομένα. Τα TF-IDF χαρακτηριστικά δημιουργούν μεγάλο αριθμό διαστάσεων, και το SVM με γραμμικό πυρήνα είναι ιδιαίτερα καλό στο να βρει την καλύτερη υπέρ-επιφάνεια που διαχωρίζει τις κατηγορίες.
- **Random Forest:**
  Χρησιμοποιήθηκε λόγω της δυνατότητάς του να χειρίζεται μη γραμμικές σχέσεις. Το Random Forest, μέσω της χρήσης πολλών δέντρων αποφάσεων, βοηθά στη σταθεροποίηση των προβλέψεων και στην ανάδειξη των σημαντικών χαρακτηριστικών. Παρέχει επίσης φυσικό τρόπο να εξηγήσουμε ποια χαρακτηριστικά συνεισφέρουν περισσότερο στην τελική πρόβλεψη.
- **XGBoost:**
  Προτιμήθηκε για την ταχύτητα εκπαίδευσής του και την ικανότητά του να μαθαίνει σύνθετα μοτίβα από δεδομένα. Το XGBoost είναι εξαιρετικά αποδοτικό και συχνά υπερέχει σε προβλήματα ταξινόμησης όπου οι σχέσεις είναι πιο πολύπλοκες.
- **Multi-Layer Perceptron (MLP):**
  Εισήχθη για την ικανότητά του να συλλαμβάνει μη γραμμικές σχέσεις μέσω των κρυφών του επιπέδων. Τα νευρωνικά δίκτυα όπως το MLP μπορούν να μάθουν από τα χαρακτηριστικά που δημιουργήθηκαν και να προσαρμοστούν σε δεδομένα με πολλές διαστάσεις, προσφέροντας μεγαλύτερη ευελιξία.

5. Βελτιστοποίηση Υπερπαραμέτρων και Cross-Validation
Κάθε μοντέλο βελτιστοποιήθηκε μέσω Grid Search, και η απόδοσή του αξιολογήθηκε με Stratified K-Fold Cross-Validation. Αυτός ο συνδυασμός διασφαλίζει ότι τα αποτελέσματα είναι αξιόπιστα, καθώς προέρχονται από πολλαπλά splits των δεδομένων, ενώ οι υπερπαράμετροι επιλέγονται βάσει της βέλτιστης απόδοσης σε metrics όπως το F1-Score και το ROC-AUC.

4. Πειραματική Μελέτη

Στόχος της πειραματικής μελέτης ήταν η διερεύνηση της απόδοσης τεσσάρων διαφορετικών αλγορίθμων—SVM, Random Forest, XGBoost και Multi-Layer Perceptron (MLP)—στην ανίχνευση ψευδών ειδήσεων. Ακολουθώντας μια ενιαία διαδικασία προεπεξεργασίας και αξιολόγησης, καταγράψαμε την ακρίβεια, το

F1-Score, και την ικανότητα διάκρισης των μοντέλων σε πραγματικές και ψευδείς ειδήσεις. Παρακάτω παρατίθενται τα αποτελέσματα και η ερμηνεία τους.

1. Multi-Layer Perceptron (MLP)
Το MLP διακρίθηκε ως το καλύτερο μοντέλο με συνολική ακρίβεια **84.2%** και ισορροπημένα F1-Scores μεταξύ των κατηγοριών.

- **Ψευδείς ειδήσεις:**
  Είχε precision **70.9%** και recall **61.9%**, πράγμα που σημαίνει ότι αναγνώριζε αρκετές ψευδείς ειδήσεις με σχετική ακρίβεια, αλλά υπήρχε χώρος για βελτίωση στον εντοπισμό όλων των ψευδών δειγμάτων.
- **Πραγματικές ειδήσεις:**
  Εξαιρετική απόδοση με precision **87.9%** και recall **91.6%**, καταδεικνύοντας ότι σχεδόν όλες οι πραγματικές ειδήσεις ταυτοποιήθηκαν σωστά, προσφέροντας μεγάλη αξιοπιστία στον εντοπισμό αληθινών πληροφοριών.

Συμπέρασμα:
Το MLP αναδεικνύεται ως το πλέον αποδοτικό μοντέλο για τον εντοπισμό ψευδών ειδήσεων. Είναι ικανό να διατηρεί μια σταθερή ισορροπία μεταξύ ακρίβειας και ανάκλησης, καθιστώντας το ιδανικό εργαλείο για προβλήματα που απαιτούν συνεπή ανίχνευση και στις δύο κατηγορίες.

2. Random Forest
Το Random Forest σημείωσε συνολική ακρίβεια **84.6%**, αποδεικνύοντας ότι μπορεί να ανταγωνιστεί το MLP σε πολλούς τομείς.

- **Ψευδείς ειδήσεις:**
  Με precision **77.9%** και recall **53.2%**, το μοντέλο ήταν αρκετά ακριβές στον εντοπισμό ψευδών ειδήσεων, αν και δεν τις εντόπιζε όλες.
- **Πραγματικές ειδήσεις:**
  Το μοντέλο ξεχώρισε εδώ, με precision **86.0%** και recall **95.0%**, εντοπίζοντας σχεδόν όλες τις πραγματικές ειδήσεις με υψηλή ακρίβεια.

Συμπέρασμα:
Το Random Forest είναι μια εξαιρετική επιλογή για περιπτώσεις όπου η προτεραιότητα είναι η σωστή αναγνώριση πραγματικών ειδήσεων. Η δυνατότητά του να χειρίζεται μη γραμμικές σχέσεις και να αποφεύγει το υπερεκπαίδευση το καθιστούν αξιόπιστη λύση.

3. XGBoost
Με ακρίβεια **84.2%**, το XGBoost απέδειξε τη σταθερότητα και την ισχύ του.

- **Ψευδείς ειδήσεις:**
  Το precision **76.9%** και recall **51.9%** δείχνουν ότι, όπως και το Random Forest, είναι καλό στο να αναγνωρίζει τις ψευδείς ειδήσεις που εντοπίζει, αλλά δεν καταφέρνει να καλύψει το σύνολο αυτών.
- **Πραγματικές ειδήσεις:**
  Με precision **85.7%** και recall **94.9%**, προσφέρει εξαιρετική απόδοση στον εντοπισμό πραγματικών ειδήσεων, παρέχοντας μια σταθερή και αξιόπιστη επιλογή για την ανίχνευση αληθινών ειδήσεων.

Συμπέρασμα:
Το XGBoost είναι μια ισχυρή επιλογή για περιπτώσεις όπου απαιτείται σταθερότητα και αξιοπιστία, ειδικά στην αναγνώριση πραγματικών ειδήσεων. Εξίσου κατάλληλο για προβλήματα με πολύπλοκες σχέσεις στα δεδομένα, προσφέρει μια καλά στρογγυλεμένη απόδοση.

4. Support Vector Machine (SVM)
Με συνολική ακρίβεια **82.1%**, το SVM είχε τη χαμηλότερη επίδοση σε σχέση με τα άλλα μοντέλα, αλλά παρέμενε αρκετά ανταγωνιστικό.

- **Ψευδείς ειδήσεις:**
  Το precision **61.7%** και recall **73.1%** δείχνουν ότι το SVM είχε την ικανότητα να ανιχνεύει περισσότερες ψευδείς ειδήσεις από άλλα μοντέλα (υψηλότερη recall), αν και το ποσοστό των λανθασμένων προβλέψεων παρέμεινε υψηλό (χαμηλότερο precision).
- **Πραγματικές ειδήσεις:**
  Με precision **90.5%** και recall **85.1%**, το SVM ήταν αρκετά αξιόπιστο στον εντοπισμό πραγματικών ειδήσεων, παρότι δεν έφτασε τα επίπεδα του MLP και του Random Forest.

Συμπέρασμα:
Το SVM αποδεικνύεται ισχυρό στη διαχείριση κειμένων υψηλής διάστασης, αλλά υστερεί σε συνολική απόδοση έναντι των άλλων μοντέλων. Ωστόσο, παραμένει χρήσιμο εργαλείο, ειδικά αν η προτεραιότητα είναι ο εντοπισμός περισσότερων ψευδών ειδήσεων.

Γενική Ερμηνεία:
Η πειραματική μελέτη ανέδειξε το MLP ως το πιο αποδοτικό μοντέλο για την ανίχνευση ψευδών ειδήσεων, λόγω της ισορροπίας μεταξύ ακρίβειας και πλήρους κάλυψης. Το Random Forest και το XGBoost έδειξαν επίσης πολύ καλή απόδοση, καθιστώντας τα εξαιρετικές επιλογές για τον εντοπισμό πραγματικών ειδήσεων. Το SVM, αν και υπολειπόμενο σε ακρίβεια, παραμένει χρήσιμο όταν η ανάκληση για τις ψευδείς ειδήσεις είναι προτεραιότητα.

5. Συμπεράσματα

Στο πλαίσιο αυτού του έργου, εξετάσαμε τέσσερις διαφορετικές προσεγγίσεις μηχανικής μάθησης—SVM, MLP, Random Forest και XGBoost—για την ανίχνευση ψευδών ειδήσεων. Η ανάλυση των αποτελεσμάτων μας επέτρεψε να εξάγουμε ορισμένα σημαντικά συμπεράσματα σχετικά με την αποτελεσματικότητα κάθε μεθόδου και την καταλληλότητά τους για διαφορετικά είδη δεδομένων.

1. Multi-Layer Perceptron (MLP): Το πιο αποδοτικό μοντέλο
Το MLP κατέγραψε την καλύτερη συνολική απόδοση, συνδυάζοντας υψηλό F1-Score και ακρίβεια. Ήταν ιδιαίτερα ικανό στο να αναγνωρίζει σωστά τόσο τις ψεύτικες όσο και τις πραγματικές ειδήσεις, διατηρώντας μια ισορροπία που καθιστά το μοντέλο ιδανικό για εφαρμογές όπου απαιτείται αξιοπιστία σε όλες τις κατηγορίες.

2. Random Forest και XGBoost: Εξαιρετικές εναλλακτικές λύσεις
Αμφότερα τα Random Forest και XGBoost προσέφεραν εξαιρετικά αποτελέσματα, κοντά στην απόδοση του MLP. Το Random Forest απέδειξε τη δύναμή του στη διαχείριση μη γραμμικών σχέσεων και τη σταθερότητά του, ενώ το XGBoost διέπρεψε με την ταχύτητα και την ικανότητά του να εκμεταλλεύεται σύνθετα μοτίβα στα δεδομένα. Αυτά τα μοντέλα είναι ιδανικά όταν οι ανάγκες απαιτούν υψηλή ακρίβεια και γρήγορη εκπαίδευση.

3. Support Vector Machine (SVM): Ανταγωνιστικό αλλά περιορισμένο
Το SVM παρότι υστέρησε σε σχέση με τα άλλα μοντέλα, κατέγραψε αξιοσημείωτη απόδοση στην ανάκληση για τις ψευδείς ειδήσεις. Αν και δεν είναι η πρώτη επιλογή για εφαρμογές υψηλής ακρίβειας, παραμένει ένα αξιόλογο

εργαλείο, ειδικά για δεδομένα υψηλής διάστασης και περιπτώσεις όπου η ανάκληση είναι κρίσιμη.

Γενικό Συμπέρασμα
Η ανίχνευση ψευδών ειδήσεων είναι ένα πολυδιάστατο πρόβλημα που απαιτεί προσεγμένες επιλογές μεθοδολογιών και τεχνικών προεπεξεργασίας. Μέσω αυτής της μελέτης, το MLP ξεχώρισε ως το πιο κατάλληλο μοντέλο για την εργασία μας. Ωστόσο, τόσο το Random Forest όσο και το XGBoost παρέχουν βιώσιμες και αποδοτικές εναλλακτικές λύσεις, ενώ το SVM παραμένει χρήσιμο εργαλείο για συγκεκριμένες ανάγκες. Στο σύνολο, αυτή η πειραματική προσέγγιση προσφέρει ένα ισχυρό πλαίσιο για την ανάπτυξη μελλοντικών εφαρμογών ανίχνευσης ψευδών ειδήσεων, ενισχύοντας την αξιοπιστία και την ποιότητα της πληροφορίας που καταναλώνουμε καθημερινά.

6. Βιβλιογραφία

1. Shu, Kai, et al. "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media." arXiv preprint arXiv:1809.01286, 2018.
2. Shu, Kai, et al. "Fake News Detection on Social Media: A Data Mining Perspective." ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, 2017.
3. Breiman, Leo. "Random Forests." Machine Learning, vol. 45, no. 1, 2001.