ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Απαλλακτική Εργασία

1.ΣΚΟΠΟΣ

Σκοπός αυτής της εργασίας είναι η αντιμετώπιση του προβλήματος των fake news σε διάφορα άρθρα στο διαδίκτυο. Συγκεκριμένα, η εργασια θα ασχοληθεί με την ανάπτυξη και τον πειραματισμό διαφόρων Recurrent Neural Network (RNN) αρχιτεκτονικών που κατηγοριοποιούν διάφορα ειδησεογραφικά άρθρα ως fake news ή μη. Για την εκπαίδευση θα χρησιμοποιηθεί το ακόλουθο dataset από το kaggle (https://www.kaggle.com/competitions/fake-news/data)[1] το οποίο δημοσιεύθηκε το 2018 και περιέχει άρθρα που μπορούν να χρησιμοποιηθούν για εκπαίδευση νευρωνικού δικτύου εντοπισμού fake news. Το dataset περιλαμβάνει το περιεχόμενο των άρθρων και ένα βοηθητικό label που τα κατηγοριοποιεί ως "πιθανώς αναξιόπιστα".
Η αξιολόγηση των αποτελεσμάτων θα γίνει με μετρικές όπως accuracy, recall και F1 score για να διαπιστωθεί η απόδοση του μοντέλου σε σύγκριση με παρόμοιες μελέτες. Ο επιθυμητός στόχος είναι μια απόδοση ισάξια ή καλύτερη των προηγούμενων ερευνών με βάση το τρέχον dataset.

2.ΣΧΕΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Τα τελευταία χρόνια, ο εντοπισμός fake news έχει γίνει κρίσιμος τομέας έρευνας λόγω της ταχείας διάδοσης παραπληροφόρησης. Διάφορα μοντέλα βαθιάς μάθησης έχουν προταθεί για την αντιμετώπιση αυτού του ζητήματος, αξιοποιώντας τις δυνατότητες διαφορετικών αρχιτεκτονικών νευρωνικών δικτύων.

Υβριδικά μοντέλα CNN-RNN

Μια σημαντική προσέγγιση στον εντοπισμό fake news είναι η χρήση Υβριδικών Συνελικτικών Νευρωνικών Δικτύων (CNN) και Επαναλαμβανόμενων Νευρωνικών Δικτύων (RNN). Η μελέτη των Nasir et al. (2021)[2] εισήγαγε ένα μοντέλο που συνδυάζει CNN με RNN για να συλλάβει αποτελεσματικά τόσο τα χωρικά όσο και τα χρονικά χαρακτηριστικά των δεδομένων κειμένου. Το στοιχείο CNN είναι υπεύθυνο για την εξαγωγή τοπικών χαρακτηριστικών από το κείμενο, τα οποία στη συνέχεια τροφοδοτούνται στο RNN, συγκεκριμένα σε ένα δίκτυο LSTM, για την καταγραφή διαδοχικών εξαρτήσεων και contextual information.

[1]Επίσης διαθέσιμο και εδώ https://www.dropbox.com/scl/fi/7gubsdfvlvtgswsmp2a7d/train.csv?rlkey=cof2cxeyek1zveki3nnkpygcp&dl=1

[2]Nasir, Jamal & Khan, Osama & Varlamis, Iraklis. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights.

Bidirectional LSTM and GloVe Embeddings
Μια άλλη σημαντική συνεισφορά είναι από ερευνητές που χρησιμοποίησαν δίκτυα Bidirectional LSTM (BiLSTM) σε συνδυασμό με ενσωματώσεις GloVe[3]. Αυτό το μοντέλο αξιοποιεί την ικανότητα των BiLSTM να επεξεργάζονται πληροφορίες τόσο προς τις εμπρός όσο και προς τα πίσω, συλλαμβάνοντας έτσι ένα πιο ολοκληρωμένο πλαίσιο του κειμένου. Οι ενσωματώσεις GloVe χρησιμοποιούνται για την αναπαράσταση λέξεων σε έναν χώρο υψηλών διαστάσεων όπου σημασιολογικά παρόμοιες λέξεις βρίσκονται πιο κοντά μεταξύ τους. Αυτή η μέθοδος ήταν ιδιαίτερα αποτελεσματική στη βελτίωση της ακρίβειας ταξινόμησης των ψευδών ειδήσεων εμπλουτίζοντας τις αναπαραστάσεις του κειμένου εισόδου πριν δοθούν στο BiLSTM.

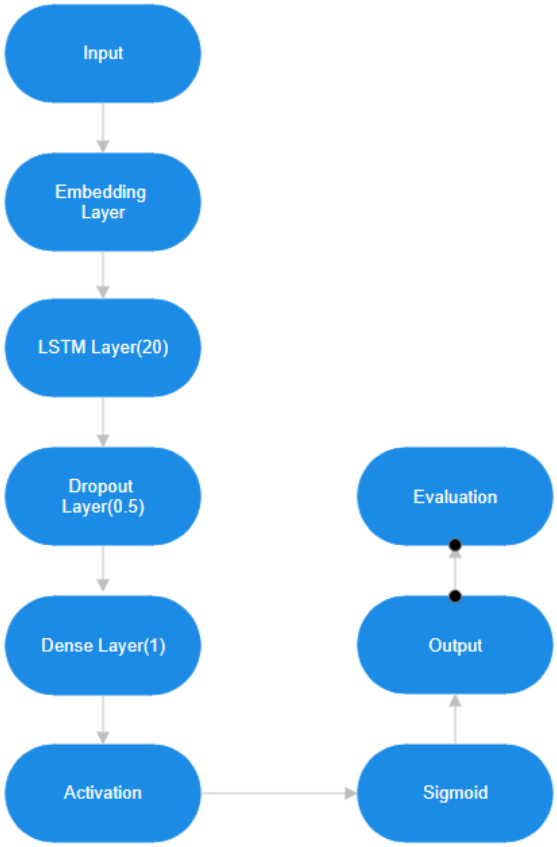
CNN + GRU
Η έρευνα των A. J. Keya et al.(2021)[4] εισάγει προηγμένα μοντέλα νευρωνικών δικτύων για την ανίχνευση fake news, χρησιμοποιώντας CNN και Gated Recurrent Unit (GRU). Η χρήση του CNN επιτρέπει την αποτελεσματική εξαγωγή χαρακτηριστικών από περιεχόμενο ειδήσεων και τίτλους, ενώ η GRU αντιμετωπίζει τις προκλήσεις του χειρισμού μακροχρόνιων διαδοχικών δεδομένων κειμένου με την επαναλαμβανόμενη αρχιτεκτονική της. Ο συνδυασμός αυτών των δύο μοντέλων νευρωνικών δικτύων, μαζί με προεκπαιδευμένες ενσωματώσεις λέξεων GloVe, ενισχύει την ακρίβεια του εντοπισμού.

3.ΥΛΟΠΟΙΗΣΗ

Το πρόβλημα αφορά την ανάπτυξη κατάλληλων classification μοντέλων τα οποία δέχονται άρθρα και εντοπίζουν μοτίβα fake news ώστε να καθορίσουν στο τέλος αν είναι πράγματι fake news ή αν περιέχουν αληθείς ειδήσεις. Για την υλοποίηση αποφασίσαμε να αξιοποιήσουμε 3 μοντέλα RNN ενώ για το embedding των λέξεων χρησιμοποιείται το Word2Vec. Επιπλέον, κάθε μοντέλο εκπαιδεύεται σε 20 epochs με batch size = 200.

α) Μοντέλο LSTM
Τα μοντέλα LSTM είναι μία εξελιγμένη μορφή μοντέλων RNN που επιτρέπουν την αποθήκευση και διατήρηση πληροφοριών για μεγαλύτερο χρονικό διάστημα.

[3]Abualigah, Laith & Al-Ajlouni, Yazan & Daoud, Mohammad & Altalhi, Maryam & Migdady, Hazem. (2024). Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe. Social Network Analysis and Mining.
[4]A. J. Keya, S. Afridi, A. S. Maria, S. S. Pinki, J. Ghosh and M. F. Mridha, "Fake News Detection Based on Deep Learning," *2021 International Conference on Science & Contemporary Technologies (ICSCT)*, Dhaka, Bangladesh, 2021



Εικόνα 1: Η αρχιτεκτονική του LSTM

Στην υλοποίηση του LSTM μοντέλου χρησιμοποιήθηκαν 20 νευρώνες, υπάρχει ένα dropout layer με τιμή 0.5, ένα πλήρως συνδεδεμένο layer με έναν (1) νευρώνα ενώ η συνάρτηση ενεργοποίησης είναι η σιγμοειδής.

β) Μοντέλο Bidirectional LSTM

Τα BiLSTM (Bidirectional LSTM) είναι μια επέκταση των LSTM που επιτρέπει την επεξεργασία δεδομένων και προς τις δύο κατευθύνσεις (προς τα εμπρός και προς τα πίσω). Αυτό επιτρέπει στο μοντέλο να χρησιμοποιεί πληροφορίες από το παρελθόν και το μέλλον, βελτιώνοντας την απόδοση σε εργασίες που απαιτούν περιβάλλον και από τις δύο κατευθύνσεις.

Στην υλοποίηση του BiLSTM μοντέλου χρησιμοποιήθηκαν επίσης 20 νευρώνες, υπάρχει ένα dropout layer με τιμή 0.5, ένα πλήρως συνδεδεμένο layer με έναν (1) νευρώνα ενώ η συνάρτηση ενεργοποίησης είναι η σιγμοειδής. Το flowchart είναι το ίδιο με την *Εικόνα 1* απλώς ο κόμβος μοντέλου είναι διαφορετικός αφού χρησιμοποιήθηκε BiLSTM.

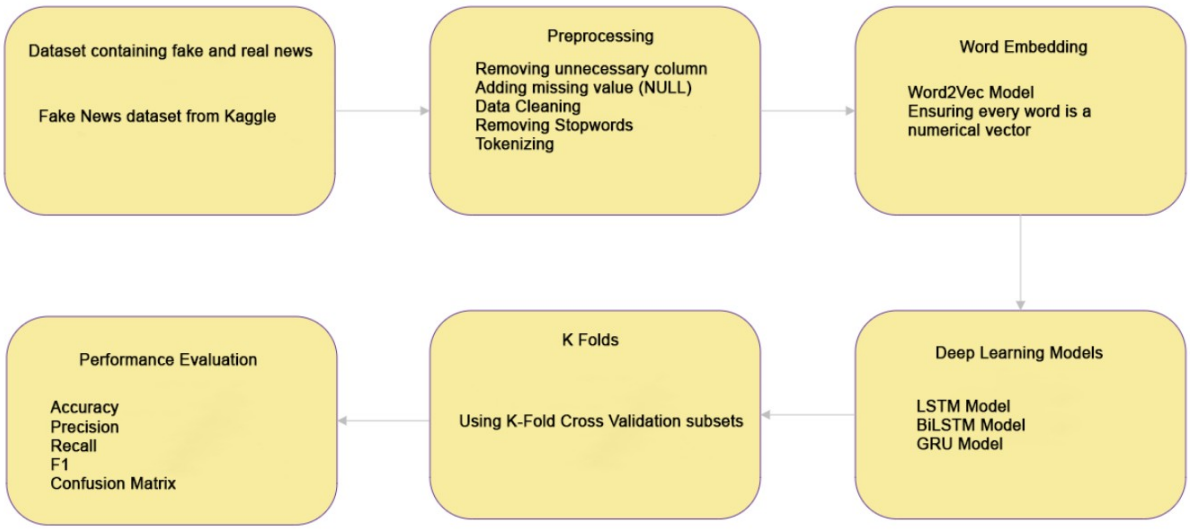
γ) Μοντέλο GRU

Τα GRU (Gated Recurrent Unit) είναι μια απλοποιημένη μορφή των LSTM, με λιγότερες πύλες και παραμέτρους. Αυτό τα καθιστά ταχύτερα στην εκπαίδευση και πιο αποδοτικά από πλευράς υπολογιστικής ισχύος, ενώ συχνά παρέχουν παρόμοια απόδοση.

Στην υλοποίηση του GRU μοντέλου χρησιμοποιήθηκαν επίσης 20 νευρώνες, υπάρχει ένα dropout layer με τιμή 0.5, ένα πλήρως συνδεδεμένο layer με έναν (1) νευρώνα ενώ η συνάρτηση ενεργοποίησης είναι και πάλι η σιγμοειδής. Το flowchart είναι το ίδιο με την *Εικόνα 1* απλώς ο κόμβος μοντέλου είναι διαφορετικός αφού χρησιμοποιήθηκε GRU.

4.ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

Η διαδικασία που ακολουθήθηκε είναι η εξής:



1) Επιλέχθηκε το Dataset “Fake News” από το Kaggle.

2) Τα δεδομένα του dataset πρέπει να επεξεργαστούν προτού δοθούν στα μοντέλα για εκπαίδευση. Σε αυτό το σημείο αφαιρούνται τα δεδομένα που δεν χρειάζονται, δηλαδή οι στήλες “id” και “author”. Στη συνέχεια προστίθεται η λέξη “None” σε όποια μεταβλητή είναι κενή. Μετά αρχίζει η επεξεργασία των κειμένων, αφαιρούνται τα URLs, οι ειδικοί χαρακτήρες, οι μη-αλφαβητικοί χαρακτήρες, τα παραπάνω κενά και μετατρέπεται όλο το κείμενο σε πεζά γράμματα. Αφαιρούνται τα stopwords και γίνεται tokenization του κειμένου.

3) Το μοντέλο σαφώς δεν μπορεί να διαβάσει χαρακτήρες. Σε αυτό το σημείο γίνεται χρήση του Word2Vec ώστε το κείμενο να μετατραπεί σε διανύσματα που μπορούν να δοθούν για εκπαίδευση στο μοντέλο.

4) Δημιουργούνται τα νευρωνικά δίκτυα, έχουμε 3 μοντέλα, ένα LSTM, ένα BiLSTM κι ένα GRU. Αρχικοποιούνται τα Weights και τα Biases και ορίζονται τα layers και η συνάρτηση ενεργοποίησης.

5) Ορίζεται η μέθοδος εκπαίδευσης και ρυθμίζονται οι εποχές και ο ρυθμός εκπαίδευσης των μοντέλων. Σε αυτό το σημείο επίσης γίνεται χρήση της μεθόδου K-Fold Cross Validation. Τα δεδομένα χωρίζονται σε splits και το μοντέλο εκπαιδεύεται μία φορά σε κάθε split. Στο τέλος υπολογίζεται το μέσο σφάλμα απόδοσης βάσει όλων των επαναλήψεων. Επιλέξαμε να χωρίσουμε τα δεδομένα σε 3 splits.

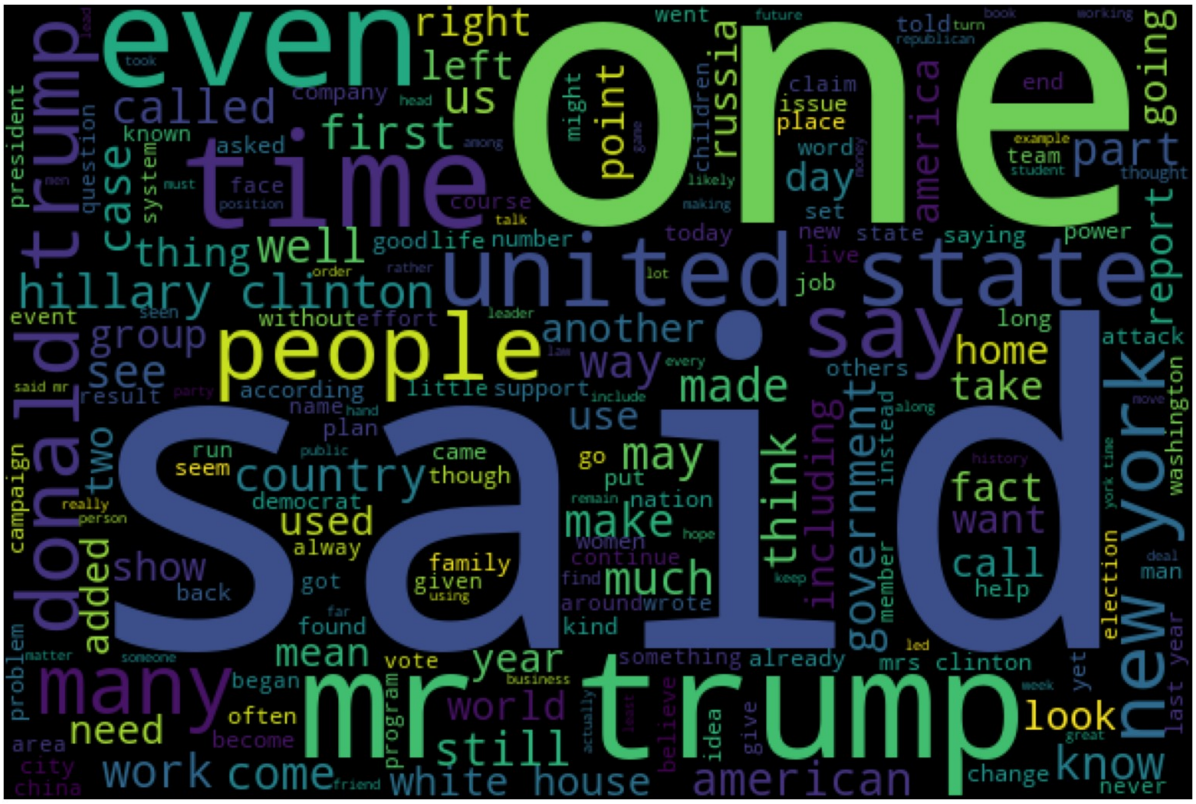
6) Η αξιολόγηση των αποτελεσμάτων γίνεται βάσει των παρακάτω μετρικών:

- **Accuracy:** Το ποσοστό των σωστών προβλέψεων από το σύνολο των προβλέψεων.
- **Precision:** Το ποσοστό των θετικών προβλέψεων που είναι πραγματικά θετικές.
- **Recall:** Το ποσοστό των πραγματικών θετικών που αναγνωρίζονται σωστά από το μοντέλο.
- **F1 Score:** Ο σταθμισμένος μέσος όρος accuracy και του recall.
- **Confusion Matrix:** Ένας πίνακας που δείχνει τον αριθμό των σωστών και λανθασμένων προβλέψεων ανά κατηγορία.

Dataset και Προεπεξεργασία

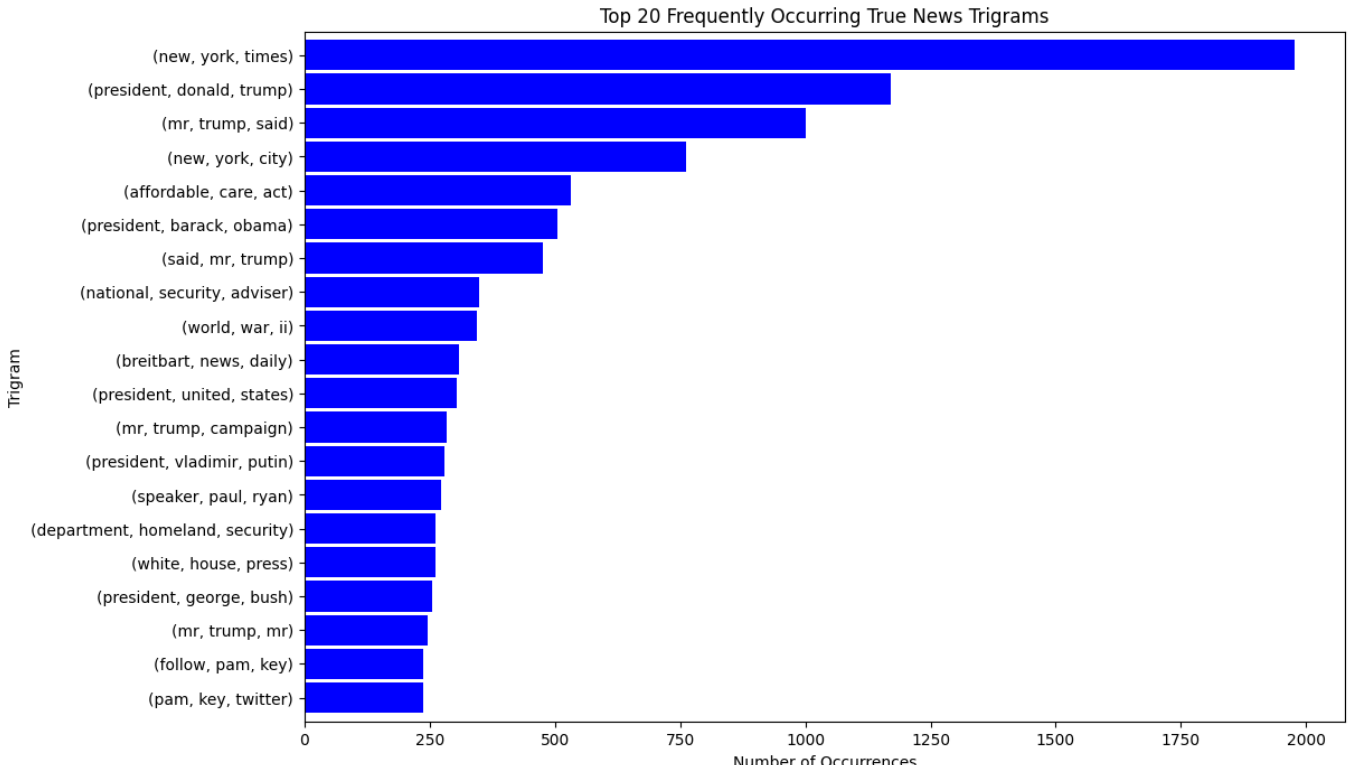
Το dataset που επιλέχθηκε προέρχεται από το kaggle και περιλαμβάνει άρθρα, τους τίτλους τους, το όνομα του συγγραφέα και το class label για την κατηγοριοποίηση ως αξιόπιστο ή μη. Η τιμή 0 συμβολίζει ένα αξιόπιστο άρθρο και η τιμή 1 μη αξιόπιστο αντίστοιχα.
Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκε μόνο το περιεχόμενο των άρθρων και το class label.

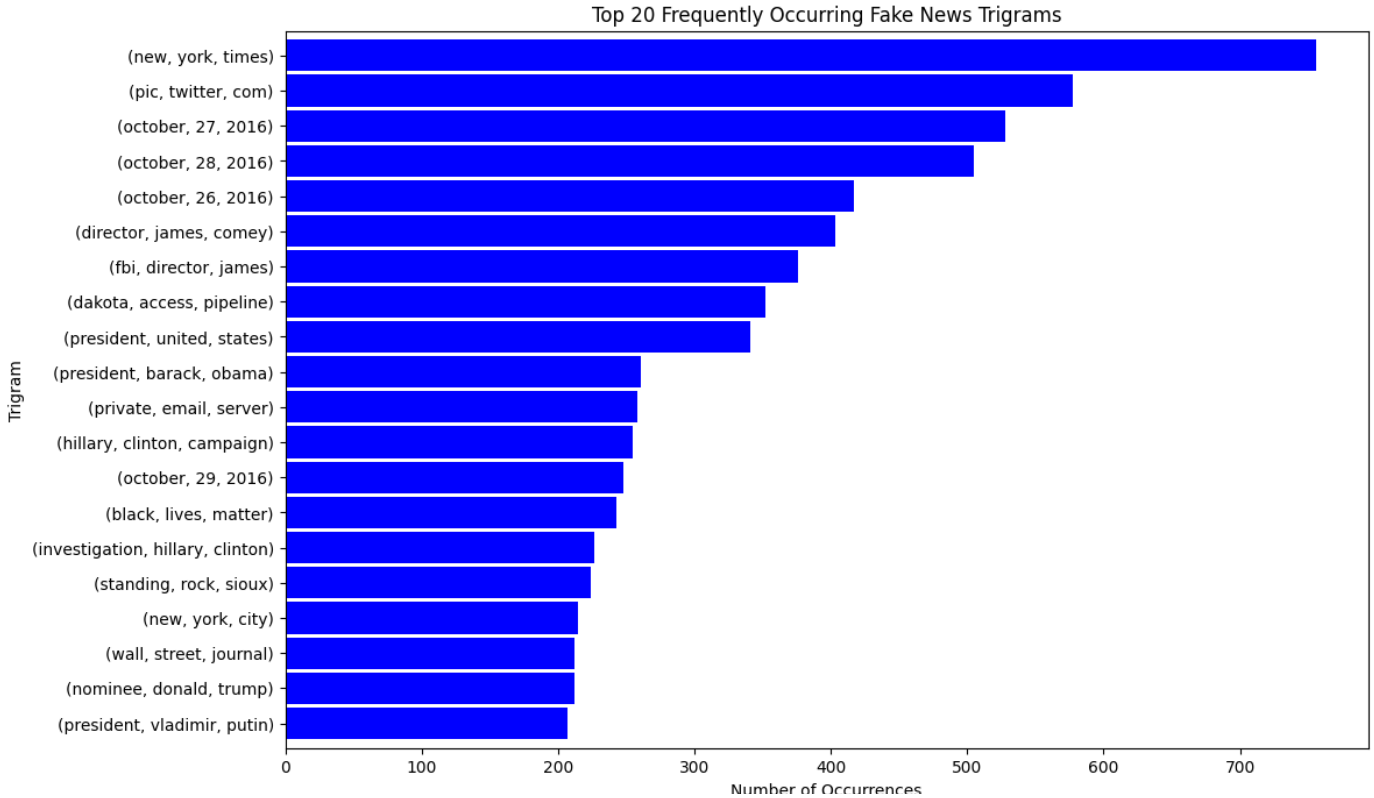
Συνολικά συμπεριλαμβάνονται 20800 άρθρα εκ των οποίων 10387 ανήκουν στην κατηγορία 0 και 10413 στην κατηγορία 1. Στην *Εικόνα 3* φαίνεται ένα "συννεφόλεξο" με τις πιο συχνές λέξεις από ολόκληρο το dataset



Εικόνα 3: Wordcloud με τις λέξεις από όλες τις κατηγορίες

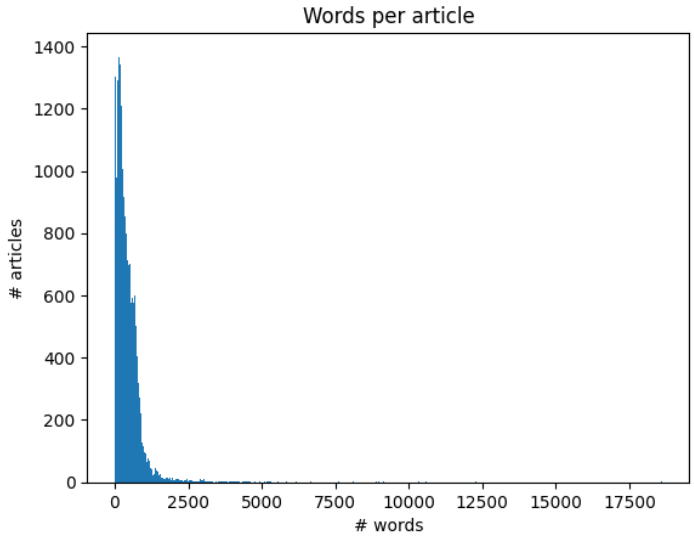
Περισσότερες πληροφορίες για τον διαχωρισμό των fake και μη άρθρων παίρνουμε από τα Trigrams, δηλαδή τις πιο συχνές τριάδες λέξεων που εμφανίζονται μέσα στο περιεχόμενο.





Αξίζει να σημειωθεί πως τα αποτελέσματα των παραπάνω εικόνων προήλθαν μετά τον "καθαρισμό" του αρχικού κειμένου, δηλαδή την αφαίρεση των stopwords και άλλων μη χρήσιμων χαρακτήρων. Στη συνέχεια, οι λέξεις δόθηκαν σε ένα μοντέλο Word2Vec για την μετατροπή τους σε διανύσματα με διάσταση 100. Επιπλέον, ακολούθησε το tokenization της κάθε λέξης και τέλος το padding της κάθε ακολουθίας token έτσι ώστε να είναι όλες στην ίδια διάσταση.

Η διάσταση της κάθε ακολουθίας είναι 1000 και αυτό προκύπτει από το παρακάτω ιστόγραμμα

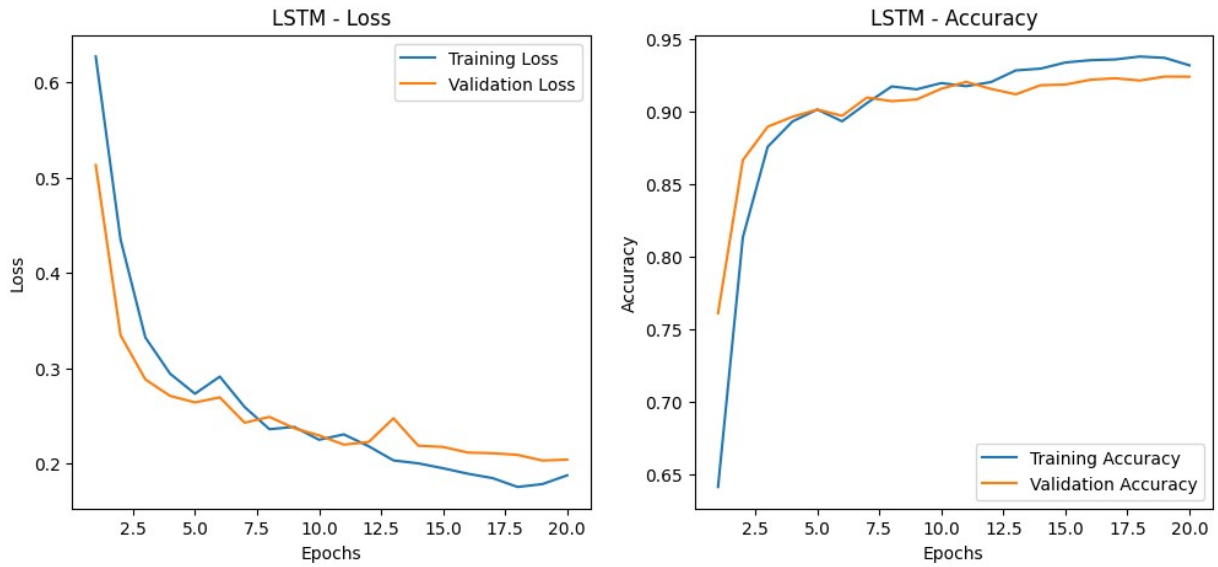


Εικόνα 6 : Λέξεις ανά άρθρο.

Εφόσον η πλειοψηφία των άρθρων (και συγκεκριμένα το 94%) περιλαμβάνουν το πολύ 1000 λέξεις δεν έχει νόημα μεγαλύτερο μέγεθος ακολουθίας. Έτσι, τα άρθρα με λιγότερες από 1000 λέξεις συμπληρώνονται με μηδενικά και στα υπόλοιπα αφαιρούνται οι παραπάνω λέξεις.

Τέλος, τα αποτελέσματα του μοντέλου Word2Vec χρησιμοποιούνται για την αρχικοποίηση των βαρών και ορίζονται οι κατάλληλες διαστάσεις για το embedding layer.

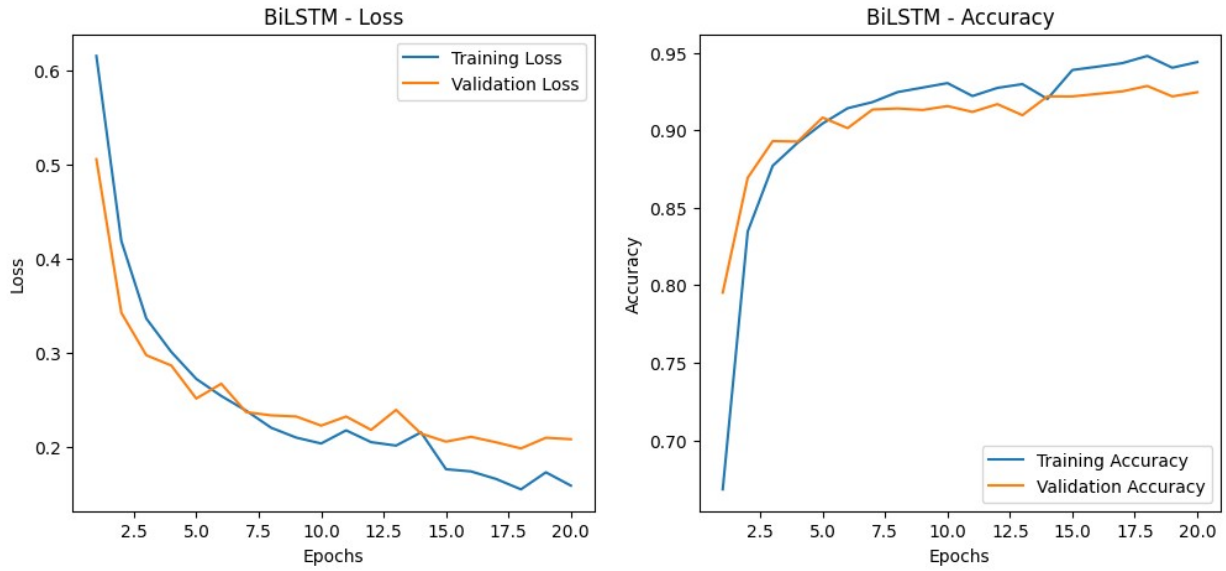
Αποτελέσματα



Εικόνα 7 : LSTM training και validation

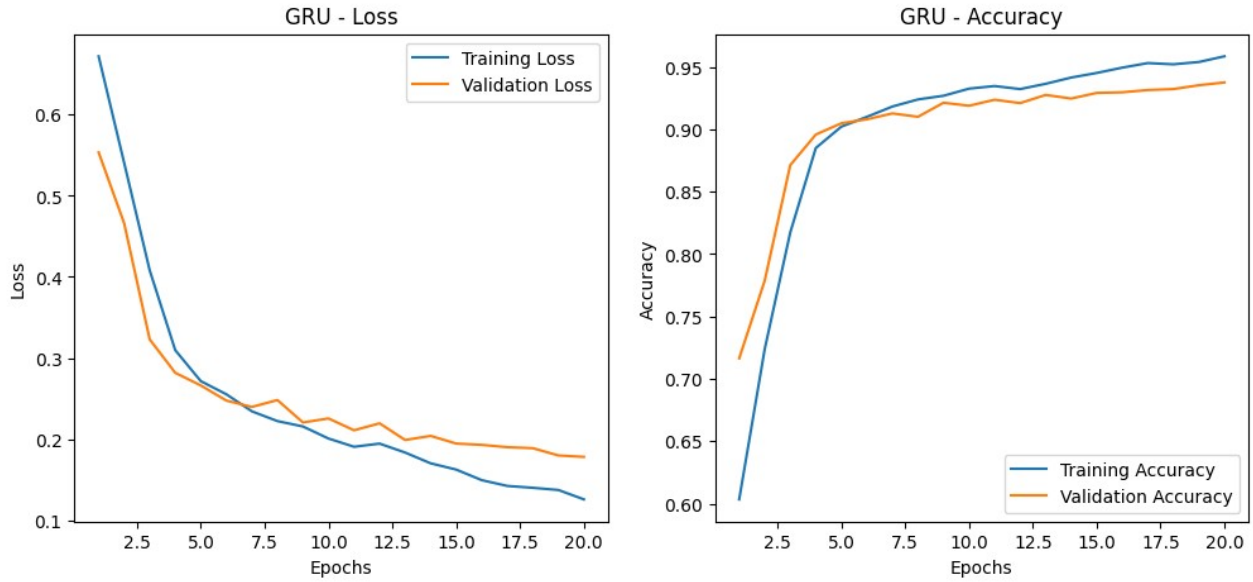
Παραπάνω βλέπουμε το Loss curve και το Accuracy curve για το μοντέλο LSTM, εκπαιδευμένο για 20 εποχές. Φαίνεται να εκπαιδεύεται σχετικά καλά στο dataset,

ξεφεύγοντας ελάχιστα σε overfit μετά τις 12 εποχές για να συγκλίνει ξανα στις 19-20. Πιθανώς με περισσότερες εποχές να ξαναυπήρχε τομή με τη γραμμή validation loss.



Εικόνα 8 : BiLSTM training και validation

Εδώ βλέπουμε τα αντίστοιχα Loss curve και Accuracy curve για το μοντέλο BiLSTM, εκπαιδευμένο για 20 εποχές. Το μοντέλο φαίνεται να γενικεύει καλώς στα validation data παρόλο που υπάρχουν αυξομειώσεις στο loss . Στις 15 εποχές το μοντέλο αρχίζει να κάνει overfit σε μερικό αλλά όχι ανησυχητικό βαθμό. Το training accuracy αυξάνεται με μερικές αυξομειώσεις και σταθεροποιείται στο 95%. Επομένως το μοντέλο αποδεικνύει πως είναι καλό για το τρέχον πρόβλημα με μερικά περιθώρια βελτίωσης.



Εικόνα 9 : GRU training και validation

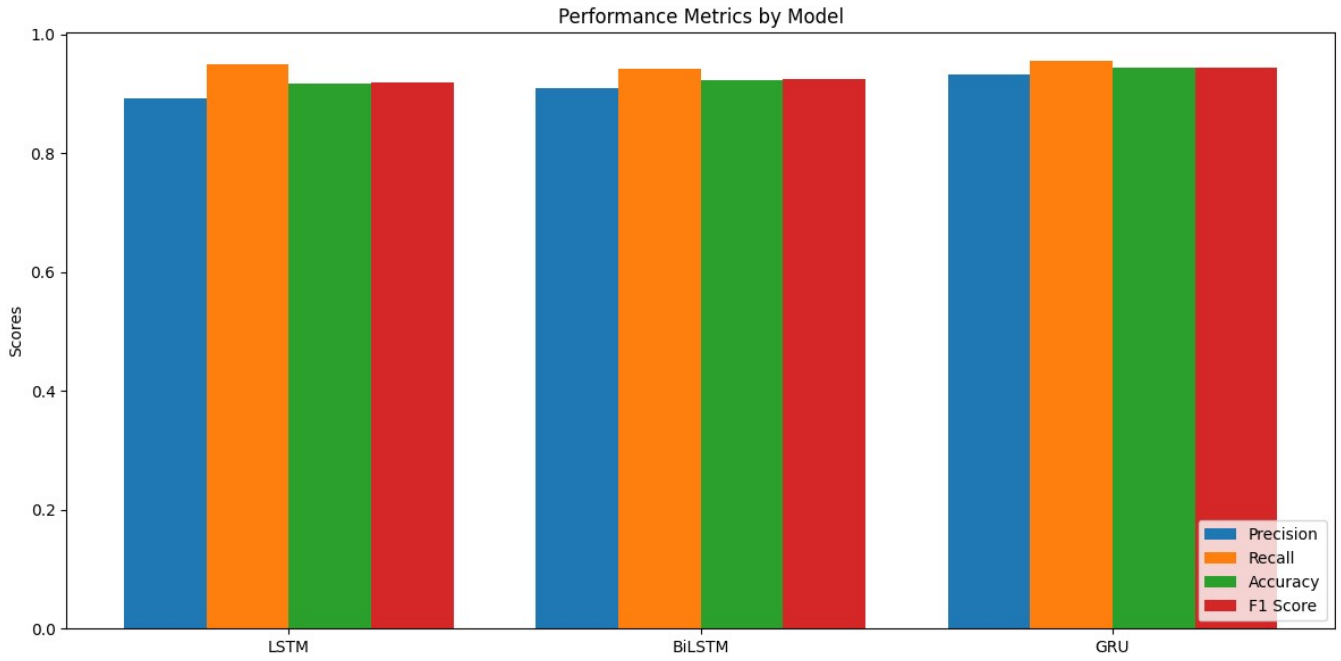
Όσον αφορά το μοντέλο GRU φαίνεται πως υπάρχει συνεχής μείωση στο training loss σε όλη την διάρκεια της εκπαίδευσης ενώ ταυτόχρονα αυξάνεται σταθερά το accuracy. Στις 7-8

εποχές αρχίζει να κάνει overfit το οποίο μεγαλώνει όλο και περισσότερο μέχρι το πέρας των 20 εποχών. Σε σύγκριση με τα προηγούμενα μοντέλα παρουσιάζει την πιο ομαλή εικόνα χωρίς αυξομειώσεις. Υπάρχει καλή γενίκευση στα validation data και το overfit που εμφανίζεται είναι σε αποδεκτά επίπεδα.

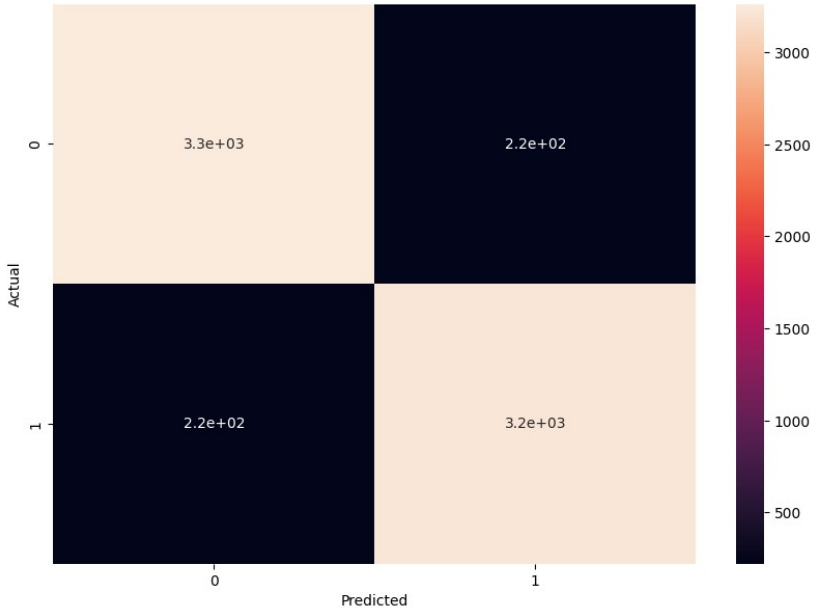
Παρακάτω παρουσιάζονται τα αποτελέσματα από το test dataset.

Πίνακας 1 : Οι αποδόσεις των μοντέλων

	Precision	**Recall**	**Accuracy**	**F1 Score**
LSTM	0.8914	0.9505	0.9161	0.9198
BiLSTM	0.9091	0.9418	0.9231	0.9248
GRU	0.9331	0.9548	0.9430	0.9438



Τα τρία μοντέλα έχουν καλή απόδοση, με πολύ μικρές διαφορές μεταξύ τους. Οι αιχμές που παρατηρήθηκαν στα αρχικά γραφήματα του training κια validation loss θα μπορούσαν να αποδοθούν σε διακυμάνσεις στη διαδικασία εκπαίδευσης, πιθανώς λόγω μικρού μεγέθους δεδομένων ή του batch size αλλά δεν φαίνεται να επηρεάζουν σημαντικά τη συνολική απόδοση, όπως υποδεικνύεται από τις τελικές μετρήσεις.



Εικόνα 11 : Ενδεικτικό Confusion matrix από το μοντέλο του GRU

Συμπεράσματα και Συγκρίσεις

Βάσει της σύγκρισης των γραφημάτων, το GRU εμφανίζει την καλύτερη συνολική απόδοση μεταξύ των τριών μοντέλων. Η καμπύλη του training και validation loss για το GRU είναι η πιο ομαλή και δείχνει σταθερή μείωση, ενώ η καμπύλη accuracy φτάνει πολύ κοντά στο 95% χωρίς σημαντικές διακυμάνσεις μεταξύ training και validation δεδομένων, υποδεικνύοντας καλή γενίκευση. Το BiLSTM δείχνει επίσης καλή απόδοση με υψηλό accuracy και σχετικά σταθερό loss, αν και παρουσιάζει ελαφρώς μεγαλύτερες διακυμάνσεις στο validation loss. Το LSTM, παρόλο που φτάνει σε υψηλά επίπεδα accuracy, έχει τις περισσότερες διακυμάνσεις στο validation loss. Συνολικά, και τα τρία μοντέλα αποδίδουν πολύ καλά, με το GRU να υπερέχει ελαφρώς σε σταθερότητα και γενική απόδοση.

Η απόδοση των μοντέλων φαίνεται να είναι σχεδόν παρόμοια έως και καλύτερη από τον ανταγωνισμό. Συγκριτικά, με την έρευνα του Yuvraj Singh[5] η απόδοση μας υπερτερεί αρκετά σε όλα τα μοντέλα εκτος του GRU. Επίσης αξιοσημείωτο είναι ότι αυτά τα αποτελέσματα προέκυψαν με μια πολύ απλούστερη αρχιτεκτονική από αυτήν που περιγράφουν οι έρευνες στο 2ο κεφάλαιο αλλα και γενικότερα στην βιβλιογραφία.

[5] Singh, Yuvraj. (2023). Fake News Detection Using LSTM in TensorFlow and Deep Learning. Journal of Applied Science and Education (JASE). 3. 1-14. 10.54060/jase.v3i2.35.

Models	Accuracy	Precisions	Recall	F1 Score	Support
LSTM	0.867349	0.87	0.87	0.87	4003
Bi-LSTM	0.882338	0.88	0.88	0.88	4003
CNN-BiLSTM	0.888083	0.89	0.89	0.89	4003

Εικόνα 12 : Τα αποτελέσματα του Yuvraj Singh. Πήγη: ResearchGate