National College of Ireland

Fake news detection using Deep Learning and Natural Language Processing

MSc Research Project
Data Analytics

Shaik Nasir Vali
Student ID: 21166421

School of Computing
National College of Ireland

Supervisor: Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shaik Nasir Vali		
Student ID:	21166421		
Programme:	Data Analytics	**Year:**	2022
Module:	MSc Research Project		
Supervisor:	Hicham Rifai		
Submission Due Date:	15 / 12 / 2022		
Project Title:	Fake news detection using Deep Learning and Natural Language Processing		
Page Count:	19		
Word Count:	4993		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shaik Nasir Vali

Date: 15 / 12 / 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Fake news detection using Deep Learning and Natural Language Processing

Shaik Nasir Vali
21166421

Abstract

Although "fake news" is a major issue today, it is not insurmountable. Multiple factors, including the proliferation of social media, may contribute to the rapid growth of disinformation campaigns. There is no way to prevent 'Fabrications' or 'Fraudulent Stories', however many different machine learning algorithms have been developed to determine if a story is true or not. Several researchers have tackled this problem using machine learning techniques like SVM classifier and KNNs. I plan on using the Decision Tree Classifier, Logistic Regression, and RNN-LSTM with GloVe and NLTK to sort out the real stories from the fake ones.

1 Introduction

1.1 Motivation and Project Background

The goal of many fake news stories is either to damage an individual's or company's reputation or to attract clickbait headlines. When the vast majority of people still relied on newspapers and ratio for their news, there was no need for the practice of altering the news. The development of sophisticated gadgets has made instantaneous access to news possible. It is undeniable that we modern social creatures spend a disproportionate amount of time glued to our smartphones, where we consume vast amounts of information through platforms like Instagram, Facebook (Tacchini et al., 2017), Twitter and Whatsapp without giving much thought to the reliability of the information we're presented with.
80% of online Americans exposed to misleading information during the recent coronavirus outbreak. The fake news detection framework compiles data from articles and evaluates its veracity. The usual method for doing so employs a transformer architecture: nevertheless, tests have revealed that the latter is less accurate. Fortunately, we have a multitude of ways to eradicate fake news. According to a study, It is possible to detect fake news using KNN, Computer Vision and SVM. (D'Ulizia et al., 2021)

This research will concentrate on comparing different models namely Logistic Regression, NLTK, Decision tree classifier and RNN-LSTM to predict the fake news. Accuracy of 80-90 was attained in predicting the fake news by few of the previous works.

1.2 Research Question

This research is based on following aspects

Qn How Deep Learning and Natural Language processing can be useful in fake news prediction?

In this research, dataset was taken from Kaggle which is a free website. The main focus is on RNN – LSTM with GloVe model when compared to other models. Similarities and differences with previous works will also be discussed thoroughly. Natural language toolkit will be used to produce word clouds which will give information of mostly used words in the news. So, this Study is aiming to build a model which can detect fake news effectively.

2 Related Work

This study presents a FDML model that increases subject categorization while enhancing the findings of short false news. The FDML model projected that news with particular topics would have a high likelihood of being classified as fake news (Zhou et al., 2020), and that some authors would post high probability fake news. FDML created a task gate as well as a dynamic method for balancing the importance of tasks in order to selectively integrate representations based on various tasks. To address the imbalance learning problem, an SDW technique is given in which the weight of each job altered during each function is dynamically adjusted.

Fake News Detectoion Model	Label	Precision	Recall	F1 Score	Accuracy	Macro-F1
FDML	Pants-fire	0.662	**0.554**	**0.604**	0.508	**0.516**
	False	**0.445**	**0.644**	**0.526**		
	Barely-True	0.540	0.383	**0.448**		
	Half-True	**0.500**	0.401	**0.511**		
	Mostly-True	**0.493**	**0.530**	**0.511**		
	True	0.567	**0.564**	**0.565**		

Table 1: Result of FDML model

The word vector representations utilized in the FNN and LSTM deep learning models differ significantly. The models were paired with a real-time data processing component that gathered auxiliary data from the newspaper item's content/title. These alternatives, as well as domain names, author details, and so on, are superimposed on the initial article prior to the word embedding step to provide a range of contexts for the information at hand. The GloVe vector model was unable to appropriately gather market information from secondary alternatives because they departed from 'natural human language' patterns. This investigation is advanced by the development of an ensemble model in which the initial item is classified using an LSTM and the auxiliary options are handled by a second model. This method has the potential to boost performance even further.

Metrics	FNN without mined features	FNN with mined features
Accuracy	0.8335	**0.8429**
Precision	0.8006	**0.8134**
Recall	0.8920	**0.8936**
F1 Score	0.8438	**0.8516**
Specificity	0.7739	**0.7914**
Negative Predictive Value	0.8757	**0.8796**
False Positive Rate	0.2261	**0.2086**
False Discovery Rate	0.1994	**0.1866**
False Negative Rate	0.1080	**0.1064**

Table 2: Result of FNN model

2.1 NLP and Deep learning studies

Gupta & Kaushal, 2015 go into detail on OSNs, which stand for Online Social Networks, and the problems they face as a result of spammers who propagate false information with malicious purpose. So just Twitter was used, and URLs, Spam terms, and a few other crucial factors were taken into account. It is not bad to get an accuracy of 87.9% after combining algorithms like Decision Tree, Naïve Bayes and Clustering to improve the model. However, since this project makes use of words, using NLTK would increase the accuracy even more.

The term "astroturfing" is typically used in political contexts. It comprises making a social media post purporting to come from "grass-roots". In this research, a support vector machine and a logistic regression classifier are used to differentiate between authentic grass-roots communications and bogus grass-roots messages, with both models achieving above – 0.5 baseline performance on the Binary Classification problem and an accuracy of 90 % (Miller, 2016). The idea behind the research is good, and if overfitting was completely taken out of the equation accuracy may be improved. Models of deep learning were researched, including CNN and R-CNN.

Models for automatically identifying bogus news are being developed with the use of AI and machine learning. In this study, researchers take a fresh approach by combining CNN with RNN to produce a hybrid technique that outperforms non-hybrid approaches. Both the ISO and FAKES datasets were used, both of which contained fabricated news stories and encouraging results were found (Nasir et al., 2021). Nonetheless, different hybrid models should have been examined given the availability of a number of hybrid approaches nowadays.

As part of the fight against disinformation, a convolutional neural network was constructed in this research to traverse a large number of hidden layers. The results of this network were compared to those of baseline models. Over 98.36% accuracy is achieved by the model (Kaliyar et al., 2020), making it state-of-the-art. Results were verified using a battery of

performance indicators. The correctness of this study may have been improved with the help of NLP.

Ruchansky et al., 2017 proposes a solution comprised of three parts: the article's text, the user's responses and the source where the users are supporting the article. Similarly, the CSI model is utilized. CSI stands for Capture, Score and Integrate. The Capture section of this research uses recurrent convolutional neural networks to implement a deep learning strategy. The correctness of the model has not been compromised in any way by this idea.

2.2 SVM, KNN & Naïve Bayes

The Naive Bayes Classifier is used in this paper to combat fake news. The dataset utilized contains new Facebook posts. The paper is about spam filtering and has achieved a modest accuracy of 74%. The results would have been better if the Natural language tool kit had been used instead of tokenization after receiving the data from the Facebook API, coupled with the Naïve Bayes classifier. Accuracy can be enhanced further by including more variables and performing better feature processing. (Granik & Mesyura, 2017)

This work presented a novel technique to text analysis using n-gram features, as well as 6 other machine learning models such as linear support vector machine, stochastic gradient descent, and so on. Following the experimental evaluation, the linear SVM model achieved the best accuracy, while the KNN model achieved the lowest. This is an interesting study with several models that were tried but only a handful were successful. To gain a second perspective on the data, at least one Deep Learning model may have been tested and trained. (Ahmed et al., 2017)s

2.3 Genuine Ideologies

The Microblog is clearly the new big thing, but not all of the information it contains is accurate. So, in this study, we'll look at how to use a hierarchical propagation model with a three-layer credibility network. Even though this proposed model improves accuracy by more than 6%, it is insufficient, and the essential concept is to analyze the news by considering events, sub-events, and messages rather than going through each word as NLTK. (Jin et al., 2014)

Recently, several methods for disseminating fake news on the internet have emerged, comparable to tabloidization – 'Click baiting', It is exactly what the name implies; when we click on it, we are taken to another web page with the major bogus information. As a result, this research investigates clickbait as a sort of deceit, surveying both text and other text click baits. The disadvantage of this paper is that it does not employ high-level strategies to

facilitate the process, and click baits are quite tough. As a result, machine learning models can produce the greatest results. (Chen et al., 2015)

People publish a lot of videos on social media platforms like Instagram and Facebook. In the case of catastrophic occurrences such as plane crashes and terrorist attacks, we must verify the authenticity, and time is of the essence, therefore the decision must be made quickly. As a result, this research introduces Media REVEALr, a scalable framework with clustering-like characteristics. This is a sophisticated procedure for distinguishing between true and false information, and they employ both reference benchmark datasets. We can also utilize K-means clustering and deep learning approaches to boost accuracy. (Andreadou et al., 2015)

The study is a survey on the automatic identification of bogus news using natural language processing. As fake news grows and spreads more rapidly, it hinders people's ability to respond to any misfortune. Therefore, automated detection of bogus news is employed. The primary objective of this study is to describe the issue formulations, datasets and Natural language processing approaches used to solve it, as well as to provide alternative solutions if anything goes awry. Therefore, this is not a conventional technique, and the instruments utilized play a significant part in forecasting bogus news. (Oshikawa et al., 2020)

Buntain & Golbeck, 2017 analyzing two Twitter datasets. CREDBANK is a crowdsourced dataset and PHEMEis an assessment dataset for probable rumours on Twitter and journalists. This research aims to demonstrate why non-experts' models perform better than journalists models for detecting fake news on twitter, The shortcomings of this article is that it only uses popular tweets from Twitter, and its accuracy of 65.29 percent falls short of expectations. This research contains numerous flaws and may have benefited from the appropriate models and methodology. (Buntain & Golbeck, 2017)

Edge computing an blockchain enable the internet of Vehicles to detect bogus news quickly. Incorrect messages on the Internet of Vehicles might cause traffic problems; consequently, it is essential to identify phony news. This study describes a framework – Rapid fake news detection – that utilizes technologies such as blockchain. The blockchain servers use vehicle reports to determine the likelihood of traffic. The implementation of blockchain was a brilliant idea, and the project's efficiency and precision are both excellent. (Xiao et al., 2020)

3 Research Methodology

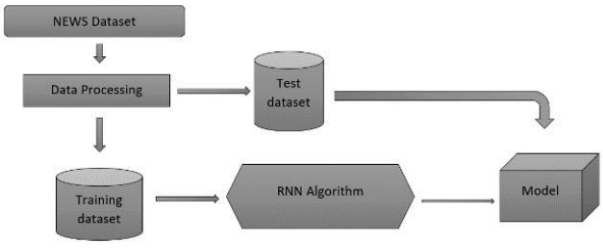

Figure 1: Research Framework

This section explains the details of the dataset, machine learning methodologies, and comparative performance metrics for the models. This project was largely inspired by Phuong and Anh Cuong le's Neural Network models's comparative study (Le-Hong & Le, 2019) for Sentence Classification; however, the focus of this study is on recurrent neural networks with LSTM and NLTK, which is an extension of the previous publication. This study employs KDD (Knowledge Discovery in Database) approach. Figure 4 depicts the phases of the Research.

3.1 Dataset

For fake news detection, proper dataset is essential. The dataset used is taken from a public site Kaggle and can be downloaded. The dataset contains 6335 feature extraction records that were divided into two types – Real and Fake. The dataset has details about news like Headline, Body , 'URL' from which the news is taken and finally label.

3.2 Data Pre-processing

The dataset containing real and fake news was taken from kaggle website. Jupyter lab is utilized for the cleansing and transformation of the dataset. Initially, Python libraries such as 'pandas', 'Numpy', 'matplotlib', and 'glob' are loaded. The 'pandas' read_csv function and the glob library are utilized to retrieve rows from each file. The dataset is examined for null and unusual values. For data collection, deep learning needs a substantial amount of historical data. The collected information contains significant historical context and cannot be used directly without being pre-processed. Coaching and testing this model to verify its proper operation and minimal inaccuracy in predictions. Figure 5 depicts distribution of fake and real news.

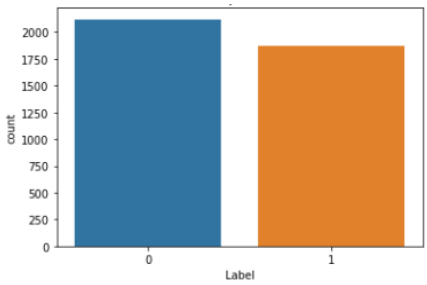

Figure 2: Distribution of Real and Fake news

4 Design Specification

The techniques used in this study are Recurrent Neural Network with LSTM (Long Short – Term Memory), NLTK, Decision Tree and Logistic Regression and the results are compared. All these models are used for fake news detection.

As there will be lot of processing going on in RNN – LSTM with GloVe embedding (Sharma et al., 2021), the accuracy expected is high. Also GloVe word embedding is superior compared to bag of words and other word embeddings. Other studies have not explored the

model building using RNN-LSTM, instead most of them have used FNN-LSTM and word embedding is an add on. The accuracy can be greatly improved from feature processing.

The headline of the news has many stopwords which do not hold any value in analysis, therefore it needs to be removed and words of utmost relevance shall be used in the preparation of word cloud using Natural Language toolkit.

ML models like KNN, SVM have shown good results in detecting the fake news. Decision tree classifier and Logistic Regression might not perform well compared to RNN but with few tweaks in the program like feature processing, commendable results can be expected.

5 Implementation

5.1 Data preparation

After data is cleaned and processed, it needs to be split into training and testing sets. The optimum split is said to be 70 % training and 30 % testing but this can vary according to the situation, model etc. Respective models are imported from sklearn and divided into training and testing sets. Accuracy and evaluation metrics are taken for each model.

5.2 Models of Machine Learning

5.2.1 Natural Language Processing (NLP)

Natural language processing allows machines to read and comprehend human language (NLP). A sophisticated linguistic communication process system would transform natural – language user interfaces by enabling the direct acquisition of information from human – written sources, such as newswire text. Information retrieval, text mining (Mahmud et al., 2021), inquiry responsiveness, and Artificial Intelligence are examples of implementations of the linguistic communication procedure.

NLP's Natural Language Toolkit is utilized in this investigation. The NLTK library is imported using pip function. Importing stopwords library enables the detection of stopwords in the text. Stopwords include 'the', 'a', 'an', 'in'; as these types of words have no analytical significance, they are eliminated during the procedure. The Porter Stemmer Algorithm is a method for removing the suffixes from an English word in order to extract its stem, which is highly valuable in the field of Information Retrieval (IR). Whitespace removal, text to lowercase, punctuation removal, tokenization of each message, and removal of single-letter terms are among the procedures conducted prior to separating the dataset int training and testing sets containing 80 and 20 percent of the data respectively.

5.2.2 Decision Tree Classifier

It is based on an algorithm that necessitates the quantitative and categorical nature of the data to be analyzed. Consequently, continuous data will not be assessed. The first option, subtree replacement, relates to the ability to change nodes in the leaves of a decision tree to reduce the number of tests along the convincing path. Subtree raising has a small impact on decision tree models in the majority of instances. In most circumstances, there is no precise way for estimating the usefulness of an option, however it may be prudent to disable it if the

induction procedure takes longer than anticipated because the subtree's raising is computationally challenging. From sklearn.tree, the required library DecisionTreeClassifier is imported and assigned to dtree variable. Just before this, the dataset is divided into training and testing sets containing 70 and 30 percent of the data respectively.

5.2.3 Logistic Regression

A split variable is utilized for calculation purposes (in that there are only two possible outcomes). Provision regression identifies the model that best explains the relationship between a split feature of interest and a number of independent factors. Provision regression is a classification technique that forecasts the probability of a category variable using Machine Learning. In provision regression, a binary variable with information coded together or a value could be employed. Alternatively, the LR model predicts the P(Y=1) function.
Assumptions:

- In binary logistic regression, the dependent variable must be binary.
- The problem level is one of the variable outputs used to represent the desired outcome in a binary regression.
- Only the applicable variables should be encapsulated.
- Each alternative should have its own set of independent variables. In other words, the model should contain a minimum amount of data.

Similarly, from sklearn, the library LogisticRegression is imported and assigned to logR variable. Even for this model the dataset is divided into training and testing sets containing 70 and 30 percent of the data respectively.

5.2.4 RNN – LSTM with GloVe word embedding

A feed- forward neural network with contained memory is generalized into a recurrent neural network. RNN is continuous because it does constant operations for each knowledge input and the output of each input is defined by previous calculations. It is traced and transmitted back to the continuous network when the output is manufactured. Before making a call, this input and the outcome learned from the previous input are considered. RNNs, unlike FNN, utilize their internal state to process input sequences. Therefore, they can be used for tasks such as unified, connected handwriting recognition and speech recognition. Inputs to distinct neural networks are fully independent of one another.

LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in machine translation etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems.

GloVe is a word vector technique that, after a brief pause, rode a wave of word vectors. Word vectors place words in a vector space in which like-words cluster together and dissimilar words repel one another. Glove's advantage over Word2vec is that, unlike Word3vec, it used global statistics in addition to local statistics to generate word vectors.

After initialising the RNN and adding the first LSTM layer and some dropout regularisation, it returns a sequence of vectors of dimension 64, a second LSTM layer is added and some dropout regularisation too. Before compiling the RNN an outer layer is also added and tested for 10 epochs. MSE and RMSE are also evaluated. With the help of Keras and tensorflow

embedding is introduced and the embedding size is taken as 100. But this time it is tested for 30 epochs. The loss, accuracy, precision and recall are calculated.

6 Evaluation

In current section, performance of the machine learning models created per Section 5 in the detecting the fake news is examined critically. These are few evaluation measures utilized in the study:

- Classification Accuracy – It is number of correct predictions divided by total umber of predictions made.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Figure 3: Classification accuracy formula

- Logarithmic Loss – It functions effectively for multi-class classification. The classifier must assign probabilities to each class for each sample while working with log loss.

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

Figure 4: Logarithmic loss formula

- Confusion Matrix – Confusion Matrix, as the name implies, describes the full performance of the model by means of a matrix.

- Sensitivity – It is true positive divided by false negative + true positive.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

Figure 5: Sensitivity formula

- Specificity – It is true negative divided by true negative + false positive

$$TrueNegativeRate = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

Figure 6: Specificity formula

- F1 Score – It is defined as the harmonic mean between recall and precision.

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

F1 Score

Figure 7: F1 Score formula

- Precision – It is true positives divided by true positives + False positives

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Precision

Figure 8: Precision formula

- Recall – It is true positives divided by true positives + false negatives

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Recall

Figure 9: Recall formula

- Mean Squared Error – It is very similar to Mean Absolute Error, with the exception that MSE calculates the average of the square of differences between predicted values and actual values.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

Mean Squared Error

Figure 10: MSE formula

6.1 Evaluation and results of Decision Tree Classification

The classification accuracy of the system in detecting false news was evaluated using a variety of evaluation metrics. This section employs the most generally employed metric for detecting fake news – confusion matrix.
The evaluation metrics - sensitivity, specificity, recall, precision, f1-score have been taken. Sensitivity for model is 0.9968 whereas specificity for the model is 0.99286.

Fake news detection model	label	Precision	Recall	F1-score	Support	Accuracy	Macro-F1
Decision Tree Classifier	0	0.99	1.00	1.00	636	0.99	0.99
	1	1.00	0.99	0.99	561		

Figure 11: Classification report of Decision Tree Classifier

```
Sensitivity :  0.9968553459119497

Specificity :  0.9928698752228164
```

Figure 12: Specificity and Sensitivity

```
Confusion Matrix result of Decision Tree Classifier : is:
 [[634   2]
 [  4 557]]
```

Figure 13: Confusion Matrix

```
Accuracy result ofDecision Tree Classifier  is: 99.49874686716792
```

figure 14: Accuracy

6.2 Evaluation and results of Logistic Regression

The accuracy of the Logistic regression model is low compared to Decision tree classifier. Anyhow the confusion matrix has been taken along with evaluation matrix – precision, f1-score, sensitivity, recall, specificity. So, specificity for logistic regression is 0.7754 and sensitivity is 0.84433.

```
Confusion Matrix result of Logistic Regression : is:
 [[537  99]
 [126 435]]
```

Figure 15: Confusion Matrix of Logistic Regression

```
Sensitivity :  0.8443396226415094

Specificity :  0.7754010695187166
```

Figure 16: Sensitivity & Specificity

Fake news detection model	label	Precision	Recall	F1-score	Support	Accuracy	Macro-F1
Logistic Regression Regressor	0	0.81	0.84	0.83	636	0.81	0.81
	1	0.81	0.78	0.79	561		

Figure 17: Classification report of Logistic Regression.

```
Accuracy result of Logistic Regression is: 81.203007518797
```

Figure 18: Accuracy of Logistic Regression

6.3 Evaluation and results of Natural language toolkit

Matplotlib and Seaborn libraries are used in plotting graphs which provide insights about the data which in return is used in evaluation of the model.

The first graph is a distribution of length plotted against every headline in the dataset.

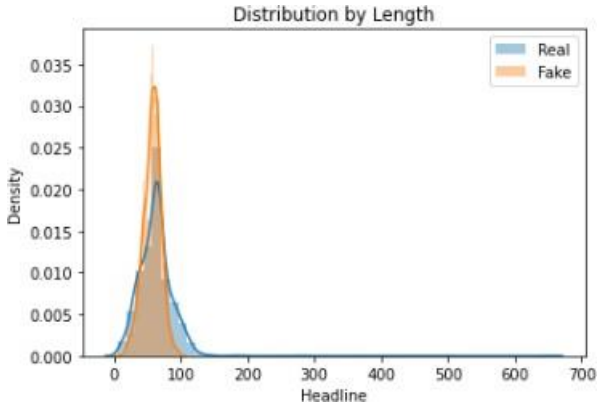

Figure 19: Distribution by length

The next graph is a distribution by digits plotted against headline in the dataset.

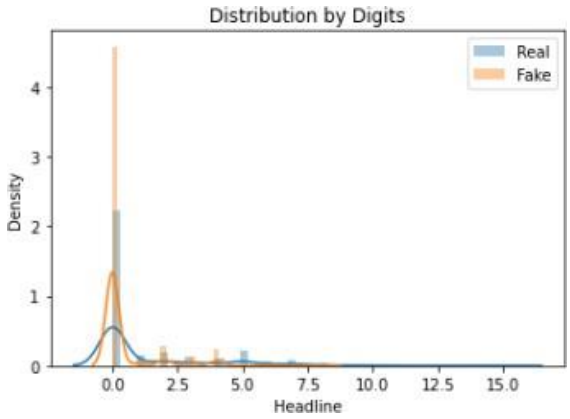

Figure 20: Distribution by Digits

The next graph is a distribution of non-digits plotted against headline in the dataset.

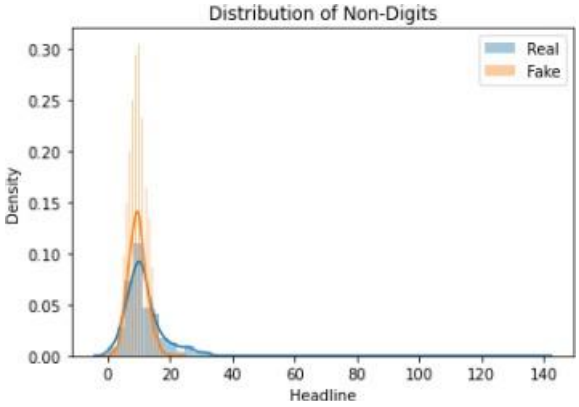

Figure 21: Distribution of Non-Digits

A new vocabulary is created for the training set with word count which is later used in the preparation of word cloud. Wordcloud, a technique that displays which words are the most frequent in the given text. Therefore, the words from actual news are placed in one variable, and the same is done for fake news. Figure 6 shows the wordcloud for fake news and figure 7 shows the wordcloud for real news.

Figure 22: Fake news word cloud

Figure 23: Real news word cloud

6.4 Evaluation and results of RNN - LSTM

The Accuracy, precision and recall are calculated. With every epoch all these values increased gradually and also word embedding is also a reason for fairly high Accuracy. MSE and RMSE are also taken for training and testing data.

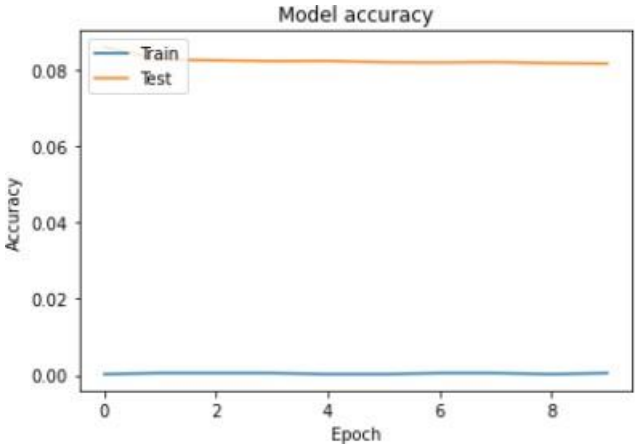

Figure 24: Accuracy vs epoch

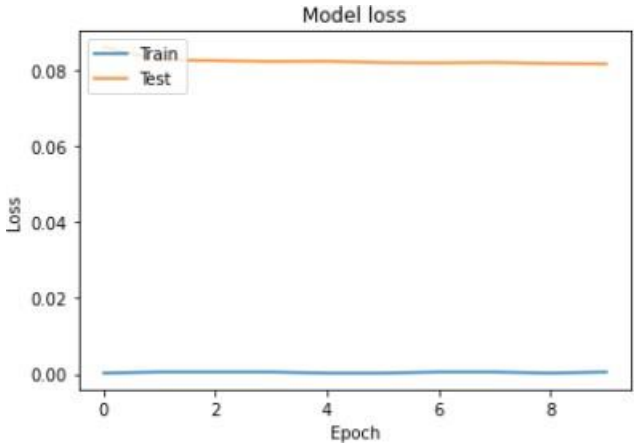

Figure 25: Loss vs epoch

```
Train Score: 0.08 MSE (0.28 RMSE)

Test Score: 0.08 MSE (0.29 RMSE)
```

Figure 26: MSE and RMSE evaluation

```
[+] Accuracy: 82.33%
[+] Precision:   86.06%
[+] Recall:   86.06%
```

Figure 27: Accuracy, Precision and Recall of LSTM

6.5 Discussion

Out of all the models, RNN – LSTM performed the best with less loss and good accuracy which was expected but the accuracy of Decision tree was not expected to be this high. Natural Language toolkit has performed well in depicting important words, finally the room for improvement is in case of logistic regression with average metrics. All the values have been recorded.

7 Conclusion and Future Work

A fruitful result was attained in three of the models which were greatly inspired from studies which are mentioned in this paper. Fake news is a complex problem for any model, but LSTM with GloVe could handle much more. Decision tree with few more tweaking could increase its accuracy a little more. With extra epochs, the accuracy for RNN could also have been increased. The essence of any news is words, with right processing and analysis, any fake news can be detected and very soon we will be building high-end models with which finding fake news will be hassle-free.

There are numerous facets to future work. These models can be built deeper to achieve a superior outcome. In addition, researchers can experiment with various model parameters and layers in the LSTM. Different word embeddings can team up with LSTM to test for higher accuracies. Also, by taking up a basic machine learning model and performing many operations on it such as feature selection and keep changing it over the time can bring great results.

8 Acknowledgement

I would like to express my sincere gratitude to my supervisor Hicham Rifai, who made this work possible. His patience, motivation and immense knowledge helped me all the throughout the research and writing of thesis. I am also thankful to God, my family, friends and colleagues for their support and motivation during the coursework.

References

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10618 LNCS*, 127–138. https://doi.org/10.1007/978-3-319-69155-8_9

Andreadou, K., Papadopoulos, S., Apostolidis, L., Krithara, A., & Kompatsiaris, Y. (2015). Media REVEALr: A social multimedia monitoring and intelligence system for web multimedia verification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9074*, 1–20. https://doi.org/10.1007/978-3-319-18455-5_1

Buntain, C., & Golbeck, J. (2017). Automatically Identifying Fake News in Popular Twitter Threads. *Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017*. https://doi.org/10.1109/SmartCloud.2017.40

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as "false news." *WMDD 2015 - Proceedings of the ACM Workshop on Multimodal Deception Detection, Co-Located with ICMI 2015*. https://doi.org/10.1145/2823465.2823467

D'Ulizia, A., Caschera, M. C., Ferri, F., & Grifoni, P. (2021). Fake news detection: A survey of evaluation datasets. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/PEERJ-CS.518

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings*. https://doi.org/10.1109/UKRCON.2017.8100379

Gupta, A., & Kaushal, R. (2015). Improving spam detection in Online Social Networks. *Proceedings - 2015 International Conference on Cognitive Computing and Information Processing, CCIP 2015*. https://doi.org/10.1109/CCIP.2015.7100738

Jin, Z., Cao, J., Jiang, Y. G., & Zhang, Y. (2014). News Credibility Evaluation on Microblog with a Hierarchical Propagation Model. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2015-January*(January), 230–239. https://doi.org/10.1109/ICDM.2014.91

Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet – A deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*. https://doi.org/10.1016/j.cogsys.2019.12.005

Le-Hong, P., & Le, A. C. (2019). A Comparative Study of Neural Network Models for Sentence Classification. *NICS 2018 - Proceedings of 2018 5th NAFOSTED Conference on Information and Computer Science*. https://doi.org/10.1109/NICS.2018.8606879

Mahmud, Y., Shaeeali, N. S., & Mutalib, S. (2021). Comparison of Machine Learning Algorithms for Sentiment Classification on Fake News Detection. *International Journal of Advanced Computer Science and Applications*, *12*(10). https://doi.org/10.14569/IJACSA.2021.0121072

Miller, B. A. P. (2016). Automatic Detection of Comment Propaganda in Chinese Media. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2738325

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1). https://doi.org/10.1016/j.jjimei.2020.100007

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. *International Conference on Information and Knowledge Management, Proceedings*, *Part F131841*, 797–806. https://doi.org/10.1145/3132847.3132877

Sharma, R., Agarwal, V., Sharma, S., & Arya, M. S. (2021). An LSTM-Based Fake News Detection System Using Word Embeddings-Based Feature Extraction. *Lecture Notes in Networks and Systems, 154*. https://doi.org/10.1007/978-981-15-8354-4_26

Tacchini, E., Ballarin, G., della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it Hoax: Automated fake news detection in social networks. *CEUR Workshop Proceedings, 1960*.

Xiao, Y., Liu, Y., & Li, T. (2020). Edge computing and blockchain for quick fake news detection in IoV. *Sensors (Switzerland), 20*(16). https://doi.org/10.3390/s20164360

Zhou, X., Jain, A., Phoha, V. v., & Zafarani, R. (2020). Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice*, *1*(2). https://doi.org/10.1145/3377478