Semester 1 – 2021/2022

Course Code	DS550
Course Name	Machine Learning
Assignment type	Project
Module	ALL

Student ID	G200002614	G200007615
Student Name	Enad S. Alotaibi	Abdulaziz M. Alqumayzi

Fake News

Enad Saud Alotaibi
G200002614@seu.edu.sa,
Saudi Electronic University,
Riyadh, Saudi Arabia

Abdulaziz Mohammed Alqumayzi
G200007615@seu.edu.sa,
Saudi Electronic University,
Riyadh, Saudi Arabia

**Corresponding Author*

ABSTRACT

Increasing in use of social media and access to every user, there is a lot of false news on social media that helps them to scam other people financially, Fake news creates economic issues to our society, we have to detect and stop using our cyber security for upcoming threads. In this work, we have collected a dataset containing title, author names, and textual information of news with their labels, we have applied NLP and machine learning with deep learning methods and optimized model attributes and engineered featured to get the highest accuracy of the model with less data input.

Literature review

With the growth of the increase in the use of social media, low of posts are related to the news without confirmation that leads to cast and post fake news. In recent years works found to be focused on detecting fake news. In 2017 online automatic fake news detection, in which they used a two-fold novel dataset for fake news detection, that leads them to analyze automatic fake news detection and manually fake news detection [1]. For fake news detection the formulations, datasets, and NLP solutions are to be applied and analyze their behavior on a massive amount of Web content [2]. Fundamental theories, formulas from various disciplines empower research in fake news detection datasets and processing techniques. Research has been conducted in 2019, developed models from various perspective that involves data mining techniques, news content and social media information, social search, machine learning, and natural language processing (NLP), introduced challenges for US presidential election to clarify fake news detection for automatic, effective and in efficient way [3]. In addition, exploring main features Julio C. S. Reiss, André Correia

proposed a new set of features and measure accuracy and performance of the model that is currently used, and finding useful information that can be used in practice, challenges, highlights, and opportunities [4]. Kai Shu; Suhang Wang construct real-world datasets gathered from users containing fake and real news and select representative users to detect fake news items and more likely to fake news and perform comparative analysis, which reveals their potential to differentiate fake news [5].

1. Introduction

As growing into social media and news spread very fast, it is very challenging for cyber security and another department to stop spreading fake news. In this work, we took data containing fake and real news with labels. That we have used to develop our TFIDF and word embedding models to detect whether the news is fake or real, the study uses machine learning and deep learning simple problem used by preprocessing our dataset using natural language processing (NLP) to remove unnecessary information. We respect complex and detailed techniques of deep learning and machine learning, but sometimes complexity is not being guaranteed for better classification and detection.

2. Body section

2.1 Data

Data contained 20078 rows and six columns with shapes of (20078, 6), in which the first 2 columns, index, and id were dropped out, both contained rows numbers. There remains column named, title, author, Text, and label. The description of the data has shown in table 1 below.

***Table 1**: Data Sescription*

Column name	Non-null values	Null values	Data Type	Description
Title	19529	549	String	Title of news
Author	18221	1857	String	Author name of news
Text	20039	39	String	Text description of news
Label	20078	0	Integer	Type of news (Real or Fake)

2.2 Methods

Data contains four columns after deleting the first two unnecessary columns, but for the classification model, only a text column was used to create the model for news classification along with their label types associated with them. Only 39 news descriptions were missing, it was better to drop empty text rows. Before model creation and training, news text was preprocessed for removing non-unique data by removing strop words, repeating characters, URLs, Emails. After that lemmatization and stemming methods were also applied to make words to their base word, and after preprocessing, data contained only unique words that were converted into vector form using TFIDF and Word Embedding methods.

Two type models (Logistic Regression and LSTM) were created and analyzed with two types of vectorization method (TFIDF and Word Embedding) that was used to reshape input data for the model. Later on, models were optimized for getting a better result. Data split into training and testing data with ratios of 80% and 20% accordingly.

Fit transform was applied on default TFIDF vectorizer for learning vocabulary and IDF that return document-term matrix. Matrix was input using supervised learning to Logistic Regression. For TFIDF vectorizer and features, max_features and n-grams were used. Max-features none to 5000 that build a vocabulary that only considers the top max_features ordered by term frequency across the corpus. N-grams value used (1, 3) that is the lower and upper boundary of the range of n-values for different n-grams to be extracted. Value of inverse of regularization strength increased to 5 in Logistic Regression.

Top 2000 maximum number of words to keep, based on word frequency from data text and turn into matrix form using word embedding techniques. Max length of the matrix keeps stable at 500. Embedding layer used for turn 2000 dimension to 50 output dimension followed by Tensor Flow input layer, fed into LSTM layer with a unit of 32 using 0.01 value of l2 type of kernel_regularizer for the linear transformation of the inputs.

Drop out layer takes input and randomly sets input units to 0.75 with a frequency of rate at each step during training time, which helps the model to prevent overfitting. The output layer consists of 1 unit with l2 type of kernel_regularizer with a value of 0.01. The architecture of the model is shown in fig 1.

Layer (type)	Output Shape	Param #
inputs (InputLayer)	[(None, 500)]	0
embedding_2 (Embedding)	(None, 500, 50)	100000
lstm_2 (LSTM)	(None, 32)	10624
dropout_2 (Dropout)	(None, 32)	0
out_layer (Dense)	(None, 1)	33

Total params: 110,657
Trainable params: 110,657
Non-trainable params: 0

Figure 1: *LSTM Model*

In optimization words embedding strategy keep same as LSTM model changes into the model by adding 64 units in LSTM, one denser layer with 16 units. L2 type of kernel_regularizer with a value of 0.01 used into a hidden layer, add Relu activation function on a hidden layer and Sigmoid function has been used at last dense layer. Drop out value into drop layer reduced to 0.50. The architecture of the model is shown in fig 2.

Layer (type)	Output Shape	Param #
inputs (InputLayer)	[(None, 500)]	0
embedding_1 (Embedding)	(None, 500, 50)	100000
lstm_1 (LSTM)	(None, 64)	29440
dropout_1 (Dropout)	(None, 64)	0
dense (Dense)	(None, 16)	1040
activation_1 (Activation)	(None, 16)	0
out_layer (Dense)	(None, 1)	17
activation_2 (Activation)	(None, 1)	0

Total params: 130,497
Trainable params: 130,497
Non-trainable params: 0

Figure 2: *Artitecture of model*

2.3 Analysis

Logistic Regression learns with an accuracy of 97.35%. TFIDF and engineered featured model gives slightly higher accuracy 1.3% train accuracy and 0.63% test accuracy compared to the previous model. LSTM model has 88.95% train accuracy with 90.36% test accuracy. Optimized LSTM gives 98.78% train accuracy and 94.89% test accuracy, trained on 2000 topmost common embedded words. LSTM and optimized LSTM accuracy graph throughout supervised learning has shown in figures 3 and 4.

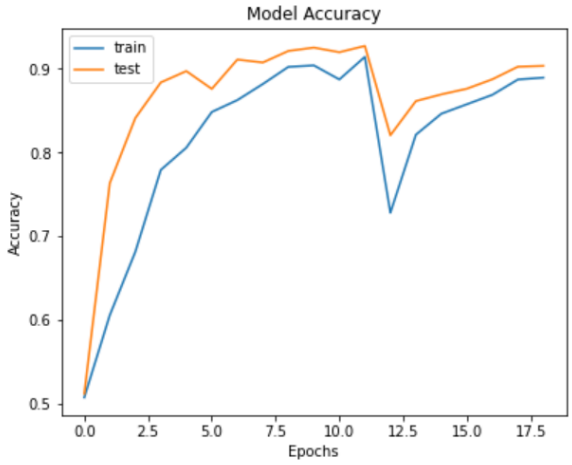

***Figure 3**:LSTM Model Accuracy*

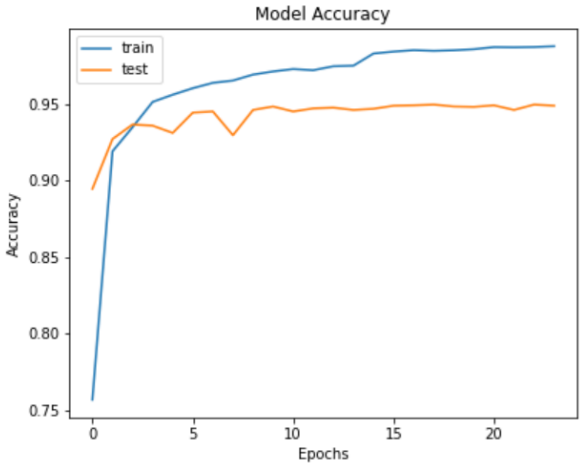

***Figure 4:** LSTM Optimized Model Accuracy*

2.4 Results

2.4.1 TFIDF Vectorizer

TFIDF Vectorizer predicts 110 fake news as the wrong prediction out of 2045 fake news and predicts 83 real news as fake news in the total of 1948 with test accuracy of 95.16%, while training accuracy of the model remains 97.35 %.

2.4.2 TF-IDF + the engineered features

This model was trained with an accuracy of 98.65 % with a test accuracy of 95.79 %. The model predicted 92 Fake news as real news while other 1953 was predicted correctly, on the other hand, the model predicted 76 real news as fake news out of 1948.

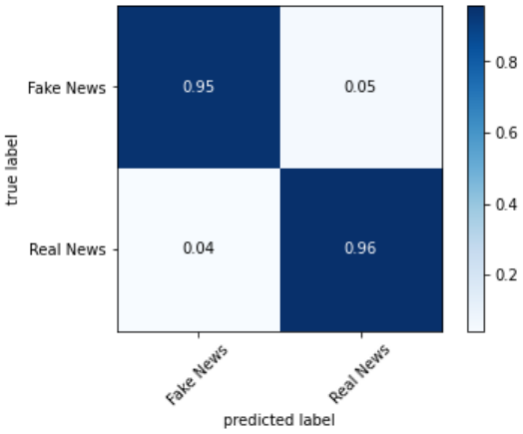

***Figure 5**: TFIDF Prediction matrix*

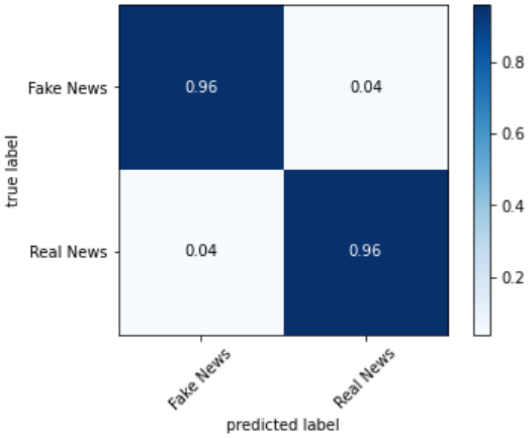

***Figure 6**: TFIDF + engineered featured Prediction matrix*

2.4.3 Word Embedding with LSTM

Word Embedding with Embedding layer attached with LSTM, Dropout, and Dense layer gave 88.95 % train accuracy, while test accuracy was 90.36%. The model predicts wrong predictions on 154 fake news and 166 real news.

2.4.4 Word Embedding with LSTM Optimized Model

Previously used LSTM model with optimization gave us better results with 98.78 % train accuracy while test accuracy was 94.89 %, while wrong was on fake news did not increase from 101 and real news category got 103 news as wrong predicted.

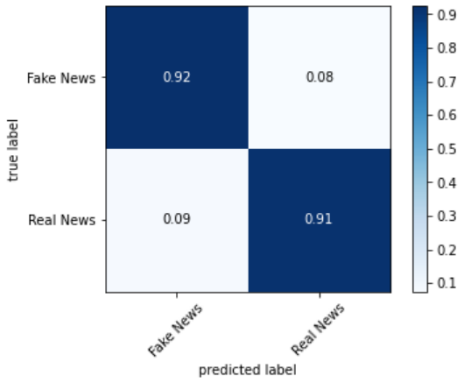

***Figure 7:** LSTM Prediction Matrix*

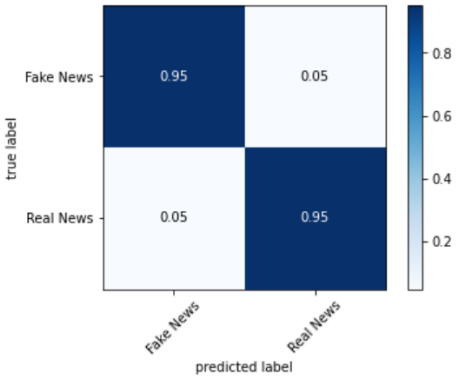

***Figure 8:** Optimized LSTM Prediction Matrix*

2.4.5 Performance Comparison

Table 2: *Model performance Comparison*

Vector Representation	Vector Representation Features	Classification Model	Model Features	Train Accuracy %	Test Accuracy %
TFIDF Vectorizer	Default	Logistic Regression	Default	97.35	95.16
TFIDF Vectorizer	max_features = 5000 ngram_range = 1, 3	Logistic Regression	C=5	98.65	95.79
Word Embedding	max_len = 500 num_words=2000	LSTM	1 x Embedding 1 x LSTM 1 x Drop Out 1 x Dense	88.95	90.36
Word Embedding	max_len = 500 num_words=2000	Optimized LSTM	1 x Embedding 1 x LSTM 1 x Activation Relu 1 x Drop Out 2 x Dense 1 x Activation Sigmoid	98.78	94.89

2.5 Further analysis

From cleaned data, separate fakes news from data and choose frequency distribution from the NLTK probability method. The top three news titles that are commonly repeated separated and choose five most similar words associated with them and visualize them on t-distributed stochastic neighbor embedding shown in below fig 9, words and most commonly three titles are shown in table 4.

Table 3: *Top Three Most Common Tittles*

Title	Frequency Distribution
What to Cook This Week - The New York Times	4
Right and Left: Partisan Writing You Shouldn't Miss - The New York Times	2
17 Great Stories That Have Nothing to Do with Politics - The New York Times	2

Table 4*: Top Five words associated with top titles*

Top five similar words	Similarity (%)
Story	0.998
TV	0.998
Rapped	0.998
Live	0.998
Happy	0.998

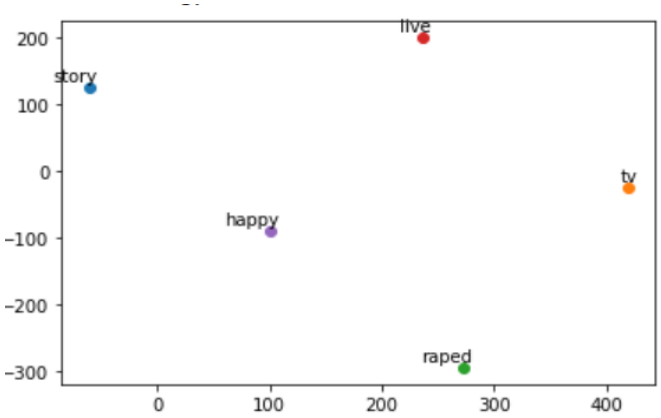

Figure 9: *T-SNE Visualization*

3. Conclusion

Logistic Regression followed by TfidfVectorizer has 95.16 % test accuracy while TFIDF with engineered featured give 0.63% higher test accuracy. LSTM close on 90.36% test accuracy, while optimized LSTM has a much better test accuracy of 94.85% compared to the LSTM. Detecting news as fake or real can be very important and useful for cyber security threats. Using more datasets including the title of news with a description of news may help to detect news with less textural data that can be linked to real-world newscasting datasets to predict if this has to happen or would it be happened.

References

[1] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv:1708. 07104.

[2] Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811. 00770.

[3] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019, January). Fake news: Fundamental theories, detection strategies and challenges. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 836-837).

[4] Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. IEEE Intelligent Systems, 34(2), 76-81.

[5] Shu, K., Wang, S., & Liu, H. (2018, April). Understanding user profiles on social media for fake news detection. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 430-435). IEEE.

[6] Gobal, M. (2017) Applied Machine Learning. 1st Edition. McGraw Hill.

[7] Bonaccorso, G. (2018). Machine learning algorithms. 2nd Edition. Packt Publishing. ISBN: 9781789347999.