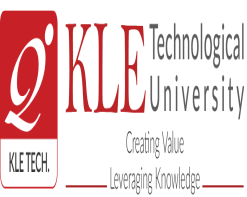
KLE Society's

KLE Technological University



**Data Mining and Analysis (18ECSC301)**

**Course Project Report**

**On**

**COVID-19 Fake news detection**

### Anup S H 01fe18bcs041

### Ankita Patil 01fe18bcs039

### Ayazahemad Shaikh 01fe18bcs058

### Abhinav Kumar 01fe18bcs008

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBLI–580 031 (India).

Academic year 2020-21

**Content Table**

|  |  |
| --- | --- |
| **SI.no** | **Content** |
| 1 | Introduction |
| 2 | Problem Statement |
| 3 | Literature Survey |
| 4 | Methodology |
| 5 | Results |
| 6 | Conclusion |
| 7 | Reference |

1. **INTRODUCTION**
2. There is no particular deﬁnition of fake news even in Journalism. Over the years, various researchers
3. have tried deﬁning fake news. Fake news in people’s lives is a source of spam which has the po-
4. tential to inﬂuence public perception and knowledge in a major way[1]. Fake news has the capa-
5. bility of reaching a broad audience, causing economic damages and manipulating in major political
6. outcomes[2]. Fake news can be described as a particular type of misinformation (false information)
7. that is purposely sent out to deceive people[3]. Fake news in social networking sites is intentionally
8. crafted fake information, which is circulating in the form of spam, conspiracy theories and hoaxes
9. on social media[4]. An article from the BBC described fake news as false information intentionally
10. spread by those who are scanty about the facts and act on the grounds of promoting political causes
11. or gaining revenue from web trafﬁc[5].
12. There is no particular deﬁnition of fake news even in Journalism. Over the years, various researchers
13. have tried deﬁning fake news. Fake news in people’s lives is a source of spam which has the po-
14. tential to inﬂuence public perception and knowledge in a major way[1]. Fake news has the capa-
15. bility of reaching a broad audience, causing economic damages and manipulating in major political
16. outcomes[2]. Fake news can be described as a particular type of misinformation (false information)
17. that is purposely sent out to deceive people[3]. Fake news in social networking sites is intentionally
18. crafted fake information, which is circulating in the form of spam, conspiracy theories and hoaxes
19. on social media[4]. An article from the BBC described fake news as false information intentionally
20. spread by those who are scanty about the facts and act on the grounds of promoting political causes
21. or gaining revenue from web trafﬁc[5].

There is no particular definition of fake news even in Journalism. Over the years, various researchers have tried defining fake news. Fake news in people’s lives is a source of spam which has the potential to influence public perception and knowledge in a major way. Fake news has the capability of reaching a broad audience, causing economic damages and manipulating in major political outcomes. Fake news can be described as a particular type of misinformation (false information) that is purposely sent out to deceive people. Fake news in social networking sites is intentionally crafted fake information, which is circulating in the form of spam, conspiracy theories and hoaxes on social media.

An article from the BBC described fake news as false information intentionally spread by those who are scanty about the facts and act on the grounds of promoting political causes or gaining revenue from web traffic. The existence of fake news has been since ages when the time when information was passed by writing on stone. The structure of the community at that time allowed knowledge and information be controlled by the individuals leading the community. Such scenario led to leaders of the group spreading false information to fame or defame someone. The development of printing press created a society in which literate people could easily spread false information to illiterate people. Mass media era led development of radio and television programs as means of spreading deceiving information. Since the late 20th century, Internet has been used as a new means of propagating news. With advancement of social media today, fake news began spreading so widely that people now realize its implications.

Fake news is just not about misleading people. It is also connected to several psychological foundations. People tend to believe stories that agree with their pre-disposed opinion. Also, people only choose to receive those information that matches with their ideology. News publishers also have their own specific goals. While reputed media tends to achieve reputation among their readers, fake news media have a very short-term goal of spreading lies, creating chaos and have some form of advantage of then developed situation. The naive realism in people and such short-term goal of publisher are the main reason why fake news exists in traditional media is prevalent. With the Internet, people believe in information on the internet to be true if it possesses approval of a large number of people. Fake news today has its impact on world politics, society and economics. Last four decades of deception research and the wide propagation of fake news can be considered as a proof that we, human beings, are not very good at differentiating what is right or wrong in the context of receiving information. The understanding of the intention of the language in news stories and their authors is also a very difficult task because different people understand language differently. For this reason, same news can be considered as both real and fake by different groups of people. The amount of data flowing in on social media and internet every day is huge. Human intervention of each and every news articles is an almost impossible task. With exabyes of content flowing every day, and the inability of humans to stop or otherwise intervene in fake news production and propagation, researchers have started exploring hybrid approaches to classification of fake news content. Research on fake news detection and classification is focused on understanding natural language and text processing using convolution neural network.

There is no particular deﬁnition of fake news even in Journalism. Over the years, various researchers

have tried deﬁning fake news. Fake news in people’s lives is a source of spam which has the po-

tential to inﬂuence public perception and knowledge in a major way[1]. Fake news has the capa-

bility of reaching a broad audience, causing economic damages and manipulating in major political

outcomes[2]. Fake news can be described as a particular type of misinformation (false information)

that is purposely sent out to deceive people[3]. Fake news in social networking sites is intentionally

crafted fake information, which is circulating in the form of spam, conspiracy theories and hoaxes

on social media[4]. An article from the BBC described fake news as false information intentionally

spread by those who are scanty about the facts and act on the grounds of promoting political causes

or gaining revenue from web trafﬁc[5].

The existence of fake news has been since ages when the time when information was passed by

writing on stone. The structure of the community at that time allowed knowledge and information

be controlled by the individuals leading the community. Such scenario led to leaders of the group

spreading false information to fame or defame someone. The development of printing press created a

society in which literate people could easily spread false information to illiterate people. Mass media

era led development of radio and television programs as means of spreading deceiving information.

Since the late 20th century, Internet has been used as a new means of propagating news. With

advancement of social media today, fake news began spreading so widely that people now realize its

implications.

1

Fake news is just not about misleading people. It is also connected to several psychological

foundations[6]. People tend to believe stories that agree with their pre-disposed opinion. Also,

people only choose to receive those information that matches with their ideology. News publishers

also have their own speciﬁc goals. While reputed media tends to achieve reputation among their

readers, fake news media have a very short term goal of spreading lies, creating chaos and have

some form of advantage of then developed situation. The naive realism in people and such short

term goal of publisher are the main reason why fake news exists in traditional media is prevalent.

With the Internet, people believe in information on the internet to be true if it possesses approval of

a large number of people[7].

Fake news today has its impact on world politics, society and economics. Last four decades of

deception research and the wide propagation of fake news can be considered as a proof that we,

human beings, are not very good at differentiating what is right or wrong in the context of receiving

information[8]. The understanding of the intention of the language in news stories and their authors

is also a very difﬁcult task because different people understand language differently. For this reason,

same news can be considered as both real and fake by different groups of people. The amount

of data ﬂowing in on social media and internet everyday is huge. Human intervention of each

and every news articles is an almost impossible task. With exabyes of content ﬂowing everyday,

and the inability of humans to stop or otherwise intervene in fake news production and propagation,

researchers have started exploring hybrid approaches to classiﬁcation of fake news content. Research

on fake news detection and classiﬁcation is focused on understanding natural language and text

processing using machine learnin

**2. PROBLEM STATEMENT**

Detection of COVID-19 related fake news in English. Given a social media post, the objective of the shared task is to classify it into either fake or real news.

If you take Crocin thrice a day you are safe. Fake

Wearing mask can protect you from the virus. Real

**3. LITERATURE SURVEY**

The author in this work, proposed to use machine learning ensemble approach for automated classification of news articles. The study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, they trained a combination of different machine learning algorithms using various ensemble methods and evaluated their performance on 4 real world datasets. Experimental evaluation confirms the superior performance of their proposed ensemble learner approach in comparison to individual learners [1].

The author purposes the work to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. They used simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4% accuracy using logistic classifier. [2]

The author in this paper addresses the challenges introduced by the unknown characteristics of fake news and diverse connections among news articles, creators and subjects. This paper introduces a novel automatic fake news credibility inference model, namely Fake Detector. Based on a set of explicit and latent features extracted from the textual information, Fake Detector builds a deep diffusive network model to learn the representations of news articles, creators and subjects simultaneously. Extensive experiments have been done on a real-world fake news dataset to compare FAKEDETECTOR with several state-of-the-art models, and the experimental results have demonstrated the effectiveness of the proposed model. [3]

**4. METHODOLOGY:**

**4.1 PREPROCESSING**

**4.1.1 Pre-Processing -1**

1. Tokenization

Tokenization is the process of turning sensitive data into no sensitive data called "tokens" that can be used in a database or internal system without bringing it into scope.

1. Lemmatization

Lemmatization is a linguistic term that means grouping together words with the same root or lemma but with different inflections or derivatives of meaning so they can be analyzed as one item

1. Lowercase
2. URL Remover

URL Remover is removed by using regular expressions.

1. Spell Correction
2. Special Characters Removal

Special characters is removed by using regular expressions.

1. Changing the labels from categorical values to binary
2. Word embedding

a. One-hot representation

**4.1.2 Pre-Processing -2**

We implemented tokenization, lemmatization, lowercase, URLremoval, Special Characters Removal, Changing the labels from categorical values to binary.

1. Word embedding

a.Word2vec.

Word2vec is a two-layer neural net that processes text by “vectorizing” words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep neural networks can understand.

**4.1.3 DATA REDUCTION:** TF-IDF

The models are not built to handle the categorical data so we converted all the text data i.e. the tweets into vector form.

**4.1.4 DATA TRANSFORMATION:**

1. “fake” is converted to 0
2. “real” is converted to 1

As the model cannot take input in categorical form, we need to convert it in numeric form.

**4.1.5 FEATURE SELECTION**

1. "id" is removed as it’s an redundant attribute.
2. "label" and "tweets" are kept as they are the most important attributes for the determination of fake news.

**4.2 LEARNING ALGORITHM**

**4.2.1 ALGORITHM:** **CONVOLUTION NUERAL NETWORK (CNN)**

Convolutional neural network is a classical deep learning technique commonly used in 2D but can be used as text classifier.

**4.2.2 ALGORITHM: LONG SHORT TERM MEMORY(LSTM)**

Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequences of data.

**4.3 EVALUATION OF MODEL BUILT**

|  |  |
| --- | --- |
| Model | F1 score |
| CNN (one hot) | 0.9228 |
| LSTM (one hot) | 0.901 |
| CNN (word2vec) | 0.93 |
| LSTM (word2vec) | 0.928 |

Table.1Evaluation table of 4 models

**5. RESULTS:**

**5.1 DATASET**

**Id:** For every Tweet there is one unique id assigned

**Tweet:** It is a categorical type of data which consists of news in the form of tweets, WhatsApp messages, news headlines, Instagram posts etc.

**Label:** It is a categorical type of data in binary form i.e. fake or real. This indicates whether the tweet attribute news is real or fake.

**Files:**

Constraint\_English\_Train - Sheet1.csv

Constraint\_English\_Validation - Sheet1.csv

Constraint\_English\_Test - Sheet1.csv

**5.2 PERFORMANCE EVALUATION METRICS**

**F1-SCORE**

F1-Score is the harmonic mean of precision and recall values for a classification problem. The formula for F1-Score is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | recall | F1 score |
| CNN (one hot) | 92.28% | 0.9228 | 0.9228 | 0.9228 |
| LSTM (one hot) | 90.1% | 0.901 | 0.901 | 0.901 |
| CNN (word2vec) | 93.40% | 0.934 | 0.9340 | 0.9340 |
| LSTM (word2vec) | 92.8% | 0.928 | 0.928 | 0.928 |

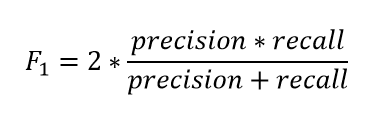


Table.2 performance evaluation matrix of 4 models

**5.3 EXPLORATIVE DATA ANALYSIS(EDA)**

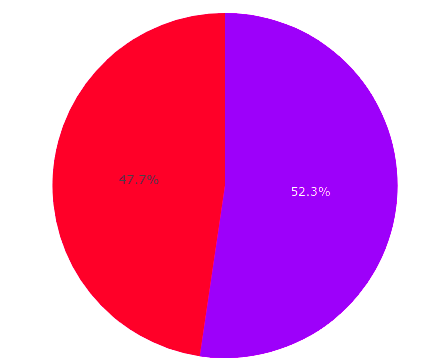


Figure2. Pie Chart showing the number of real news(purple) and number of fake news(red)

The dataset is balanced having 3360 real news and 3060 fake news.

**Number of characters**

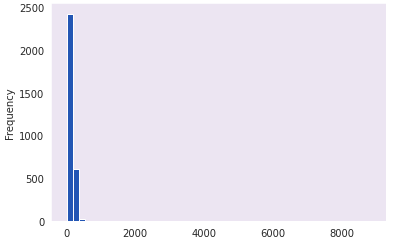


Figure 3. Number of characters in fake news

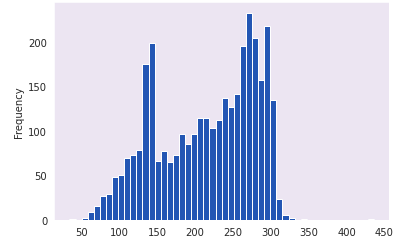


Figure 4 .Number of characters in real news

**Number of words**

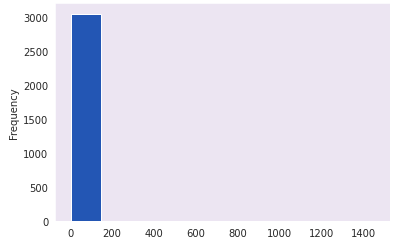


Figure 5. Number of words in fake news

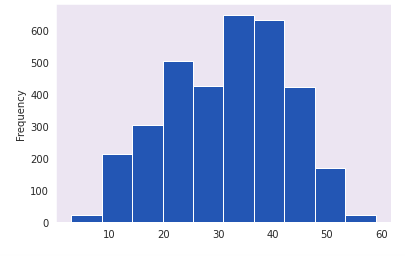


Figure 6. Number of words in real news

Insight: The average number of characters and words are higher in real news than fake news as the fake news are taken from some article without context therefore the number of words and characters are less.

**Number of unique words**

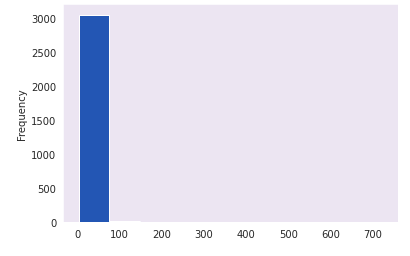


Figure 7. Number of Unique words in fake news

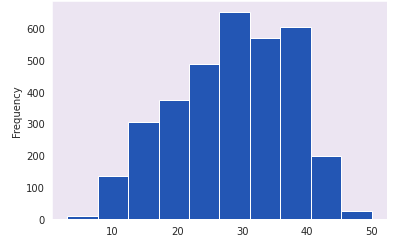


Figure 8. Number of Unique words in real news

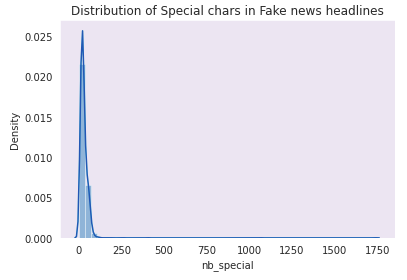


Figure 9. Number of Special character in fake news

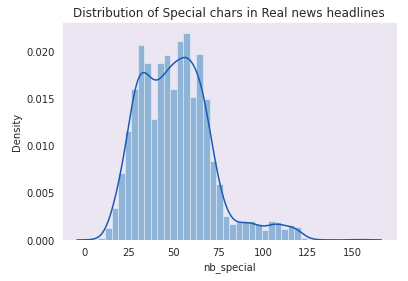


Figure 10. Number of Special characters in real news

Insight: Real news have more special characters than fake news therefore we can say that the use of special characters are more in real news than in fake news.

Word Cloud:

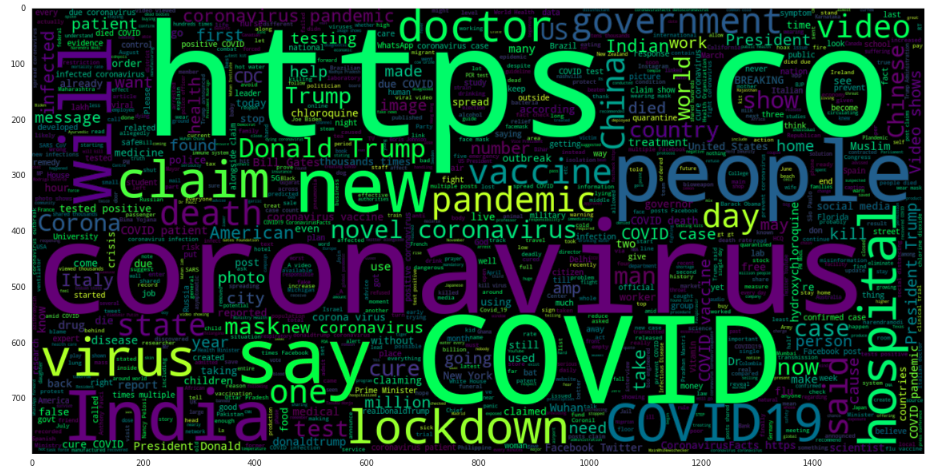


Figure 11.Most frequent words in fake news

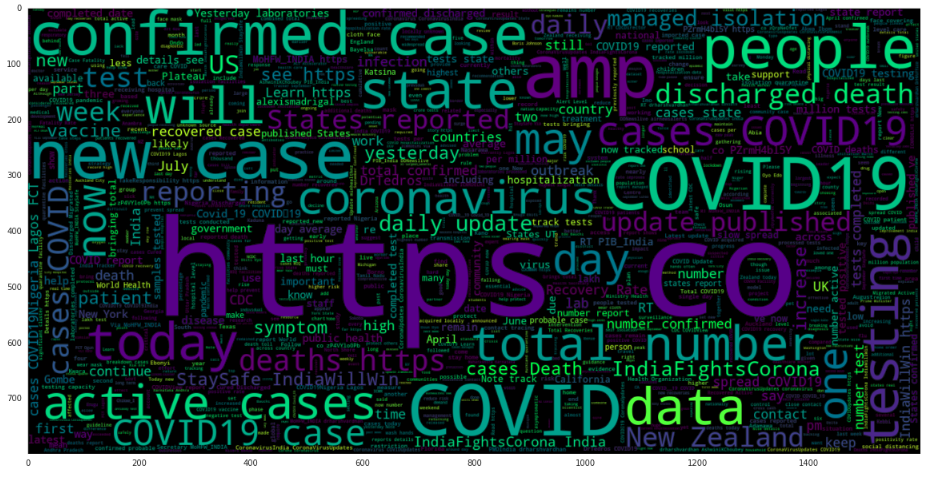


Figure12. Most frequent words in real news

Insight: Most frequent words of the real news and fakes are almost the same.

**5.4 RESULT**

After preprocessing 1

|  |  |
| --- | --- |
| MODEL | ACCURACY |
| CNN | 93.04% |
| LSTM | 92.62% |

Table 2. Results using pre processing1

After preprocessing 2

|  |  |
| --- | --- |
| MODEL | ACCURACY |
| CNN | 94.02% |
| LSTM | 93.16% |

Table 3. Results using pre processing2

**5.5 Bias-Variance Graphs**

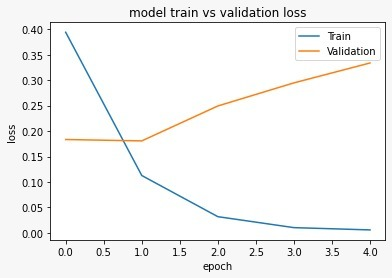


Figure 13. Bias-Variance Tradeoff for LSTM with one hot encoding

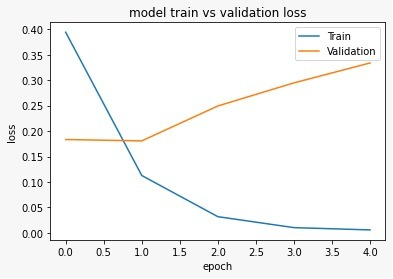


Figure 14. Bias-Variance Tradeoff for CNN with one hot encoding

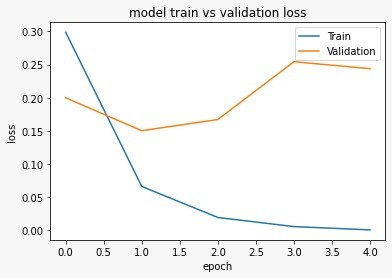


Figure 15. Bias-Variance Tradeoff for LSTM with Word2Vec

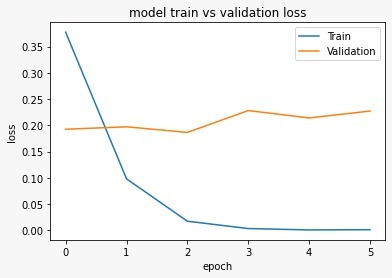


Figure 16. Bias-Variance Tradeoff for CNN with Word2Vec

Out of the four models for every model the training loss reduces as the number of epoch increases but the best model chosen i.e. CNN with word2vec has the validation loss lower than the other 3 difference models therefore CNN with word2vec has the lowest Total error.

**6 Conclusion**

Using the dataset provided, we created a model using word2Vec and CNN for predicting fake or real tweet from Covid-19 dataset. The model uses several preprocessing techniques, feature extraction using word2Vec and prediction using CNN. Our model gives an accuracy of 94%.

**Reference**

1. Ahmad, Iftikhar, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. "Fake News Detection Using Machine Learning Ensemble Methods." *Complexity* 2020 (2020).
2. Aldwairi, Monther, and Ali Alwahedi. "Detecting fake news in social media networks." *Procedia Computer Science* 141 (2018): 215-222.
3. Zhang, Jiawei, Bowen Dong, and S. Yu Philip. "Fakedetector: Effective fake news detection with deep diffusive neural network." In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1826-1829. IEEE, 2020.
4. Researchgate.net 2020:<https://www.researchgate.net/publication/341287001_Language-Independent_Fake_News_Detection_English_Portuguese_and_Spanish_Mutual_Features> .
5. Towardsdatascience 2018: <https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f> .
6. data-science-group-iitr 2017: <https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285>.
7. Machinelearningmastery 2017: <https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/>.
8. Towardsdatascience 2020: <https://towardsdatascience.com/text-normalization-7ecc8e084e31>.