JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Fake News Prediction using Deep Learning Techniques

CH.SAI NIKHIL, Department of Computer Science and Engineering, AVIT, Paiyanoor. nikhil.csn09@gmail.com

ANAGURTHI KUNAL, Department of Computer Science and Engineering, AVIT, Paiyanoor. kunalanagurthi0999@gmail.com

JOHN SANDESH NETHALA, Department of Computer Science and Engineering, AVIT, Paiyanoor. johnsandeshnethala@gmail.com

S.LEELAVATHY, Assistant Professor GR II, Department of Computer Science and Engineering, AVIT, Paiyanoor

ABSTRACT:

There has been a rapid increase in the spread of fake news in the last decade, most prominently observed in the 2016 US elections. Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science. One such area affected by fake news is the financial markets, where a rumour can have disastrous consequences and may bring the market to a halt. This project is about designing a fake news predictor using the deep learning techniques in which we will be classifying the original and the fake news of same topic by applying different techniques, there is fake news in every industry in the society which mis leads the people and creates confusion in the society. Due to the advancement of technology the fake news is circulated very quickly which creates panic in the public, the deep learning model we create will identify original and fake news so that the people will not fall into the trap of fake news.

Keywords: Online fake news, Machine learning, fake news, Text Classification, social media.

1. INTRODUCTION:

Fake news is a term which is used to fabricate news or news which is not true. Fake news is circulated through traditional media like news and newspapers and un-traditional social media platforms like twitter, Facebook, Instagram, WhatsApp etc. generally fake news is circulated to damage the reputation of any organization or institution or lack of knowledge on the subject. Fake news is increasingly becoming a menace to our society. It is typically generated for commercial interests—to attract viewers and collect advertising revenue. However, people and groups with potentially malicious agendas have been known to initiate fake news in order to influence events and policies around the world. Whatever may be the reason the fake news is considered as one of the threats to democracy. The root cause of this problem lies in the fact that none of the social networking sites use any automatic system that can identify the veracity of news flowing across these platforms. A possible reason for this failure is the open domain nature of the problem that adds to the intricacies. The recently organized Fake News Challenge is an initiative in this direction. The aim of this challenge is to build an automatic system that has the capability to identify whether a news article is fake or not. Due to the advancement of the technology and availability of the internet the fake news is circulated faster than the genuine news in social media platforms. Research conducted on velocity of fake news spread it is concluded that the tweets containing the fake or misinformation reached to people six times faster than the truthful tweets. The adverse effects of inaccurate news range from making people believe that Hillary Clinton had an alien baby, trying to convince readers that President Trump is trying to abolish first amendment to mob killings in India due to a false rumor propagated in WhatsApp. Technologies such as Artificial Intelligence (AI) and Natural Language Processing (NLP) tools offer great promise for researchers to build systems which could automatically detect fake news. However, detecting fake news is a challenging task to accomplish as it requires models to summarize the news and compare it to the actual news in order to classify it as fake. Moreover, the task of comparing proposed news with the original news itself is a daunting task as its highly subjective and opinionated. Technologies like Natural language processing and deep learning help us to detect the fake news..

2. LITERATURE REVIEW:

Fake news is increasingly becoming a menace to our society. It is typically generated for commercial interests—to attract viewers and collect advertising revenue. However, people and groups with potentially malicious agendas have been known to initiate fake news in order to influence events and policies around the world. It is also believed that circulation of fake news had material impact on the outcome of the 2016 US Presidential Election [1]. From an NLP perspective, this phenomenon offers an interesting and valuable opportunity to identify patterns that can be coded in a classifier. In our experiments, we will be ignoring all other signals (e.g., the source of the news, whether it was reported online or in print, etc.), and instead focus only the content matter being reported.

With COVID-19 emerging as a pandemic that has affected everyone worldwide, people have become more reliant on news to make everyday decisions to ensure their safety of themselves and their loved ones. However, fake news is almost becoming a "second pandemic" or "infodemic," posing as a health hazard to people worldwide. Given that coronavirus-related fake news is such a new phenomenon, prior work has not applied fake news detection to coronavirus. In an effort to tackle this issue, we utilize a modified LSTM that considers features relevant to fake news including the Jaccard index between the title and text, polarity, and frequency of adjective use. Our model was trained on a 600 article dataset containing 300 fake news articles and 300 real news articles. It achieved an overall accuracy of 0.91 with F1 scores of 0.89 and 0.92 for real and fake news, respectively.

Internet is one of the important inventions and a large number of persons are its users. These persons use this for different purposes. There are different social media platforms that are accessible to these users. Any user can make a post or spread the news through these online platforms. These platforms do not verify the users or their posts. So some of the users try to spread fake news through these platforms. These fake news can be a propaganda against an individual, society, organization or political party. A human being is unable to detect all these fake news. So there is a need for machine learning classifiers that can detect this fake news automatically. Use of machine learning classifiers for detecting the fake news is described in this systematic literature review

3. SYSTEM ANALYSIS:

3.1 EXISTING SYSTEMS:

Many models like logistic regression, RNN, GRU, LSTM AND BILSTM are used, 300 dimensions pretrained glove is used for vectorization and GRU got highest value in both precision and F1 score with 0.79 and 0.84. The disadvantage is learning rate is high for the GRU models.[1] Hybrid CNN-RNN deep learning approach. CNN is used for feature extraction from the text and the RNN is used for classification the accuracy results are high for only few data sets. In IOST data set with hybrid model achieved 0.99 precision, accuracy and F1 score.[2] used FNC1 data set which has 4 categories unrelated, related, agree, disagree. Used cosine similarity and deep neural networks for classification. Many models were tried out of which model with TFIDF vectorization and cosine similarity achieved accuracy of 94.31%. The disadvantage is the data set is unbalanced.[3] Used 4 different data sets, existing ensemble learners are used with IOST data set achieved highest accuracy of 99% with random forest and perze-LSVM average accuracy on data set is 97.67 disadvantage is performance varies for different data sets.[4] Used FNC-1 data set and skip through vectors for the sentences and n-grams with TF vectors used for headings and text and attention mechanism is used, highest score achieved is 82.05 and disadvantage is being a small data set and un even distribution led to poor performance of the model.[5] Model is a blend of CNN and RNN. Used pretrained glove for vectorization, using CNN helped to extract more features in less time and LSTM is used for classification achieved accuracy of 99.5% and use of Glove instead of traditional bag of words made huge difference.[6] Used TI-CNN data set and 3 different models CNN, LSTM and BERT are used. Word2vec is used for vectorization and hyper parameter tuning is performed and SGD optimizer used for CNN and LSTM, Adam optimizer for BERT. Accuracy 0.97, 0.98, 0.98 achieved for CNN, LSTM and BERT. Loss function is binary cross entropy.[7]

For data set 1356 news instances were collected form users in twitter and other social media. Models like CNN, LSTM were used out of this CNN with bidirectional LSTM and attention mechanism achieved highest accuracy of 88.33%.[8].

3.2 PROPOSED SYSTEM:

Collected the data set from Kaggle which have 18000 records of two different classes fake and original represented as 0 and 1 in the data set, loaded the data set into the google colab and removed all the null values present in the data set and re-indexed it and divided the data into independent and the dependent features x as independent feature and y as depended feature. In data pre-processing vectorization is done used one hot representation, TFIDF and count vectorizer as vectorization techniques for four different models implemented in the project and used padding for LSTM and GRU before embedding into the algorithms and imported NLTK library and downloaded stop words and removed all the stop words in LSTM and GRU used stratified k-fold as a training method for LSTM and 70-30 split technique for GRU, multinomial naïve bayes and passive aggressive classifier. For training LSTM and GRU used 10 epochs and twitter data is used for testing the LSTM model.

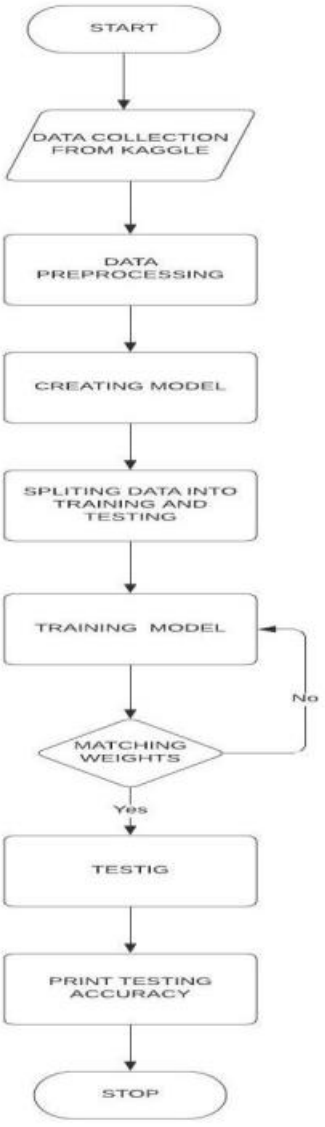
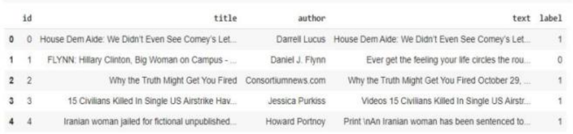


FIG 1 flow chart

DATA SET:

Data is collected from Kaggle fake news detection competition. Data set consists of testing and training in training there are 18000 records and in testing there are 7000 records in data set there are two class fake and original classes represented as 0 and 1 labels the data set has 4 features title, text, Author and label.



	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

FIG 2 data set

4. SYSTEM DESIGN:

4.1 METHODOLOGY:

The collected data is cleaned by removing the null values and re indexing the data. In data preprocessing the data is further cleaned by removing the stop words and apply stemming and lemmatization techniques which convert the similar words into the root word. After data preprocessing the text data is converted into number vectors because machine learning or deep learning algorithms cannot understand the text information this process is called vectorization. vectorization is the important step in any natural language processing it even decides the performance of the model there are many techniques to this like bag of words, TF-IDF, word2vec etc but we used one hot representation because it is easy and fast. We also use the pad sequencing to convert the text information in the data set to same length so that it is easy for the algorithm for processing. We will be selecting the number of vectors features so that our word corpus will be categorised into these features. After this the data is passed into the algorithm here, we used the LSTM an RNN algorithm in deep learning. We used hyper parameters and optimizers like sigmoid and Adam and loss functions like binary cross entropy and accuracy as metrics.

4.2 ARCHITECTURE:

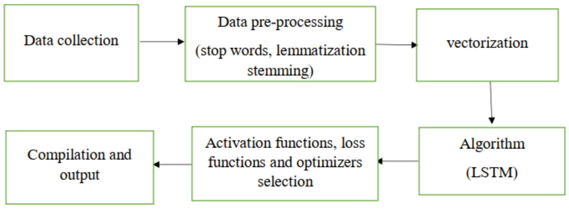

Data collection
Data pre-processing
(stop words, lemmatization stemming)
vectorization
Algorithm
(LSTM)
Activation functions, loss functions and optimizers selection
Compilation and output

FIG 3: ARCHITECTURE DIAGRAM

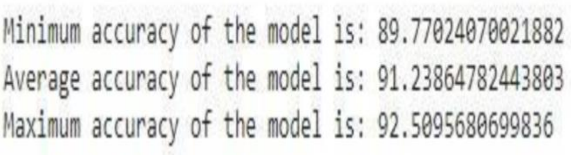
The above architecture diagram shows the different steps involved in developing the fake news detection model. Different steps shown in the architecture diagram is explained in the next section of this report.

5. IMPLEMENTATION:

Import all the required modules and import the data set and remove all the null values from it and reindex the data set select the vocabulary size of 5000. Import the nltk library for data preprocessing from nltk module download the stop words English library, from nltk import the porter stemmer (It is a type of stemming technique) for stemming process, using the regular expression remove all the unwanted details or symbols other than alphabets from a-z or A-Z, convert all the text available in to lowercase and apply stemming and remove stop words and store the pre-processed text in a list. For vectorization we used one hot representation with 5000 vocabulary size, to make the text in the data set of equal size we used pad sequencing and converted the text into equal length of 20 to make text to equal size zeroes are added at the end or at the back, so that it will be easy for computing and we selected 40 features for our model. We used sequential model in this project and imported all the required modules like tensor flow, NumPy and Sklearn from tensor flow we imported the drop out regularization technique and added 100 LSTM and GRU layers we created a dense layer with sigmoid as activation function and binary cross entropy as loss function and Adam as optimizer and accuracy as performance metrics. We used 70% data for training and 30% data for testing and selected random state as 42 with 64 as batch size for running model for 10 epochs we achieved accuracy of 99.54 and validation accuracy of 91.12 for GRU and used stratified k-fold for training and testing of LSTM model and achieved maximum accuracy of 93 and minimum accuracy of 90.

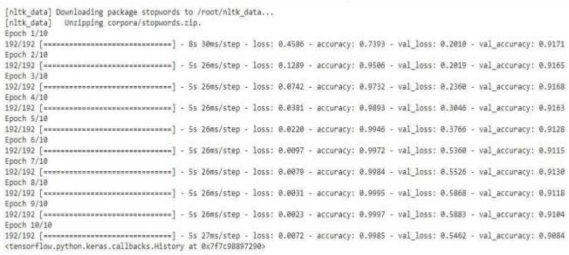
Used supervised naïve bayes algorithm with count vectorization and TF-IDF vectorization for vectorization techniques and used MultinomialNB from naive bayes Passive Aggressive Classifier from linear model. Achieved 90% and 92% accuracy in naïve bayes and linear model with count vectorization and achieved 88% and 91% accuracy in naïve bayes and linear model using TF-IDF vectorization..

6. RESULTS:



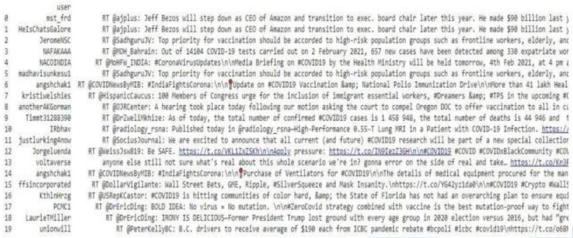
```
Minimum accuracy of the model is: 89.77024070021882
Average accuracy of the model is: 91.23864782443803
Maximum accuracy of the model is: 92.5095680699836
```

FIG 4 (LSTM Output)



```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
Epoch 1/10
192/192 [==============================] - 8s 30ms/step - loss: 0.4586 - accuracy: 0.7393 - val_loss: 0.2010 - val_accuracy: 0.9171
Epoch 2/10
192/192 [==============================] - 5s 26ms/step - loss: 0.1289 - accuracy: 0.9506 - val_loss: 0.2019 - val_accuracy: 0.9165
Epoch 3/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0742 - accuracy: 0.9732 - val_loss: 0.2360 - val_accuracy: 0.9168
Epoch 4/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0381 - accuracy: 0.9893 - val_loss: 0.3046 - val_accuracy: 0.9163
Epoch 5/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0220 - accuracy: 0.9946 - val_loss: 0.3766 - val_accuracy: 0.9128
Epoch 6/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0097 - accuracy: 0.9972 - val_loss: 0.5360 - val_accuracy: 0.9115
Epoch 7/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0079 - accuracy: 0.9984 - val_loss: 0.5526 - val_accuracy: 0.9130
Epoch 8/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0031 - accuracy: 0.9995 - val_loss: 0.5868 - val_accuracy: 0.9118
Epoch 9/10
192/192 [==============================] - 5s 26ms/step - loss: 0.0023 - accuracy: 0.9997 - val_loss: 0.5883 - val_accuracy: 0.9104
Epoch 10/10
192/192 [==============================] - 5s 27ms/step - loss: 0.0072 - accuracy: 0.9985 - val_loss: 0.5462 - val_accuracy: 0.9084
<tensorflow.python.keras.callbacks.History at 0x7f7c98897290>
```

FIG 5 (GRU Output)



```
               user
0           mst_frd        RT @ajplus: Jeff Bezos will step down as CEO of Amazon and transition to exec. board chair later this year. He made $90 billion last y
1    HeIsChatsGalore        RT @ajplus: Jeff Bezos will step down as CEO of Amazon and transition to exec. board chair later this year. He made $90 billion last y
2          JeromeNSC        RT @SadhguruJV: Top priority for vaccination should be accorded to high-risk population groups such as frontline workers, elderly, anc
3           NAFAKAAA        RT @MOH_Bahrain: Out of 14104 COVID-19 tests carried out on 2 February 2021, 657 new cases have been detected among 330 expatriate wor
4          NACOINDIA      RT @MoHFW_INDIA: #CoronaVirusUpdates\n\nMedia Briefing on #COVID19 by the Health Ministry will be held tomorrow, 4th Feb 2021, at 4 pm a
5    madhavisunkesu1        RT @SadhguruJV: Top priority for vaccination should be accorded to high-risk population groups such as frontline workers, elderly, anc
6         angshchak1 RT @COVIDNewsByMIB: #IndiaFightsCorona:\n\n📍Update on #COVID19 Vaccination &amp; National Polio Immunization Drive\n\nMore than 41 lakh Heal
7     kristiwelshies     RT @HispanicCaucus: 100 Members of Congress urge for the inclusion of immigrant essential workers, #Dreamers &amp; #TPS in the upcoming #C
8     anotherAKGorman        RT @DJRCenter: A hearing took place today following our motion asking the court to compel Oregon DOC to offer vaccination to all in cu
9       Timmt31288390        RT @DrZweliMkhize: As of today, the total number of confirmed #COVID19 cases is 1 458 948, the total number of deaths is 44 946 and  t
10            IRbhav        RT @radiology_rsna: Published today in @radiology_rsna—High-Performance 0.55-T Lung MRI in a Patient with COVID-19 Infection. https://
11   justlurking4now        RT @SociusJournal: We are excited to announce that all current (and future) #COVID19 research will be part of a new special collectior
12       Jorgeluenda     RT @WeissJsw819: Be SAFE. https://t.co/VKL1ZxZSKh\n\nApply pressure: https://t.co/IN9IezZ3GH\n\n#COVID19 #COVID #COVIDxBlackCommunity #COV
13        voltaverse        anyone else still not sure what's real about this whole scenario we're in? gonna error on the side of real and take… https://t.co/Kn3F
14        angshchak1     RT @COVIDNewsByMIB: #IndiaFightsCorona:\n\n📍Purchase of Ventilators for #COVID19\n\nThe details of medical equipment procured for the man
15    ffsincorporated     RT @DollarVigilante: Wall Street Bets, GME, Ripple, #SilverSqueeze and Mask Insanity.\nhttps://t.co/YG42yz1da0\n\n#COVID19 #Crypto #WallS
16          KthlnHrzg     RT @USRepKCastor: #COVID19 is hitting communities of color hard, &amp; the State of Florida has not had an overarching plan to ensure equi
17             PCMC1      RT @DrEricDing: BOLD IDEA: No virus = No mutation. \n\n#ZeroCovid strategy combined with vaccine is the best mutation-proof way to fight
18      LaurieTMiller        RT @DrEricDing: IRONY IS DELICIOUS—Former President Trump lost ground with every age group in 2020 election versus 2016, but had "gre
19         unionwill           RT @PeterKellyBC: B.C. drivers to receive average of $190 each from ICBC pandemic rebate #bcpoli #icbc #covid19\nhttps://t.co/o68M
```

FIG 6 (Twitter Data)

```
accuracy:   0.902
Confusion matrix, without normalization
```

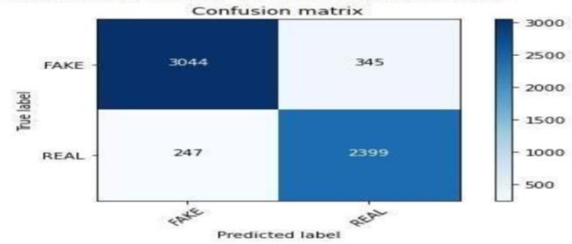


FIG 7 (Naïve Bayes Count Vectorization)

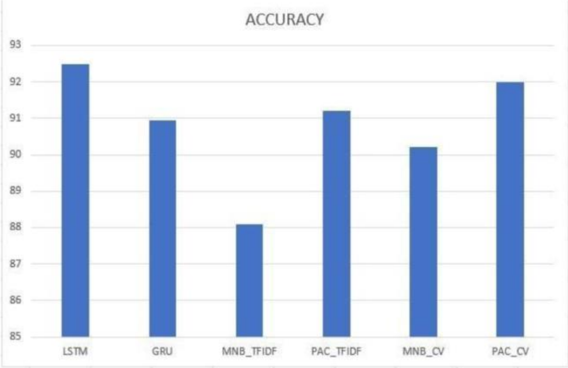


FIG 8 (PERFORMANCE ANALYSIS)

7. CONCLUSION:

Performing all the pre-processing techniques like removal of null values and stop words and performing stemming techniques and with fixed vocabulary size of 5000, using one-hot representation and adding embedding layer to get sentences of equal size and 40 vector features and sigmoid as a hyper parameter and binary cross entropy, Adam as loss function and optimizer our fake news detection model using LSTM algorithm achieved accuracy of 99.35% and validation accuracy of 91.12%. In future we will be collecting the data set from twitter and analysing the tweets and classifying them into fake or genuine tweet with the model we designed above and more different like CNN, GRU etc models will be implemented and model with good performance and high accuracy will be selected. Will try other hyper parameters and optimizers and reduce the over fitting and improve the overall performance of the model and use attention mechanism to improve the accuracy of the model. Despite the relative abundance of extant works addressing fake news detection, there is still plenty of space for experimentation, and the discovery of new insights on the nature of fake news may lead to more efficient and accurate models. In addition, this paper is, to the best of my knowledge. we hope to get one step closer towards building an automated fake news detection platform. This study provides a baseline for the future tests and broadens scope of the solutions dealing with fake news detection. I would like to further dig deep and evaluate the effects of such news propagation on the readers and come up with simple techniques for faster prediction. The research can borrow qualitative models built on similar tasks by other disciplines and reevaluate feature engineering and pre-processing techniques used.

FUTURE SCOPE:

The system will be implemented with different models and different training techniques will be used to train the model better to achieve the benchmark accuracy.

8. REFERENCES:

[1] Samir Bajaj, Stanford university, fake news detection using deep learning.

[2] Jamal Abdul Nasir, Osama Subhani Khan, Iraklis Varlamis CNN-RNN approach for fake news detection.

[3] Aswin Thota, Priyanka Tilak, Simrat Ahluwalia, Nibrat Lohia, a deep learning approach to fake news detection southern Methodist university.

[4] Ftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais, fake news detection using machine learning ensemble approach.

[5] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal, On the Benefit of Combining Neural, Statistical and External Features for Fake News Identification, Indian Institute of Technology, Roorkee, India.

[6] Aman Agarwal, Mamata Mittel, Akashat Pathak, Lalit Mohan Goyal, Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning.

[7] Álvaro Ibrain Rodríguez, Lara Lloret Iglesias fake news detection using deep learning university of cantabria spain.

[8] Sachin Kumar, Rohan Astana, Shashwat Upadhyay, Nidhi Upreit, Mohammad Akbar fake news detection using deep learning a novel approach.

[9] Sanket Doshi various optimization algorithms for training neural networks.

[10] Shiva Varma understanding different loss functions for neural networks..