International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Fake News Detection System

Aditi Raut[1], Aleena Marium[2], Ruchika Navandar[3], Shraddha Chitte[4], Harsha sonune[5], Kedarnath Dixit[6]

[1]*UG Student, Department of Information Technology, Cummins College of Engineering for Women, Savitribai Phule University, India.*
[2]*UG Student, Department of Information Technology, Cummins College of Engineering for Women, Savitribai Phule University, India.*
[3]*UG Student, Department of Information Technology, Cummins College of Engineering for Women, Savitribai Phule University, India.*
[4]*UG Student, Department of Information Technology, Cummins College of Engineering for Women, Savitribai Phule University, India.*
[5]*Assistant Professor, Department of Information Technology, Cummins College of Engineering for Women, Savitribai Phule University, India.*
[6]*Engineering Specialist at Persistent Systems, Pune, Maharashtra, India.*

Abstract - *Fake news detection is very difficult while its spread is simple and has vast effects . To tackle this problem we propose a model which detects fake information and news with the help of Deep Learning and Natural Language Processing. A Deep Neural Network on a data set is trained and by using Natural Language Processing the correlation of words in respective documents is found and these correlations serve as initial weights for the deep neural network which predicts a binary label to detect whether the news is fake or not. In this work we have successfully used RNN and Long Short-Term Memories to test for classification. Tensorflow is used for implementation of the proposed framework and provides visualizations for the neural network. Confusion matrix and classification reports show that we can achieve an accuracy score of 99% by using Long Short-Term Memories and Recurrent Neural Network respectively .*
Also, to check the headline or fact of the news, we have created a module to check for contradiction between inputted news and web scraped news related to it.

Key Words: **Fake News, Machine Learning, Deep Learning, Web Scraping, Neural Network, RNN, LSTM, Adam Optimizer, Beautiful Soup, Spacy**

1. INTRODUCTION

Fake news is a false piece of information. In this day and age, with the increment in spread of phony news from web-based media and different sources it is getting vital to have the option to classify between genuine news and phony news. Fake news is a main consideration in actuating riots, mob lynching and other social-monetary aggravations. Any piece of Fake news can be made to intentionally mislead or delude a person, advance a one-sided perspective, specific reason or plan for the entertainment. It may also hamper one's health if the fake news is about how to cure a disease. Variety of misinformation is being spread about the same[20]. Fake news can be promoted by unauthenticated person, web-based media, printing of phony news in papers perhaps because of political pressing factor and some more. The severity of the problem has increased substantially as 2019 was declared as the year of fake news[21]. News and media inclusion gets enormously misshaped because of the introduction and spread of phony news. Where news can be a shelter, counterfeit news is a plague to the general public. In any case, the qualification between certifiable news and a fake one is troublesome. Its examination consequently assumes a vital part in the present standing.

2. CURRENT STATE OF ART

There exist a couple of models for identification of Fake News [12] – [17]. In the paper, by Bajaj [14], the creator moves toward the issue from an absolutely NLP point of view. An examination from numerous various models execution is done and an investigation is introduced. A Convolution Network is examined with another plan that joins an "consideration like" component and a few designs are investigated. The creator analyzed models of Logistic Regression, Two-layer Feed forward Neural Network, Recurrent Network, Long Short-Term Memories, Grated Recurrent Units, Bidirectional RNN with LSTMs, Convolution Neural Network (CNN) with Max Pooling, Attention-Augmented CNN and the perceptions and results are noted. As indicated by creator's perception, the RNN design with GRUs beat one with LSTM cells. This holds in spite of the way that a positive inclination was added to the LSTM's neglect entryway.

Author Gilda [16] applied Term Frequency-Inverse Document Frequency, a strategy ordinarily applied in Document examination. This technique comprised of bi-grams and probabilistic setting free language (PCFG) recognition to an assortment of practically 11000+ stories. He tried dataset on different classifiers also including,

Random Forest, Support Vector Machines, Bounded Decision, Stochastic Gradient Descent, and others. His use of Term Frequency-Inverse Document Frequency of bi-gram, contribution to Stochastic Gradient Descent classifier model decided phony source with practically 77% exactness. In the paper, [17], creators Singhania, Fernandex and Rao, introduced a computerized identifier utilizing profound learning techniques were utilized. It had a three level progressive consideration organization (3HAN). This brought about a quick and very exact discovery of phony news stories. 3HAN was utilized for making a news story with the assistance of 3 vectors, each for words, sentences and features and cycles the information news story in a base up way. The feature, not many words and sentences are recognizing highlights of the articles and are moderately of more significance than the rest to which 3HAN gives the necessary consideration with the assistance of 3 layers. Creators noticed practically 96.5% exactness on a sizeable genuine world dataset. 3HAN gives a justifiable yield through consideration boundary esteems given to different segments of the articles.

3. ALGORITHM DESCRIPTION

3.1 RNN- RECURRENT NEURAL NETWORK

Repetitive Neural Network (RNN) is an assortment of Artificial Neural Network (ANN) wherein the hubs are associated with structure a successive coordinated chart. RNNs can utilize their memory to deal with a successive information. They are amazing and hearty neural organizations. They are powerful and robust neural networks. Also, they are equipped with an internal representation that acts as memory or storage. This allows them to be widely used for ordered dataset such as text, audio, speech, video, etc.

3.2 LSTM- LONG SHORT TERM MEMORY

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. This comprises cells, information, yield, and neglect entryways. Cells are utilized to hold recollections and doors are utilized for controlling the data stream all through these individual cells. Vanishing gradient descent problem can be effortlessly fixed by inclination cutting by restricting the factor. Disappearing slope plunge issue is confronted when yield is created by passing contribution to an enormous number of covered up layers.

4. PROPOSED METHODOLOGY

We have used two different approaches to check the authenticity of the given news and to categorize it into real or fake.

4.1 LINGUISTIC BASED APPROACH

To check news for authenticity, we have used RNN using LSTM. The dataset used is kaggle fake news dataset. Here, we have 17903 fake news and 20826 true news. For pre-processing the data, the famous NLTK toolkit is used. Tasks like stop word removal, tokenization, and padding are performed using NLTK. A sequential RNN Model along with LSTM is used. Binary Cross Entropy is used. The performance was increased by using Adam Optimizer. Here, the model works by focusing on content tone, grammar, popularity and pragmatics.

4.2 FACT BASED APPROACH

This approach is used to check a one liner news headline or news fact. To check the fact, we gather information by web scraping using beautiful soup. The top 3 results of the google search are extracted and then checked for contradiction of the inputted text. To check the contradiction, we have used spacy linguistic features. spaCy is an open-source software library for advanced natural language processing. To find whether the inputted news and extracted news contradict with each other, we check for the dependency of the word with the other words in the sentence. Same process is applied for the other sentences. Along with it, the presence of antonyms are also checked. Depending on the counts of negative dependency and antonyms, the news are found to be fake or true. Also, to check for the scalar facts, use of regex can be done.

5. PROPOSED SYSTEM

We have designed a web application which enables the user to access it anytime anywhere.

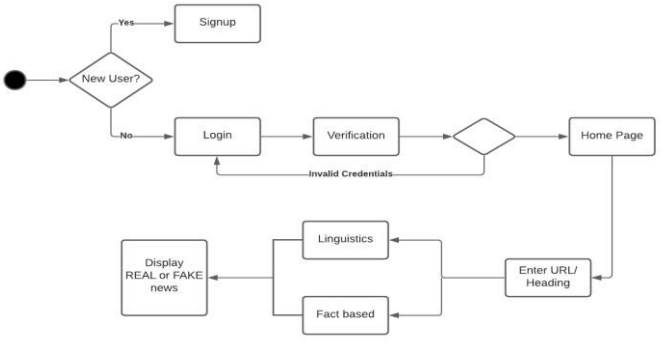

Fig -1: System flow

5.1 MODULE 1- LOGIN AND SIGNUP

For signup the user needs to add all its credentials like name, email address, create a new username and a strong password. The signup information is validated and then stored in the firebase dataset that was created by us. We created a firebase because it's easy to access and provides proper authentication. For the login page the user needs to enter the username and password that they have created during the signup. If he/she is the authenticated user he can then access our website to check whether the news is real or fake.

5.2 MODULE 2- LINGUISTIC CHECKING

In this module the user needs to copy paste the contents of the entire article that he/she is not sure of whether it's fake or real. He has to paste the contents in the text area provided on our web page.
After pasting the contents. The contents are sent to the algorithm and after the classification it then returns the result as real or fake.

5.3 MODULE 3- FACT CHECKING

In this module also we have our textbox into which the user can put into its question or it makes it easy for the user to check any fact. For example "5G network spreads corona virus" this was the fake news which was circulating in our social media sites, so after performing web scrapping on the above news it would be returned real or fake.

6. SOFTWARE TOOLS AND LIBRARIES

6.1 DJANGO

Django is an high level Python Web system that energizes fast turn of events and spotless, sober minded plan. Worked by experienced engineers, it deals with a large part of the problem of Web advancement, so you can zero in on composing your application without expecting to waste time. It's free and open source.

6.2 FIREBASE

It is a platform which was developed by Google for developing web applications. It provides us with variety of tools to develop our application. We used it for authentication and maintaining our database of our users. As its on cloud it can be accessed anywhere at anytime.

6.3 FASTAPI

FastAPI is a high performance web framework and modern way of building API for python 3.6+. Some of its key features are that its fast, it has fewer bugs, it has a great editor support, it minimizes the code duplication and its robust

6.4 PYTHON

Python is a popular programming language which is basically used for server side web development, it can connect to various database systems, it can read and modify files. It can also be used to handle big data and perform complex mathematic operations. We have used this to program the various algorithms that are RNN and LSTM.

6.5 GOOGLE COLAB

Google colab allows you to write and execute your python code on your browser with zero configuration required, free access to all the GPUs, and it enables easy sharing so all the team mates can see the same code at the same time and make changes to it.

6.6 NATURAL LANGUAGE TOOLKIT

NLTK also known as Natural Language Toolkit is a suite of libraries and programs for statistical natural language processing for English written in the python language. It supports classification, tokenization, stemming, tagging and parsing.

6.7 GOOGLE NEWS

Google News is a news aggregator administration created by Google. It presents a constant progression of connections to articles coordinated from a huge number of distributers and magazines. Google News is accessible as an application on Android, iOS, and the Web. Google delivered a beta adaptation in September 2002 and the authority application in January 2006.

6.8 TENSORFLOW

TensorFlow is a free and open-source programming library for machine learning. It tends to be utilized across a scope of assignments however has a specific spotlight on training and testing of profound neural networks. Tensorflow is an emblematic numerical library dependent on dataflow and differentiable programming. It is utilized for both research and production at Google.

6.9 BEAUTIFUL SOUP

It is a python library which deals with pulling out contents from HTML and XML files. It works with parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

6.10 SPACY

spaCy is a free, open-source library for cutting edge Natural Language Processing (NLP) in Python. In case you're working with a great deal of text, you'll in the end need to find out about it. For instance, what's it about? What do the words mean in setting? Who is doing what to whom? What organizations and items are referenced? Which writings are like one another? spaCy is planned explicitly for creation use and assists you with building applications that interaction and "see" huge volumes of text

7. RESULT

7.1 LINGUISTIC APPROACH

For the RNN-LSTM model, 80% of the data was used for training purposes and 20% of data was used for testing purposes. The accuracy of the model came out to be 0.99.For linguistic based approach, for the better result, the content should be large.

7.2 FACT BASED APPROACH

Self made data set was created to check the working of the approach.14 news were checked against the algorithm, out of which 13 came out to be correctly predicted.

News to be checked	Actual	Predicted
5G responsible for coronavirus spread	FALSE	FALSE
Delhi ganga ram hospital, 25 dead	TRUE	TRUE
IPL 2021 has been canceled	FALSE	FALSE
Blobfish voted world's ugliest animal	TRUE	TRUE
Spider man no way home trailer leak	TRUE	TRUE

Table-1: result

8. CONCLUSION

The spread of fake news can adversely affect our surroundings. Henceforth, to distinguish counterfeit news we have proposed a computational model. An enormous dataset from kaggle was used and methods such as stop word removal, lemmatization, tokenization was used. We have utilized RNN with LSTM units for neural organization. Also to check fact or headline, web scraped news is checked for authenticity.

REFERENCES

[1] Dataset:https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset?select=True.csv
[2] Farzi News URL: http://farzinews.com/,.
[3] Teekhi Mirchi URL: http://teekhimirchi.in/
[4] Germany investigating spread of fake news online. URL: https://www.theguardian.com/world/2017/jan/09/germany investigating-spread-fake-news-online-russia-election.
[5] The Spoof URL:https://www.thespoof.com/,
[6] Fun Funny Khez URL: http://www.funfunnykhez.com/,
[7] Dapaan URL: http://dapaan.com/.
[8] Farzi News URL: http://farzinews.com/,.
[9] The Hindu URL: https://www.thehindu.com/,
[10] The Times of India URL: https://timesofindia.indiatimes.com/,
[11] Rising Kashmir URL: http://www.risingkashmir.com/,
[12] Ruchansky, N., Seo, S., & Liu, Y. (2017, November). CSI: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 797- 806). ACM.
[13] Zhang, J., Cui, L., Fu, Y., & Gouza, F. B. (2018). Fake News Detection with Deep Diffusive Network Model. arXiv preprint arXiv:1805.08751.
[14] Bajaj, S. (2017). The Pope Has a New Baby!. Fake News Detection Using Deep Learning, Stanford.
[15] Trovati, M., Hill, R., & Bessis, N. (2015, November). A Non-genuine Message Detection Method Based on Unstructured Datasets. In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015 10th International Conference on (pp. 597-600). IEEE.
[16] Gilda, S. (2017, December). Evaluating machine learning algorithms for fake news detection. In Research and Development (SCOReD), 2017 IEEE 15th Student Conference on (pp. 110-115). IEEE.
[17] Singhania, S., Fernandez, N., & Rao, S. (2017, November). 3HAN: A Deep Neural Network for Fake News Detection. In International Conference on Neural Information Processing(pp. 572-581). Springer, Cham.
[18] Germany investigating spread of fake news online. URL: https://www.theguardian.com/world/2017/jan/09/germany investigating-spread-fake-news-online-russia-election.
[19] Natural Language Toolkit. URL: https://www.nltk.org/. Accessed on 21st, Jan 2019.
[20] https://economictimes.indiatimes.com/news/international/world-news/twitters-cracking-down-on-coronavirus-misinformation/marking-tweets/slideshow/81290697.cms
[21] https://economictimes.indiatimes.com/news/politics-and-nation/fake-news-still-a-menace-despite-government-crackdown-fact-checkers/articleshow/72895472.cms