FAKE NEWS DETECTION USING NLP

ABSTRACT:

The spreading of fake news has given rise to many problems in society. It is due to its ability to cause a lot of social and national damage with destructive impacts. Sometimes it gets very difficult to know if the news is genuine or fake. Therefore it is very important to detect if the news is fake or not. \"Fake News\" is a term used to represent fabricated news or propaganda comprising misinformation communicated through traditional media channels like print, and television as well as non-traditional media channels like social media. Techniques of NLP and Machine learning can be used to create models which can help to detect fake news. In this paper we have presented six LSTM models using the techniques of NLP and ML. The datasets in comma-separated values format, pertaining to political domain were used in the project. The different attributes like the title and text of the news headline/article were used to perform the fake news detection. The results showed that the proposed solution performs well in terms of providing an output with good accuracy, precision and recall. The performance analysis made between all the models showed that the models which have used GloVe and Word2vec method work better than the models using TF-IDF. Further, a larger dataset for better output and also other factors such as the author ,publisher of the news can be used to determine the credibility of the news. Also, further research can also be done on images, videos, images containing text which can help in improving the models in future.

I. INTRODUCTION:

The fake news has been rapidly increasing in numbers. It is not a new problem but recently it has been on a great rise. According to Wikipedia Fake news is false or misleading information presented as news.[1] Detecting the fake news has been a challenging and a complex task. It is observed that humans have a tendency to believe the misleading information which makes the spreading of fake news even easier. According to reports it is found that human ability to detect deception without special assistance is only 54%[2].

Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted. Recent reports showed that due to the rise of fake news that was being created online it had impacted the US Presidential Elections. Fake news might be created by people or groups who are acting in their own interests or those of third parties.

The creation of misinformation is usually motivated by personal, political, or economic agendas.[3]

Since a lot of time is spent by users on social media and people prefer online means of information it has become difficult to know about the authenticity of the news. People acquire most of the information by these means as it is free and can be accessed from anywhere irrespective of place and time. Since this data can be put out by anyone there is lack of accountability in it which makes it less trustable unlike the traditional methods of gaining information like newspaper or some trusted

source. In this paper, we deal with such fake news detection issue. We have used the techniques of NLP and ML to build the model .We have also compared text vectorization methods and obtained the one which gives a better output.

II. LITERATURE SURVEY:

M. Granik et.al [4] proposed a simple approach for the detection of fake news by using Naive Bayes Classifier. They tested it against a dataset of Facebook news posts. They also made use of the BuzzFeed news dataset. They achieved classification accuracy of approximately 74% on the test set.

Niall J,Conroy et.al [5] designed a basic fake news detector that provides high accuracy for classification tasks. They used the linguistic cues approaches and network analysis approach in it. Both approaches adopt machine learning techniques for training classifiers to suit the analysis. They achieved an accuracy of 72% which could be improved. This could be done if the size of the input feature vector is reduced and also by performing cross-corpus analysis of the classification models.

R. Barua et.al [6] identified if a news article is real or misleading by using an ensemble technique using recurrent neural networks (LSTM and GRU). An android application was also developed for determining the sanctity of a news article. They tested this model on a large dataset which was prepared in their work. The limitation of this method was that it required the article to be of a particular size. It would give wrong predictions if the article was not enough to generate a summary.

B. Bhutani et.al [7] used sentiment as an important feature to improve the accuracy of detecting fake news. They have used 3 different datasets. They used Count vectorizer, Tf-Idf vectorizer along with cosine similarity and Bi-grams ,Tri-grams methods. The methods used to train the model are Naive Bayes and Random forest. They used different performance metrics to evaluate the model. They got an accuracy of 81.6%.

M. Vohra et.al [8] proposed, a rumor detection system which determine the authenticity of an information and classify it as rumor or not a rumor. Data was collected by Twitter API. To generate topics from the preprocessed data, topic modelling was performed via Latent Dirichlet Allocation(LDA).They did web scraping on 4 trusted news website. After scraping these sites for articles the links of these articles are save and displayed in the GUI. These keywords were searched on their selected four news websites and news articles were extracted from the results. If no article was found in all the four sites the new assigned that topic as rumor otherwise if article was found its was assigned as not a rumor.

III. PROBLEM STATEMENT:

Since a lot of time is spent by users on social media and people prefer online means of information it has become difficult to know about the authenticity of the news. People acquire most of the information by these means as it is free and can be accessed from anywhere irrespective of place and time. Since this data can be put out by anyone there is lack of accountability in it which makes it less trustable unlike the traditional methods of gaining information like newspapers or some trusted source.

Fake news is dangerous as it can deceive people easily and create a state of confusion among a community. This can further affect the society badly .The spread of fake news creates rumors circulating around and the victims could be badly impacted.

IV. PROPOSED SOLUTION:

As we have seen that the problem of spreading fake news is a serious issue therefore, there is a need to detect this fake news. The main aim of the project is to obtain a model which will help in detecting if a news article is fake or not. The problem of detecting fake news is a very difficult task and many researchers are trying to obtain a solution to it.

Since there are not many datasets which are available publicly to perform this task. We have considered three different datasets which will be merged together to obtain a master dataset which will help in training the models to find if a news is fake or not.

Firstly, the datasets are collected. The datasets are then merged to obtain a master dataset. T his dataset is then preprocessed. Preprocessing of the datasets include lowering of the data, stop word removal, stemming, tokenization and padding is also performed in order to obtain the same length. The dataset is then split into training data and testing data.

To overcome the problem of detecting fake news this project proposes 6 similar LSTM models which are to be trained and each model will be fed with the different text vectors of news headline and news content. This will help in obtaining a good model which will tell if the news is true or it is fake. In this project we have used six similar LSTM models.

Three text vectorization techniques are used which are GloVe, Word2vec and TF-IDF. The first LSTM model will be fed with the vectors of the title of the news using GloVe. The second model will be fed with the vectors of the content of the news using GloVe. Similarly, two models will be built using the Word2vec technique each for the title of the news and the content of the news respectively. Lastly, the LSTM model will be fed with the text vectors of the title of the news using TF-IDF and another model will be fed with the text vectors of the content of the news using TF-IDF. By doing so we can identify which technique gives better results and identify which model performs well. Lastly, the performance is measured using the performance metrics accuracy, precision and recall.

A. DATASET:

There are very few datasets which are available publicly for the detection of fake news. In this paper we have used three different datasets which are available online. The first dataset ISOT Fake News dataset is obtained from a website[17]. The second data that is used in the project is the Fake News Detection dataset from Kaggle[18]. The third dataset used is the Real and Fake News Dataset which is obtained from Kaggle[19].

B. MERGING THE DATASET:

The first dataset ISOT Fake News dataset is obtained from a website[9]. This dataset was created using data from real world news sources. This dataset consists of two types of articles: fake and real. The dataset consists of two CSV files. First file contains all the news which is true and the second file

contains the news which is fake.

Each article contains the following information: article title, text ,type and the date the article was published on. The second dataset used in the project is the Fake News Detection dataset from Kaggle. This dataset consists of 4 columns which are the URLs of the news source, the Headline of the news, the Body of the news that is the content of the news and the last column contains the Label of the news which tells whether the news is fake or not. Next, the two datasets are merged together to obtain a single dataset. After the merge we obtained a dataset with 10344 records. Finally, we obtain a master dataset by merging the first dataset with the above merged dataset [dataset with 10344 records], hence the final obtained master dataset consists of 54726 records and three columns , Title, text and Class.

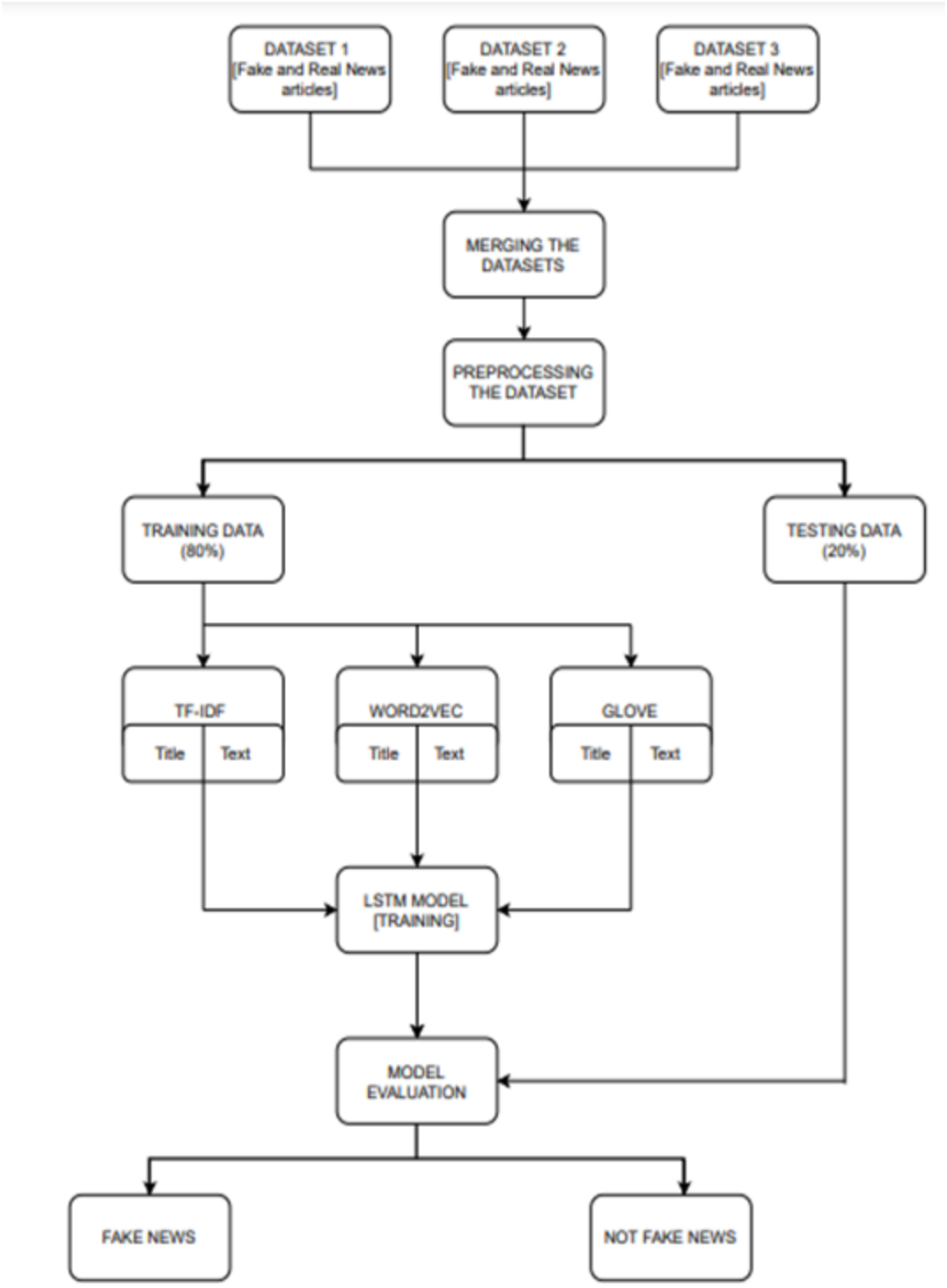

Fig.1 Proposed System Model for Detection of Fake News.

C.PREPROCESSING THE DATASET:

Data preprocessing is a data mining technique that involves transforming raw data into understandable form. In natural language processing, text preprocessing is the practice of cleaning and preparing text data. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing methods such as tokenization, lemmatization, stop word removal and lowercasing.

D. TRAIN-TEST SPLIT:

The next step in the process is to split the data into train and test data. Here, we have done 80% train data and 20% test data split.

E. FEATURE EXTRACTION:

The next step is feature extraction. Machine Learning algorithms learn from a predefined set of features from the training data to produce output for the test data. But the main problem in working with language processing is that machine learning algorithms cannot work on the raw text directly. So, we need some feature extraction techniques to convert text into a matrix(or vector) of features.

Three feature extraction methods will be used

TF-IDF: TF-IDF stands for term frequency-inverse document frequency. It highlights a specific issue which might not be too frequent but holds great importance. The TF–IDF value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. TF-IDF(Term Frequency/Inverse Document Frequency) is one of the most popular IR(Information Retrieval) techniques to analyze how important a word is in a document. TF-IDF is the product of TF and IDF. A high TF-IDF score is obtained by a term that has a high frequency in a document, and low document frequency in the corpus. For a word that appears in almost all documents the IDF value approaches 0, making the tfidf also come closer to 0. TF-IDF value is high when both IDF and TF values are high i.e the word is rare in the whole document but frequent in a document.

WORD2VEC: Word2Vec produces a vector space, typically of several hundred dimensions, with each unique word in the corpus such that words that share common contexts in the corpus are located close to one another in the space. That can be done using 2 different approaches: starting from a single word to predict its context (Skip-gram) or starting from the context to predict a word (Continuous Bag-of-Words). Word2vec is one of the most popular implementations of word embedding, which was invented by Google in 2013. It describes word embedding with two-layer shallow neural networks in order to recognize context meanings. Word2vec is good at grouping similar words and making highly accurate guesses about meaning of words based on contexts. It has two different algorithms inside: CBoW (Continuous Bag-of-Words) and skip gram model.

GLOVE: GloVe, a very powerful word vector learning technique GloVe does not rely just on local statistics (local context information of words), GloVe (Global Vectors for Word Representation) is an alternate method to create word embeddings. It is based on matrix factorization techniques on the word-context matrix. A large matrix of co-occurrence information is constructed and you count each “word” (the rows), and how frequently we see this word in some “context” (the columns) in a large corpus.

F. MODEL:

The model that will be used in this project is the LSTM model. The features extracted from the above feature extraction methods will be given to the LSTM model.

All the pre-processed news titles and content in vector form are given to the LSTM model. We have used the Tensorflow framework to perform this task of detecting fake news. Long Short Term Memory [LSTM] Long short-term memory networks are an extension for recurrent neural networks, which basically extends the memory. The units of an LSTM are used as building units for the layers of a RNN, often called an LSTM network.

LSTMs enable RNNs to remember inputs over a long period of time. This is because LSTMs contain information in a memory, much like the memory of a computer. The LSTM can read, write and delete information from its memory. This memory can be seen as a gated cell, with gated meaning the cell decides whether or not to store or delete information (i.e., if it opens the gates or not), based on the importance it assigns to the information. The assigning of importance happens through weights, which are also learned by the algorithm.

This simply means that it learns over time what information is important and what is not. In an LSTM you have three gates: input, forget and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it isn't important (forget gate), or let it impact the output at the current timestep (output gate).

V. RESULTS:

'Fig.2' shows thee number of fake news and real news in the dataset. We have used word clouds to check which are the words which appear frequently in the fake and real news. 'Fig.3' shows the Word Cloud for real news and Fig. 5.4 shows the Word Cloud for fake news.

and recall. 'Table.I' shows the performance metrics of the models.

Model	Accuracy	Precision	Recall
MODEL 1 [Fed with text vectors of 'Title' obtained by GloVe]	89.71%	91.84%	86.61%
MODEL 2 [Fed with text vectors of 'Text' obtained by GloVe]	92.8%	92.3%	92.9%
MODEL 3 [Fed with text vectors of 'Title' obtained by Word2Vec]	86.49%	87.14%	84.04%
MODEL 4 [Fed with text vectors of 'Text' obtained by Word2Vec]	96.26%	95.40%	96.86%
MODEL 5 [Fed with text vectors of 'Title' obtained by TF-IDF]	80%	70%	70%
MODEL 6 [Fed with text vectors of 'Text' obtained by TF-IDF]	81%	71%	76%

The performance is measured using the accuracy , precision and recall.

Accuracy: It shows the overall accuracy of the instances which are correctly classified to the total number of the instances. It is calculated by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Where, TP = true positive, TN = true negative, FP = false positive, FN = false negative.

Precision: It represents the percentage of relevant sarcastic headlines. That is, it measures the amount of headlines categorized as sarcastic against the total number of headlines classified as sarcastic. It is calculated by the following formula:

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

Recall: It represents the percentage of relevant sarcastic headlines that have been searched. That is, against the total number of sarcastic headlines, measured the number of headlines that are normally classified as sarcastic. It is calculated by the following formula:

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

"Table 1" shows the results obtained by the models. From the results obtained we can observe that the model trained using the content of the news gives better output than the other models. Also, we can see that the models which have used GloVe and **WordVec method work better than the models using TF-IDF.**

CONCLUSION:

Fake news have increased in recent years and it has caused a lot of harm to the society. This research project aimed to develop a model using the techniques of NLP and ML to detect if a news article/headline is fake or not and identify which methods give better output. In this paper, we have presented six LSTM models and three different methods were used for feature extraction. We have used different attributes like the title and text of the news to perform fake news detection. For future work we can work on larger dataset and also future research can be done on images , videos which can help in improving the models. The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are

self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance. Causal Productions permits the distribution and revision of these templates on the condition that Causal Productions is credited in the revised template as follows: “original version of this template was provided by courtesy of Causal Productions .