INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

A Review on Falsified News Detection using Deep Learning Approach

Sudesh Pawar[1], Ghanshyam Patil[2], Ketan Patel[3], Piyush Pawar[4], Shilpa Khedkar[5], Bhagyashree More[6]

[1, 3, 4, 5, 6] *Dept of Computer Engineering, Modern Education Society's College of Engineering, Pune*

Abstract: Technology is changing the way that we live our lives. With the advances in technology Social Media attains a lot of attention of the people around us. With the advancement of the technology the fake news with both political or commercial motive is growing very fast on internet. Users might be affected easily by falsified news, which has brought about mammoth effects on humanity. The foremost goal is to stop the spread of rumours and focus on the true, substantiate news columns. This paper focuses on techniques such as training of deep learning neural network models and natural language processing (NLP) techniques for text analytics based on news title or news content. These techniques can be used for the treatment of the data. Some of the treating steps are tokenizing the input, lemmatizing it ,Stemming and removal of irrelevant words (stop words) are done prior to converting them into N-gram vectors using a technique known as TF-IDF that is term frequency inverse document frequency. With the help of Machine learning and natural language processing, we will try to aggregate the news and later determine whether the news is real or fake using deep neural networks. Experimental evaluation yields the simplest performance using Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction technique.
*Keywords: Fake News, Social Media, Natural Language Processing, Neural Network, TF-IDF, N-Gram Vectors*Introduction

I. INTRODUCTION

Falsified news, or more commonly known as fake news, is one of the largest menaces today. It is defined as a media item comprising but not limited to news, reports & material that is completely false or partially true. Falsified news is very dangerous to the society and can cause serious trouble if not dealt with.

It is similar to a forest fire and influences many gullible every day. Every day a huge number of articles are generated on the internet and a large percent of these articles are malicious in nature, community may be unknowingly helping spread these sans having a knowledge about this news if the news is legit or false.

In India, there are no direct laws to deal with this menace but multiple laws under IPC, IT Act and News Broadcaster Authority Act curb these up to a large extent. There are a lot of social media applications available like Twitter, Telegram etc which can be used to spread misinformation.

Many Applications such as WhatsApp have taken steps to curb fake news by imposing restrictions on forwarding of messages. Our proposed system will help detect these news articles with reliability and accuracy using deep learning techniques. The system is developed to differentiate between genuine news and malicious news.

In today's world due to the presence of many different social medias the fake news spreads like fire which is quite dangerous to the society. The applications like WhatsApp, hike, signal, telegram are the fastest carrier of the fake news and the fake news on these applications get viral over the world in less time. The proposed system tries to authenticate the news and thus helping to stop the spread of fake news.

II. PROBLEM STATEMENT

Fake news also called as bluff news occupies large sphere of cyber security space today's date. Cyber technology's broad reach and quick outspread contributes to its risks. Publicity through such fake news on internet today has been adopted by institutions as well as individuals for various reasons. emotional news is created and spread through social media to realize intended end. On the other side, it may also require narration of a true fact however being purposely overestimated. This may also include labelling the web-pages with ambiguous title or tag-lines in order to grab attention of readers. Such fake information may lead in committing offences, social unrest, financial frauds upon such misrepresentation, political gain, to extend number of readers, gain revenue related to click, etc. This may also affect the affect the importance of great journalism. The problem is to identify the genuineness of the news and online content and at the same time it is also important to identify the bots involved in spreading rumours.

III. LITERATURE SURVEY

There is a lot of existing work in the domain of fake news detection using ML techniques, however for the scope of this survey we predominantly focused on the researches which specifically used deep learning techniques in the desired application area.

Kong et.al [1] in their article, aim to make a deep learning model for identifying misinformation & apply processing techniques on natural language for textual analytics based on title or content of the news.

In case of natural language processing techniques, various pre-processing techniques on textual data are used before converting them in a vector i.e., into use of TF-IDF to convert into N-Gram Vectors. As per the paper, model fed with content of the news are comparatively slower computationally, however, they are more accurate and have a better F1-Score than the ones fed with just the headline of the news articles.

In the next paper by Wu.et.al [2], the deep learning model uses natural language processing techniques and attention mechanism to strengthen emotional semantic features and obtain deep text semantic features. Major focus has been on establishing credibility of the information obtained through news articles, in this paper.

Elhadad et.al. in their paper [3], focused specifically on hoaxes about the Covid-19, the pandemic which caused a lockdown and therefore giving an incentive for people with malicious intentions to spread misinformation. One can understand how ensemble approach can be used to tackle the problem of fake news as in this research, multiple learning algorithms are combined to develop an ensemble ML classifier that uses voting, which focuses purely on detecting misinformation related to the pandemic and this approach can be generalized and transformed to the broader domain of falsified news detection efficiently.

Umer et.al., in their paper, a composite neural network architecture, that amalgates the potential of both LSTM as well as CNN, along with different reduction approaches are discussed to tackle the issue of falsified news. The non-linear features are provided as input to chi-square as well as PCA which provides better context-specific data for detection of misinformation and hoaxes. After careful evaluation, it is found that the model with PCA performs better than the model using Chi-Square for reduction. [4]

Asr et.al. in their paper, focus on how the challenge to find quality data for training the model without any underfitting or overfitting is tackled. The paper explains how perspective of NLP helps in approaching the problem. The dataset generated by the researchers in this research has its foundations in the art of web scraping. The authors suggest that taking support of the fact-checker websites is a good solution to reduce efforts while upholding factual accuracy. [5]

Zhang et.al., in their paper, present a novel deep-diffusive neural network-based model which based on a set of latent and explicit features extracted from texts, learns representations of news structures, content and authors simultaneously. They compare their model with state-of-the-art models and it is observed that their model outperforms traditional models. [6]

Kumar et.al. [7] in their paper, conclude that Convolutional Neural Net + bidirectional LSTM ensembled network with attention mechanism achieved the highest accuracy amongst all the models tested.

Thota et.al. in their paper [8], present a novel architecture for a neural net from scratch which outperformed traditional models by a significant margin and provided us a good idea about how to use NLP techniques available at our end to optimize the performance further.

Kaliyar et.al, in their research article [9], propose a novel deep CNN model termed as FNDNet, which aims to learn multiple features at each layer of training, and this performance is further validated using parameters such as Wilcoxon, F1 Score, and accuracy.

Girgis et.al., in their research article [10], compare multiple deep learning models such as RNN, CNN etc. and then conclude that RNN with Gated Recurrent Unit outperforms CNN in classification of fake news.

Agarwal et.al., in their article [11], propose usage of GloVe for word embedding as a tool for pre-processing in order to establish a lingual relationship via construction of vector space. This NLP technique acts as a perfect precursor to the author's proposed neural net techniques which when both combined give us a state-of-the-art model.

IV. DISCUSSION

The fundamental idea of our discussion is to build a model that can predict the credibility of real time news media.

As shown in Fig. 1, the proposed framework consists of five major phases:

1) Data collection from the news aggregators and storing it into a dataset
2) Data pre-processing such as cleaning and data preparation is done on this Dataset
3) An Application Layer (Web App) which will take URL as input to process
4) The Deep Learning Model consisting of Neural Networks will processed the data
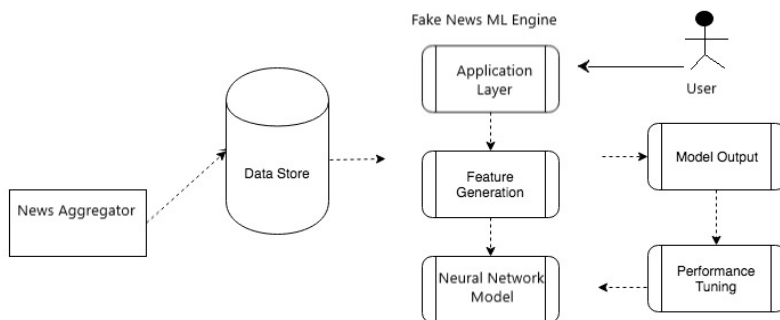5) Model output will be displayed on screen.

Fig. 1 System Architecture

A. Data Collection

Data collection is the process where we collect information from different sources. This data can be of Structured, Semi structured, or Unstructured data. As in our case the data which will be collected will be semi structured in nature. Dataset consisting of news items can be sourced from sources such as Kaggle, RSS (Really Simple Syndication) feeds and more. The data can also be scrapped from the web or trusted new sites using web scrapper tools like Beautiful soup,scrappy,etc. However, the web scrapped data is difficult to analyse. The emphasis of our proposed system is to teach the deep learning model segmenting genuine or falsified articles on basis of heading and/or its material without revealing where the actual origin of the news articles is coming from to avoid bias due to the source of news. Since the project is mainly targeted towards Indian Subcontinent, the scrapped/sourced dataset is also mainly consisting of relevant news media.

B. Figures and Tables

Data cleaning is the process to identifying and correcting errors in the dataset that may negatively impact machine learning model. According to recent works, it is verified that cleaning the data itself is improving the machine learning results[12].The Data Cleaning phase of the proposed system mainly emphasizes on retaining similar attributes of news headline.

The dataset should be cleaned appropriately, here the fields like Uniform Resource Locator, original source of data and Serial Identifiers like ID are discarded from this data in the cleansing phase as they are irrelevant data.

Similarly, the focus is on removal of irrelevant items, duplicate items and incomplete data in the dataset by either removal or imputation of the affected tuples. The troublesome tuples are discarded because few of the article's title or internal content contain extra line breaks, spaces or any kind of irrelevant composition that results in the textual matter in dataset slip into new columns or tuples, causing introduction of disorderly formatting into the resultant data.

After Cleaning, the dataset should also be pre-processed using textual filtering methods like Stemming of the data, Tokenization, removing stop words, Lemmatization, Removing unnecessary punctuation, tags. We can use NLTK to perform the data pre-processing operations as discussed above.

C. Data Preparation

Data preparation is the process of 3 steps assemble, merge, structuring and grouping data which can be used in data visualization applications. Data needs to be transformed into text which is understandable by the machine and feed into the model which is going to be trained. In natural language processing text search can be done very efficiently using a tool name Regular expression. It executes a task which includes segmentation of word, sentences along with the normalization and stemming. Using Regular expression stop words i.e. the words which are frequently used in the sentences along with the punctuations and special characters will get removed.

D. Application Layer

The Application Layer is an important phase in the proposed system for UX Purposes. After training the model on the cleaned dataset, the new input has to be provided to the trained model and providing this via command line interface is possible but since majority of the target audience for such systems are people from non-technical backgrounds, it is always advised to have a proper UI based frontend, either web, desktop or mobile for ease of usage. Here, our application layer, proposed to be developed using JavaScript Framework and connected to the trained model in the backend is extremely important. The proposed UI of the system will take the user input which will be the news URL and then classify it at the backend and then display the output on the screen.

E. Deep Learning Model

In our proposed system a neural network will be fed with N-gram vectors generated on title, news content from the dataset and trained for proposed classification. N-gram is a contiguous series of N items of given data in textual format \& it is quite useful in NLP. In simpler terms n grams is simply a sequence of n words. N-gram simply tells the probability of the occurrence of current words according to the occurrence of the previous or n-1 words. TF-IDF, also can be defined as the prediction of how relevant a word in a series is to a text/data. The meaning of the word increases proportionally to the number of times in the text a particular word appears but is replaced by the word frequency in the dataset.

The most commonly used machine learning classification algorithms such as Naive Bayes, Random forest, SVM, Decision trees are not able to perform well and are not able to get at least 90\% of accuracy. But here the Keras NN model with all the pre-processed media titles and/or media content in vector format, KNN can be used for detecting fake news with some dense layers. In the proposed model the KNN will be used and the layers 1,3,5 are using Rectified Linear Unit with 64 nodes, 16 nodes and 2 nodes respectively as the activation function. There will be an addition of one or two dropout layers of 15 to 20\% dropout rate to the proposed neural model. This is done to avoid the overfitting of the model and to be able to generalize the dataset.

F. Output Phase

In the last layer of the deep learning model, we have added the sigmoid activation function (formula stated below). Sigmoid function helps us map any output value between 0 to 1. As the value of probability also lies between 0 to 1, sigmoid acts as a good choice for activation function in case of probabilistic models. This is done to make sure that the final result of the neural model is either true or false i.e., binary format. For better generalization and to avoid the overfitting of the model we have also implemented early stopping. After this output is obtained, it would be passed to our application layer to display it to the end user who provided the news article as an input to our system.

V. ADVANTAGES

A. Social networking sites contain false information and people are often fallen prey to it.
B. Detecting those fake contents on these sites will safeguard the authenticity of the information available online.
C. Detecting Fake news that is forwarded can prevent the spread of rumors.
D. The proposed model provides better output than plain vanilla machine learning models.
E. It is easier to detect red flags in the news articles than manual methods of detection.
F. The proposed model emphasizes on avoiding biased treatment towards any news source.

VI. LIMITATIONS

A. Proposed model uses the N-gram approach which might fail to model more complicated contextual dependencies of the text once implemented.
B. Accuracy can be further improved by feature engineering methods.
C. As the rate of producing news is increasing day by day so it is challenging to effectively detect fake news.
D. Detection of Sarcasm is not optimal using the proposed model and there is a scope of improvement.

VII. FEASIBILITY STUDY

There are multiple datasets such as BuzzFeed, Kaggle Dataset and Sourjyar Indian News Dataset which have been cleaned for practical usage. There are GPUs like Nvidia RTX, Nvidia Titan and AMD Vega available to train the proposed model, thus we can say that no hardware bottleneck exists. Frameworks like PyTorch and Tensorflow exist to convert proposed model into actual implementation, which means there are no software or framework related bottlenecks. Thus, there are no bottlenecks and the proposed system is feasible to implement.

VIII. CONCLUSIONS

Many people consume news from social networks instead of mainstream news media. However, social media has also been used to spread misinformation, which has a bad impact on individuals and communities. In this paper, an innovative model for falsified news detection using deep learning algorithm \& N-Gram Vectorization has been proposed. This model takes news events as an input and based on the features observed by the deep learning model, it predicts whether news is false or true and displays it to the user via a web app interface.

IX. FUTURE SCOPE

The proposed model can be used by Social Media Websites to check authenticity of posts or the news content, which will have better recall and more accurate output compared to existing systems and a better choice to detect correctly and stop spreading of rumors. In future research, the researchers can consider use of GPT-3 to further enhance the results of this model. Future researches can consider using rating of the various news vendors as per the results of the authenticity of posts. Government Agencies can use this model to find people who spread misinformation regularly with malicious intents and arrest them. It can also be used for detection in other languages by appending an additional layer for transliteration/translation to our model. There is a huge need of credible research in this domain and the authors request research community to contribute via their research.

X. ACKNOWLEDGMENT

REFERENCES

[1] S. H. Kong, L. M. Tan, K. H. Gan and N. H. Samsudin, "Fake News Detection using Deep Learning," 2020 IEEE 10th Symposium on Computer Applications \& Industrial Electronics (ISCAIE), Malaysia, 2020, pp. 102-107, doi: 10.1109/ISCAIE47305.2020.9108841.

[2] L. Wu, Y. Rao, H. Yu, Y. Wang and N. Ambreen, "A Multi-semantics Classification Method Based on Deep Learning for Incredible Messages on Social Media," in Chinese Journal of Electronics, vol. 28, no. 4, pp. 754-763, 7 2019, doi: 10.1049/cje.2019.05.002.

[3] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[4] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. -W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," in IEEE Access, vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[5] F. T. Asr and M. Taboada, "Big Data and quality data for fake news and misinformation detection", Big Data Soc., vol. 6, no. 1, pp. 1-14, May 2019.

[6] Jiawei, Zhang Dong, Bowen & Yu, Philip. (2020). FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network. 1826-1829. 10.1109/ICDE48307.2020.00180.

[7] Kumar, S, Asthana, R, Upadhyay, S, Upreti, N, Akbar, M. Fake news detection using deep learning models: A novel approach. Trans Emerging Tel Tech. 2020.

[8] Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10.

[9] R. Kaliyar, A. Goswami, P. Narang, S. Sinha, FNDNet – A deep convolutional neural network for fake news detection, Cognitive Systems Research. Vol 61, 2020, Pg. 32-44, ISSN 1389-0417.

[10] S. Girgis, E. Amer and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," 2018 13th Intl. Conference on Computer Engineering and Systems, 2018, pp. 93-97.

[11] Agarwal, A., Mittal, M., Pathak, A. et al. Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning. SN COMPUT. SCI. 1, 143 (2020).

[12] D. J. \&. J. H. Martin, Regular Expressions, Text Normalization, Edit Distance, Jurafsky, Edition 3, 2018.

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089   (24*7 Support on Whatsapp)