2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-3701-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/Confluence52989.2022.9734129

Classifying Fake News Detection Using SVM, Naive Bayes and LSTM

Parita Jain
Computer Science and Engineering
KIET Group of Institutions
paritajain23@gmail.com

Swati Sharma
Computer Science and Engineering
KIET Group of Institutions
swati.sharma@kiet.edu

Monica
Information Technology
ABES Engineering College
monica@abes.ac.in

Puneet Kumar Aggarwal
ABES Engineering College
puneetaggarwal7@gmail.com

ABSTRACT- Web-primarily based media is a stage to state one's perspectives and viewpoints unreservedly and has made correspondence simpler than it became previously. This moreover opens up a risk for people to get out counterfeit phrase intentionally. The trustworthy entry to an collection of statistics sources on the net likewise brings the issue of people being provided to counterfeit facts and doubtlessly accepting such information. This makes it considerable for us to understand furthermore, banner such substance through on-line media. With the modern-day tempo of information created via online media, it's far hard to split between certifiable information and fabrications without knowing the wellspring of the news. In this paper, we propose various techniques to verify that the collected news is fake or not. For this, the approach named Natural Language Processing (NLP) is used. Various other methodologies like text classification, classification modeling is also used, and analysis of results has been done. Data from various sources was collected and to verify that the news is correct or not various techniques like SVM, Naive Bayes & LSTM are used.

KEYWORD: Fake News, Social Media, SVM, Naïve Bayes, LSTM

I. INTRODUCTION

Online media has emerge as an essential piece of our life. In the current day's modern-day atmosphere, these ranges are very helpful in spreading the news to the remotest of the a ways-off areas of the world. In addition, the expanding digitization has validated its activity in public and global legislative issues. In the advanced time of online media, individuals are continuous in offering photographs on stand out enough to be observed, which brings approximately the dissemination of new records over the web.

In today's world, where everything is news and people are interested more in other's lives, it is very difficult to stop spreading fake news. A lot of people use fake news to manipulate the minds of people. Fake news increases with the use of social media these days. The malicious & controversial messages are very much needed to stop at the origin [14]. Otherwise, it sometimes creates serious issues. A lot of companies like Facebook, Google, etc. are working on fake news detection and avoidance these days. People spread fake news to manipulate the minds of the crowd during the election. Fake news also created a lot of controversial mindsets because of its incompleteness and incorrectness. It is very much needed to stop this news, so that one can provide or spread only genuine & authentic information to the world. The approach just tried to make a contribution in this field to grow in the right direction.

As in step with Zubiaga, Aker, Bontcheva, Liakata, and Procter [1] any post what shares combined media content material that doesn't dependably addresses the event that it claims to is taken into consideration as a misleading or phony picture. On the alternative hand, the pictures which can be altered via making use of strategies like photo joining, replica circulate photograph changing, and so forth are delegated managed snap shots.

In the 2016 US legitimate elections, counterfeit news become found to have drawn in greater prominent engagement from on line media customers than the news published via normal news resources [2]. Bovet and Makse [3] seen that 25% of the 30 million tweets that contained connections to information resources in the course of the five months until the political race date have

been either especially one-sided or fake. Bogus statistics has been determined to have quicker dissemination, especially for legislative problems related information, and it summons sentiments like repugnance and dread, as contemplated within the answers given via the readers [4]. There had been several events in which peruses accept such information with-out confirming the realness of the information content material from dependable resources [5] [6] [17].

In this paper, we propose various techniques to verify that the collected news is fake or not. For this, the approach named Natural Language Processing (NLP) is used. Various other methodologies like text classification, classification modeling is also used, and analysis of results has been done. Data from various sources was collected and to verify that the news is correct or not various techniques like SVM, Naive Bayes & LSTM are used.

II. RELATED WORK

Numerous methods have been proposed to understand counterfeit news, which incorporate statistics mining and interpersonal business enterprise evaluation techniques. Shu et al. [7] signify counterfeit information location models into information content material models and social setting fashions. Conroy et al. [8] endorse operational guidelines for making plans a framework for test of data. The creators see promising results with the aid of giving an resourceful mixture technique that joins semantic signs and machine mastering, with community-primarily based conduct records.

Regular language preparing techniques for the invention of phony information have been assessed by Gilda [9]. Term recurrence reverse archive recurrence (TF-IDF) [10] of bigrams and probabilistic putting loose language structure (PCFG) discovery become applied with specific fashions which includes stochastic angle drop and slope boosting. TFIDF of bigrams with stochastic angle drop version distinguished phony information with a precision of 77.2%. Be that as it can, just the vector-primarily based method can't be applied to take a look at explicit highlights and teach the classifiers as these are specific to the specific making ready dataset.

Buntain and Golbeck [11] utilized primary, content based, client and fleeting highlights to plan a framework to distinguish counterfeit news in famous Twitter strings. The substance-based highlights incorporate extremity, subjectivity and conflict. Their machine′s relevance is constrained to exceedingly re-tweeted strings of Twitter discussions, and all things considered, most tweets are occasionally re-tweeted [18]. Their excessive acting model applied at the BuzzFeed dataset carried out a precision of 65.29%.

Jin et al.,[12] made one of the big endeavors to apply images for test of facts through using visual highlights like lucidity and rationality rating, and measurable highlights of pix like tally and picture percentage. Utilizing these picture highlights, they accomplished the most noteworthy check exactness of 83.6%. This precision was helped by more than 7% contrasted and different methodologies that utilization non-picture includes as it were [15] [16].

A blend of textual content based totally and consumer highlights became utilized by Krishnan and Chen [13]. Client highlights included number of partners, quantity of adherents, companions to devotee share moreover, if the client has a checked URL. Printed highlights comprise tweet duration, word tally, wide variety of query marks, quantity of interjection marks, range of URLs, wide variety of capital letters, wide variety of hash tags, and so forth. With a precision of 80.Sixty eight% for the Hurricane Sandy dataset, they acquire a high review without settling a lot on exactness. Be that as it may, their chose highlights are simply relevant to interpersonal organizations that have an idea of companions, adherents and client check [19].

In Section III, we explain our proposed methodology. We use three different algorithms and after comparison we found that the LSTM algorithm is best for fake news detection among all the three algorithms. The specified effects, together with discussions are provided within the effects and discussions segment. Finally, we conclude our work.

III. PROPOSED METHODOLOGY

To analyze things properly, we use three different algorithms and after comparison we found that the LSTM algorithm is best for fake news detection among all the three algorithms. In the procedure, we follow several steps like, first we convert the train dataset into embeddings and then analyze the things over the test

dataset. All the modelling techniques are defined as shown in Figure 1:

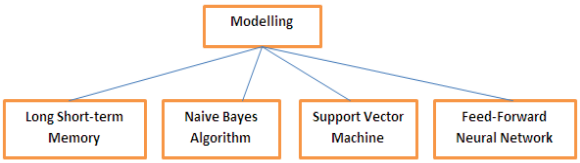


Figure 1: Modelling Techniques

Step 1: Modelling- We have collected the data from Kaggle and try to develop mathematical models using various machine learning algorithms, to analyze the precision & authenticity of various NLP approaches.

i) Naive Bayes Algorithm: We were looking for the best accuracy rate for our data, so we implemented a Naive Bayes classifier in the beginning. The Naive Bayes Algorithm uses the concept of Bayes theorem.

$$P(c|x) = (P(x|c)*P(c)) / P(x)$$

We use the maximum likelihood approach for estimation purposes. The advantage here is that it needs a very small amount of training data to estimate the parameters. Naive Bayes algorithms can be implemented in different ways, but here we use Gaussian Naive Bayes' approach using scikit-learn library of python. We draw the confusion matrix and analyze the facts.

(ii) Support Vector Machine (SVM): SVM is a supervised machine learning technique used for classification into two groups. Here we classify the news into two parts: reliable & unreliable, for this we use SVM. We can also separate the outliers using this modelling technique. It has a concept of a hyperplane that divides the set of points into two parts. We use this algorithm to classify the news and then we draw the confusion matrix to analyze the facts.

(iii) Feed-Forward Neural Network: Feedforward neural network is a kind of ANN (Artificial Neural Network) that creates a self-learning model learned from the past using activation function. The choice of the activation function is very crucial here. It was the first, finest and simplest ANN. It has various implementations like Single-layer perceptron, multi-layer perceptron & some other feedforward networks too. We use this approach to train the dataset [20].

(iv) LSTM: It stands for Long Short-term Memory and it is a complex area in deep learning. These Networks are a type of Recurrent Neural network which are best utilized in the domain of problems that require learning order and sequential dependency. Due to the traits of this model, it is best suited for our project because our project's central characteristic is text data classification. In the pre-processing part, we will use word embeddings because the order of words is essential so we cannot use doc2vec. The various processes which will be used in preprocessing are stemming, lemmatization, a bag of words intuition, etc. After preprocessing we feed the data to the LSTM model to train it and then we can test our model and check its accuracy using a confusion matrix.

Step 2: Data

The datasets used for this project were drawn from Kaggle [1]. The training dataset has about more than 15000 rows of data from a lot of information available over online platforms. We do some pre-processing of the collected data because it contains some part of source code.

A full training dataset has the following attributes:
1. id: unique id used for news content
2. title: the title used for news content
3. author: author of the news content
4. text: information in the form of text
5. label: • 1: unreliable • 0: reliable

IV. RESULT

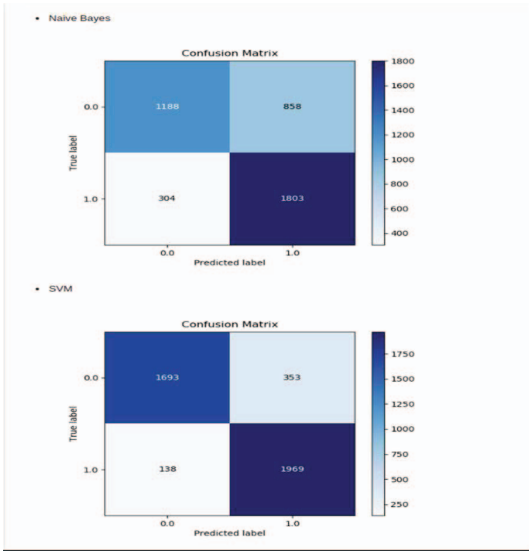
The model is compared with other models using their Confusion Matrices to calculate the Precision, Recall and the F1 scores. Table 1 shows our results.

Table 1: Performance of various models on Test Set

Name	Precision	Recall	F1 Score
Naive Bayes	0.68	0.86	0.76
SVM	0.85	0.93	0.89
LSTM	0.94	0.94	0.94

After considering all the observations we got the best results by using LSTM. Using LSTM, we got the best F1 score, precision & recall as compared to the other models.

Confusion matrix of SVM, Naive Bayes & LSTM defined as given below in Figure 2:



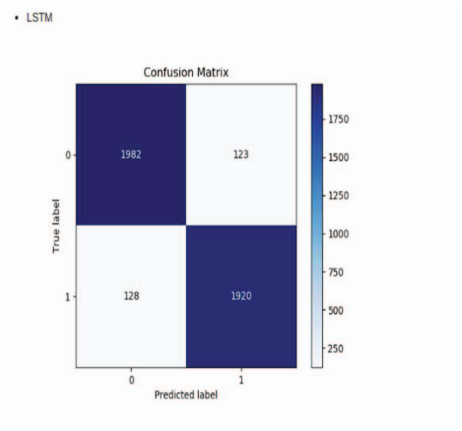


Figure 2: Confusion matrix of SVM, Naive Bayes & LSTM

V. CONCLUSION

Numerous procedures had been proposed to apprehend counterfeit news, which contain records mining and interpersonal agency analysis strategies. In this paper, we suggest diverse techniques to affirm that the collected news is fake or now not. For this, the approach named Natural Language Processing (NLP) is used. Various other methodologies like text classification, classification modeling is also used, and analysis of results has been done. Data from various sources was collected and to verify that the news is correct or not various techniques like SVM, Naive Bayes & LSTM are used. After considering all the observations we got the best results by using LSTM.

REFERENCES

[1] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey", ACM Computing Surveys, Vol. 51, pp.1–32, 2018.

[2] P. Dong et al., "A systematic review of studies on cyber physical system security", International Journal of Security and Its Applications, Vol. 9, pp. 155-164, 2015.

[3] A. Bovet, H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election", Nature Communications, Vol.10, 2019.

[4] S. Vosoughi, D. Roy, S. Aral, "The spread of true and false news online", Science, Vol.359, pp.1146–1151, 2018.

[5] P. K. Aggarwal, S. Sharma, Riya, P. Jain, Anupam, "Gaps Identification for User Experience for Model Driven Engineering", in Proceedings of Confluence, 2021.

[6] C. Kang, A. Goldman. In Washington Pizzeria Attack, FakeNewsBroughtReal Guns.

[7] K. Shu, A. Sliva, S. H. Wang, J. L. Tang, H. Liu. Fake news detection on social media: A data mining perspective. SIGKDD Explorations Newsletter, Vol.19, pp.22–36, 2017.

[8] P. K. Aggarwal, P. Jain, J. Mehta, R. Garg, K. Makar, and P. Chaudhary, "Machine Learning, Data Mining and Big Data Analytics for 5G-Enabled IoT", Blockchain for 5G enabled IoT: the new wave for Industrial Automation, pp. 351-375, Springer, 2021

[9] S. Gilda. Notice of violation of IEEE publication principles: Evaluating machine learning algorithms for fake news detection. In Proceedings of the IEEE 15th Student Conference on Research and Development, pp.110–115, 2017.

[10] P. Chaudhary, S. Goel, P. Jain, M. Singh, P. K. Aggarwal, Anupam, "The Astounding Relationship: Middleware, Frameworks, and API", In Proceedings of ICRITO, pp. 1-4, 2021.

[11] P. Jain, P. K. Aggarwal, P. Chaudhary, K. Makar, J. Mehta, and R. Garg, "Convergence of IoT and CPS in Robotics", Emergence of Cyber Physical Systems and IoT in Smart Automation and Robotics, pp.15-30, Springer, 2021

[12] Z. W. Jin, J. Cao, Y. D. Zhang, J. S. Zhou, Q. Tian. Novel visual and statistical image features for microblogs news verification. IEEE Transactions on Multimedia, Vol.19, pp.598–608, 2017.
[13] S. Krishnan, M. Chen. Identifying tweets with fake news. In Proceedings of International Conference on Information Reuse and Integration, pp.460–464, 2018.
[14] P. K. Aggarwal, P.S. Grover, and L. Ahuja, "Security Aspect in Instant Mobile Messaging Applications," In Proceedings of RAETCS, pp.1-5, 2018.
[15] P. Jain, Anupam, P. K. Aggarwal, K. Makar, V. Shrivastava, S. Maitrey, "Machine Learning for Web Development: A Fusion", in Proceedings of AIST, 2020.
[16] K. Shu, D. Mahudeswaran, H. Liu, "FakeNewsTracker: A tool for fake news collection, detection, and visualization", Computational and Mathematical Organization Theory, Vol.25, no.1, pp.60–71, 2019.
[17] P. Jain, P. K. Aggarwal, "Mobile Phone Maintainability Prediction using MCDM Methodology", International Journal of Recent Technology and Engineering, Vol. 8, Issue 3S2, pp. 410-415, 2019.
[18] C. Buntain, J. Golbeck, "Automatically identifying fake news in popular twitter threads", In Proceedings of International Conference on Smart Cloud, pp.208–215, 2017.
[19] P. K. Aggarwal, P. Jain, P. Chaudhary, R. Garg, K. Makar, J. Mehta, "AIIoT for Development of Test Standards for Agricultural Technology", Intelligence of Things: AI-IoT Based Critical Applications and Innovations, pp. 77-99, Springer, 2021.
[20] M. Singh, N. Sukhija, A. Sharma, M. Gupta, P. K. Aggarwal, "Security and Privacy Requirements for IoMT-Based Smart Healthcare System". Big Data Analysis for Green Computing, 17–37, Taylor & Francis, 2021.