iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 Issue: IV Month of publication: April 2020

DOI: http://doi.org/10.22214/ijraset.2020.4043

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

Comparison of Techniques to Detect Fake News

Dimple Rathod[1], Harsh Salvi[2], Manthan Mehta[3], Prof. Anusha Vegesna[4]
[1, 2, 3]Student, [4]Faculty, Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

***Abstract:** Authenticity of Information is becoming a longstanding issue affecting businesses and society, both for digital as well as printed media. On social networks, the information spreads at such a fast pace and is so amplified that inaccurate, distorted or wrong information acquires a tremendous potential to cause real world impacts for millions of users, within minutes. Currently a number of concerns about this problem were raised and consequently some solutions to mitigate this problem were put forward. There are various dynamic predictive models such as Random Forest, XGBoost, Naive Bayes, Logistic Regression that can be used for predicting whether a news article is fake or not. In this paper, we describe four different machine learning models. The performance of these models has been examined. Each model was tested against the header and content of the articles and the results were noted.*
Keywords: Fake news, Logistic regression, XGBoost, Naive Bayes, Random Forest

I. INTRODUCTION

In recent years, with the development of new media, competition in the news industry is heating up. In order to attract readers, many media start to make some eye-catching fake news, which has caused trouble to our daily life. Fake news in various fields is emerging endlessly, which has a bad influence on society. As news media, news reports should pursue authenticity, rather than "headline party", exaggerated and eye-catching. In today's century, when someone shows us the news, video or pictures of any event, how can we believe its authenticity? With generation of millions of news articles every day, we are facing the problem of fake news at a global level; and it has become one of the most urgent issues to be tackled in recent times. The surge in fake news was noticed during the 2016 US presidential elections that determined the fate of these elections. It is seen that the sharing of hoax news has been more than that of accurate news in many cases. Research indicates that WhatsApp and Facebook are the platforms that are most utilized for spreading fake news. An approximate of one in two Indians have agreed to have received fake news during the 2019 Lok Sabha elections [4]. One way to tackle fake news is to manually classify news as real or fake. Even though that seems like the simplest solution it is not practical. Hence, there is a need to look for a pragmatic technical solution to do the same. Machine learning models have been previously used with varying success to predict the authenticity of news articles. This paper looks at the performance of various such techniques and draws a comparison among them.

II. DATA DESCRIPTION

The dataset used was published on Kaggle. The labelled dataset was derived from various other datasets containing fake and real news published online. The real news data set has been collected from various publications like New York Times, Washington Post. Fake news articles are sourced from a total of 244 websites where articles were identified and reported as fake by BS Detector. The final compiled dataset has 4 main columns: title, body, publication and the label "fake". The title and body columns represent the textual unformatted content of each news article. A total of 28711 articles are present out of which 12999 are labelled as fake and 15712 labelled as real.

III.METHODOLOGY

The implementation involves the use of a dataset that has been split into a training set and test set for the purpose of creating a model. A 90-10 split is used. This results in 2871 test articles and 25839 training articles. The models are trained on headline and body separately. Since our parameters consist of textual data, this data cannot be directly used to train our machine learning algorithms. These algorithms take data in the forms of numbers and hence we need to convert our textual parameters to fixed length vectors which will act as input to the algorithm. For the same purpose, we adopted the method of TF-IDF Vectorizer (Term Frequency Inverse Document Frequency). This method tokenizes terms, learns the vocabulary and inverse frequency weightings, and then encodes the new terms. In layman terms, TF- the term frequency checks how many times a particular word appears in one news article; IDF – the inverse document frequency takes the log of { the number of articles in our corpus divided by the number of articles in which this term appears }. Both these new terms are multiplied to get the TF-IDF value of each word. TF-IDF are word frequency scores that find out the most important words that are to be taken into consideration for our model; that are more

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

interesting that are more frequent in a document but not across documents. Before performing any sort of vectorization, the stop words are removed from all textual data. Based on the TF-IDF values, this paper studies the implementation of four different classification algorithms: Random Forest, Logistic Regression, XGBoost, Naive Bayes in order to tackle the problem of fake news.

A. Logistic Regression

Logistic Regression is a classification algorithm and assigns observations to discrete sets of classes. The model's basis is a logit transformation and maximum likelihood estimation to estimate the probability of unreliability in relation to our predictor variables. The logit function gives an output with a range of 0 to 1. A threshold of 0.5 is set to convert the output to the classification label. The algorithm calculates the posterior p(x|y) in order to learn the labels to be assigned to the feature inputs. This is a prime example of a discriminative approach and although this is technically not a statistical classification, it gives the conditional probability of class membership which we use to assign a value. The sklearn library is used to implement this algorithm [2].

B. Random Forest

Random forest techniques are capable of performing both regression and classification by using a technique called Bootstrap Aggregation. Multiple decision trees are used to form multiple branches in order to determine the final output. Using a set of decision trees helps in increasing the accuracy of our model. It lies at the base of the Boruta algorithm, which selects important features in a dataset. Random forest algorithm has been implemented using the sklearn library [5].

C. Naïve Bayes

Naïve Bayes approach in its simplest form uses the basis of the Bayesian rule. Each feature is considered independent of any other feature. The algorithm learns a model of joint probability p(x,y), from each labeled article. Predictions are made using the Bayes' rule by computing the conditional probability, and then assigning labels based on the max probability [1] [3].

D. XGBoost

XGBoost (eXtreme Gradient Boosting**)** is an implementation of gradient boosting algorithm. It increases the speed and the efficiency of the model. The underlying method includes parallelization by creation of multiple decision trees. These decision trees are built by using information from previous decision trees one after the other, this sequential process is called boosting.The underlying concept aims to reduce the misclassification rate in subsequent iterations. It is an iterative process and does not use conventional ensemble techniques. Models are trained in succession. Each new model attempts to correct the mistakes of the previous model. This basically leads to avoidance of repetition of mistakes across models [2].

TF-IDF is not the only vectorization technique which can be used. In addition to TF_IDf, vectorization is performed using Google's pre-trained news Word2Vec model. Word2Vec creates a vector from a word that is in its vocabulary. It is a word vector representation. Its input is a large corpus of text which is mapped into a vector space which is several hundred of dimensions. Each unique word in our news articles is further assigned a vector in the area. Word2Vec uses the technique of similarity between two words that share common contexts in the corpus are located in close proximity to one another in the vector space. It uses neural networks to produce word vectors. To utilize the word vectorization technique for entire sentences and paragraphs, the mean of all the vectors in the words are taken. The same algorithms namely Logistic Regression, Naïve Bayes, Random Forest and XGBoost were trained on the vectors obtained from Google's Word2Vec word embeddings.

IV. RESULTS

TABLE 1: Results – Accuracy using TF-IDF Vectorization

Algorithm	Headline	Body
Logistic Regression	65.54%	82.77%
Random Forest	65.80%	84.39%
XGBoost	65.86%	85.78%
Naïve Bayes	64.80%	75.84%

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

The Table 1 above shows the accuracy obtained by all the algorithms when using the tf-idf vectorization. When it comes to detecting from the headline, we can see that almost all the algorithms perform similarly and there is no difference. When detecting from the body, we can see that tree type algorithms perform a bit better than simple logistic regression and a whole lot better than Naive Bayes. This shows that Naive Bayes is not able to effectively use the data contained in the body field compared to other algorithms.

TABLE 2: Results – Accuracy using Google Word2Vec

Algorithm	Headline	Body
Logistic Regression	70.93%	82.32%
Random Forest	71.43%	87.01%
XGBoost	71.82%	90.50%
Naïve Bayes	61.18%	65.54%

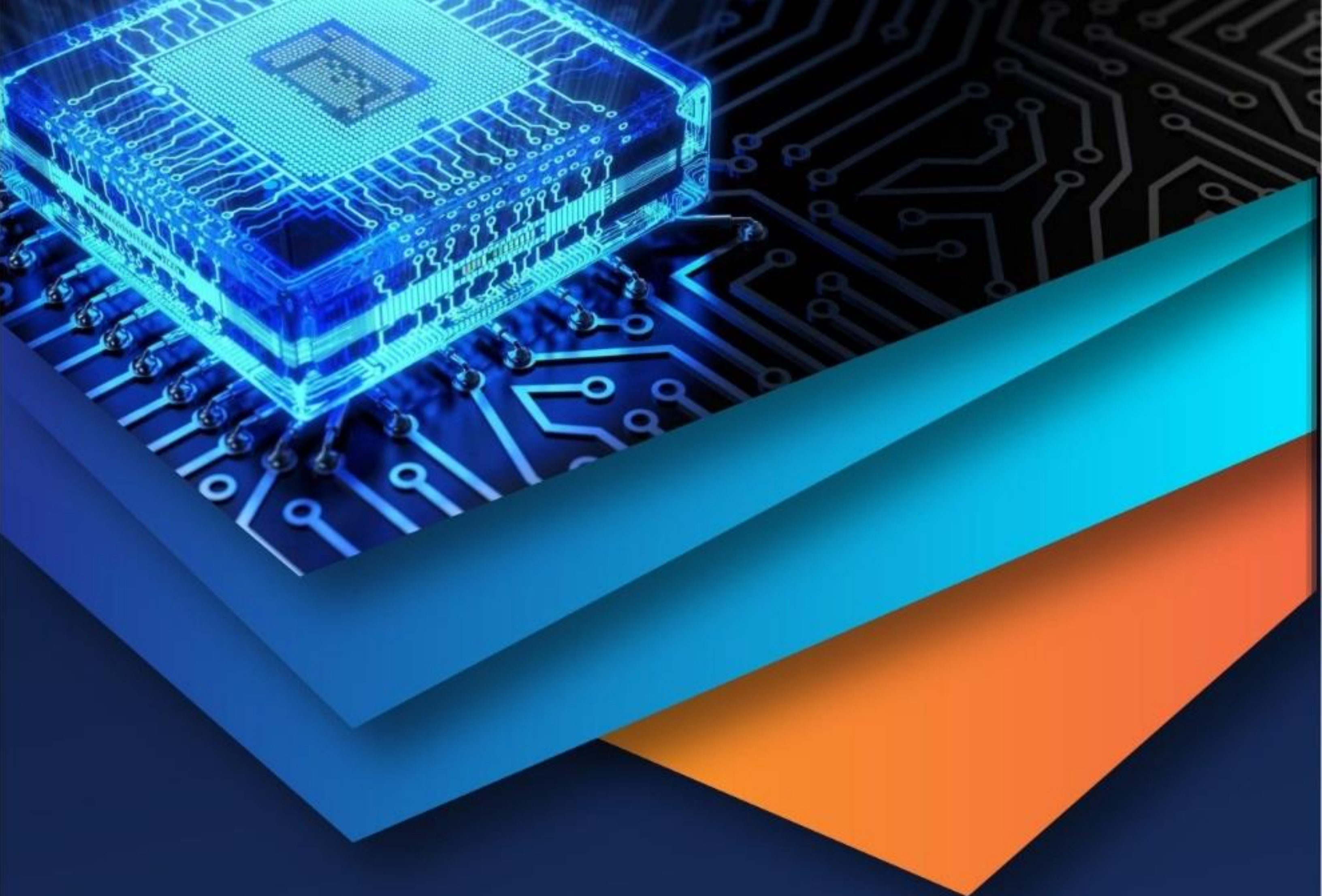
The Table 2 above shows the accuracy obtained when google word2vec is used as the vectorization technique instead of the standard tf-idf method. It can be noted that using word2vec significantly improves the accuracy on headline for all but Naive Bayes. The increase in accuracy with the body models is to a lesser extent but still relevant. Since word2vec adds meaning to the words, models seem to perform better. Naive Bayes algorithm is based on the assumption that all the columns in the data are independent. Since google word2vec outputs an array of 300 numbers which are linked, Naive Bayes' ignorance of inter column dependence leads to a lower accuracy.

V. CONCLUSION AND FUTURE WORK

Identifying and differentiating a fake news article from authentic news has become extremely crucial in today's day and age. We conclude that basic machine learning models have the capacity to identify sometimes subtle language patterns that may be difficult for humans to identify for the same purpose at a satisfactory rate. The development of these models has a wide scope of application in the real world and can prove to be beneficial in various sectors for distinction between increasing deceptive information and authentic information. This methodology has a scope of further improvement by including more parameters other than just the headline and body content while training the models for predicting the authenticity of the article. By tuning more parameters, one can achieve better accuracy. Improved models could also be developed by use of much more enlarged datasets. With larger datasets, neural networks can provide much better results if they are efficient in utlising the high amount of data. While these algorithms are effective in detecting the language used in these new articles, it cannot check for factual content.

REFERENCES

[1] Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). IEEE, 2017.
[2] Aphiwongsophon, Supanya, and Prabhas Chongstitvatana. "Detecting Fake News with Machine Learning Method." 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE, 2018.
[3] Mykhailo Granik, Volodymyr Mesyura, "Fake News Detection Using Naive Bayes Classifier", 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).
[4] Rubin, V., Conroy N., Chen Y., Cornwell S., "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News", Proceedings of the Second Workshop on Computational Approaches to Deception Detection Association for Computational Linguistics, 2016.
[5] S. Gilda, Evaluating machine learning algorithms for fake news detection - IEEE Conference Publication, 2019.



10.22214/IJRASET

INDEX COPERNICUS 45.98

ISRA JIF IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY