iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** X **Month of publication:** October 2020

DOI: https://doi.org/10.22214/ijraset.2020.31985

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

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 X Oct 2020- Available at www.ijraset.com

Classification of Fake News: A Comparative Analysis using NLP Techniques

Joseph Alexander [1], Reena Raj [2]
[1]*Student, SBM, CHRIST (Deemed to be University), Bangalore, India*
[2]*Assistant Professor, SBM, CHRIST (Deemed to be University), Bangalore, India*

Abstract: ***As news on social media is becoming more sought after, fake news has become a major public and government issue. The fake news uses interactive material to deceive readers and get exposure, thereby causing negative consequences and exploiting public events. The pervasive dissemination of fake news has the potential to have highly negative impacts on people and culture. Consequently, the identification of false news on social media has recently become an evolving research that attracts considerable interest. This paper attempts to investigate and compares the accuracy of supervised learning techniques which are Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest and Multinomial Bayes to find the best fit for the model.***

Keywords: fake news, deception, dissemination, classification, social media, SVM

I. INTRODUCTION

In today's era of big data, social media is amongst one of the primary ways through which people obtain information. The speed with which the fake news is circulating around the globe it has inevitably led to the need to minimise the public vulnerability it is causing by its dissemination in different fields. Fake news purposefully created to misguide the readers. It is a form of propaganda which is claimed to be genuine news. It is shared through mainstream news and social media. There are 3.5 billion users on different social media platforms which accounts for 40% of the population. Statistics indicate that people assume that the dissemination of fake news has risen dramatically (87%) by the use of the Internet, and that the bulk (62%) of fake news is believed to be generated by online news websites and platforms. Fake news detection on social media poses specific features and obstacles that make existing detection algorithms unreliable or non-applicable from conventional news media. It is deliberately written to persuade readers to believe false facts, making it impossible and nontrivial to identify on the basis of news content; thus, to help make a decision, we need to provide auxiliary information, such as social media user interactions.

And before the advent of the internet there were false news and hoaxes. The widely accepted definition of fake news from the Internet is, false articles purposely created to mislead users. Social networking companies and news agencies are publishing fake news to increase readership or as part of a strategic war. It was a longstanding issue. Social media is a double edge weapon for the distribution of news. On one hand, its low cost, convenient access, and rapid information sharing lead people to search out and absorb social media news. On the other hand, it allows for the wide dissemination of "fake news," that is, low-quality news with intentionally false facts. Therefore, detecting and curbing fake news is essential for social media sites, in order to provide credible information to users.

II. LITERATURE REVIEW

According to Nicole Brian in his research he has focused primarily on fake news which was defined as, "fabricated content that intentionally poses as news coverage of actual events." The author has attempted to find the efficacy and drawbacks of language-based approaches for detecting false news by the use of machine learning algorithms, including but not limited to convolutional neural networks and recurring neural networks. The result of this paper is to decide how much can be done by observing patterns found in the text and blinding the world to external knowledge in this assignment. To construct complete sets of positive and negative examples for document-level classification, available datasets for sentence-level classification have been searched and combined. The purpose of this project was to determine whether and how machine learning could be helpful in identifying patterns that are characteristic of real and fake news articles, and this was done by tracking important trigrams.[1]

This paper by J Zhang et al., is based on a collection of explicit and latent features extracted from the textual information, it implements a novel automated fake news credibility inference model, namely fake detector. It builds a deep, diffuse network model to simultaneously learn the representations of news storeys, creators, and topics. In order to compare fake detector with other state-of-the-art models, comprehensive experiments have been performed on a real-world fake news dataset, and the experimental results have shown the efficacy of the proposed model based on the categorical labels of news articles.

In the experiments, fake detector has extensively been compared with many baseline methods which are Hybrid CNN, LIWC (Linguistic Inquiry and Word Count), TriFN, Deepwalk, LINE, Propagation, RNN and SVM. [2]

According to Mykhailo Granik et al., in their paper they have exhibited a simple approach for fake news detection using the naive bayes classifier. This technique was applied as a software framework and checked against a data set of news posts from Facebook. Three broad Facebook pages, each from the right and left, as well as three mainstream political news pages (Politico, CNN, ABC News) were gathered from them. Fake news articles sometimes use the same collection of terms, which may mean that the particular article is actually a fake news article. The main concept is to treat each word of the news article separately. The paper also addressed ways to boost precision, which were to get more information and use it for training, eliminating stopwords, using stemming, separately handling uncommon words, and using group of words to measure probabilities instead of separate words. [3]

In this research paper by E Tachhini et al., have addressed how Social Network Sites (SNSs) have revolutionised the way data is disseminated by encouraging users to exchange information openly. As a result, as vectors for the diffusion of disinformation and hoaxes, SNSs are also increasingly used. 15,500 Facebook posts and 909,236 users make up the dataset. They have shown that, based on the users who "liked" them, Facebook posts can be categorised as hoaxes or non-hoaxes with high accuracy. Two classification methods have been used, one being logistic regression, which tests the accuracy of the algorithms as a function of the number of posts available as a training set. In general, since the training set can only be generated through a laborious manual post-inspection process, these findings tell us how much we need to invest in manual labelling to reap the benefits of automated classification. The other is a recent adaptation of Boolean crowdsourcing algorithms, which demonstrates how much knowledge our learning is about. The other is a recent implementation of the Boolean algorithms of crowdsourcing, which demonstrates how much data our learning can move from one set of pages to another. [4]

In this research paper by Perez Rosas et al., have tried to concentrate the automated recognition in fake material in the online news. Firstly, they have introduced two new datasets for the purpose of detecting fake news, covering seven different news domains, and secondly, they have performed a series of learning tests to establish reliable fake news detectors. Several sets of linguistic features have been extracted like Ngrams, Punctuation, Psycholinguistic features, Readability and Syntax. The research has been conducted on several experiments with different (combinations of) feature sets. They have used a linear SVM classifier and five-fold cross-validation, with accuracy, precision, recall, and F1 measures averaged over the five iterations. [5]

III.METHODOLOGY

A. Dataset Description

This dataset has 77964 rows and was taken from Kaggle. The dataset consisted of two independent variables which consisted the news title and the news article. The target variables had two labels namely real and fake. The articles in this dataset were related to the 2016 US Presidential elections.

B. Pre-Processing

1) *Data Cleaning:* In the initial phase punctuations and symbols were removed from the title and the text columns using the regex library, since these do not add much value to the NLP model.
2) *Tokenization:* This means breaking down of sentences into individual words or tokens. We do this because we need an individual meaningful entity to work upon and that can only be a word and not a complete sentence. These tokens are sometimes loosely referred to as words or phrases, but making a token distinction is often necessary. A token is an example of a series of characters that are grouped together as a useful semantic processing unit in some specific text.To perform tokenization, we used the nltk library and its dependencies, such as punkt and wordnet.
3) *Stop Words Removal:* There are words that do not hold much meaning and are used vaguely and abundantly. They are used, literally, for the construction of the sentence. Some instances of stopwords are, and, my, be, been, him, her, was, he, she etc. We imported a list of stopwords using the nltk corpus module.
4) *Lemmatization:* It refers to translating a word to its root form. It usually refers to converting words to its root form with the use of vocabulary and morphological analysis of words, normally aiming to remove the inflectional endings only and to return the dictionary form of the word. This was done using the WordNetLemmatizer in the nltk library.
5) *Count Vectorizer:* The count vectorizer provides a simple method to both tokenize the collection of texts in the documents and build a vocabulary of words which is known, but also to encode new documents using that vocabulary. In count vectorizer we only count the number of times a word appears in the document which results in biasing in favour of most frequent words. this ends up in ignoring rare words which could have helped is in processing our data more efficiently.

6) *TF-IDF (Term Frequency-Inverse Document Frequency):* Here TF acts same as the count vectorizer. But IDF gives more weightage to words which are rare and less weightage to common words. The tf-idf vectorizer will tokenize the data, learn the present and inverse document frequency weightings of the vocabulary, and allow you to encode new records. Alternatively, if we have implemented count vectorizer, we can use it with a tf-idf transformer to just calculate the inverse document frequencies and start encoding documents.

C. Modelling

The dataset was split into 70:30 ratio for training and testing purpose and the following models were implemented:

1) *Support Vector Machine (SVM):* It is a supervised technique in machine learning which is used for the challenges in classification and regression. It's mostly used in classification issues though. In the SVM algorithm, we map each data element as a point in an n-dimensional space with the value of each characteristic being the value of a unique co-ordinate. SVM works reasonably well when there is a clear separation margin between classes, such as the separation between classes which exists in this case i.e., real and fake classes. The high generalization ability of the method makes it particularly suited for high dimensional data such as text. [6]
2) *Logistic Regression:* It is one of the baseline supervised machine learning algorithm which is used for classification in natural language processing, and is also closely related to neural networks. It is a supervised machine learning classifier that extracts from the input real-valued characteristics, multiplies each by a weight, sums them, and passes the sum to produce a probability via a sigmoid function. To make a decision, a threshold is used. A positive weight on a function is pointed towards y=1 in the binary classification and a negative weight towards y=0.
3) *Random Forest:* Random Forests is a bagging type of ensemble model in which the base model for bagging is decision tree. The forest selects the classification with the most votes over all the trees in the forest and takes the average performance of different trees in the case of regression. In text classification, random forest classifiers are suitable for dealing with high-dimensional noise data.[7]
4) *Decision Tree:* Decision tree is a type of supervised learning algorithm often used in classification problems. It functions for the input and output variables both categorical and continuous. It learns from data to approximate a sine curve with a set of if-then rules and is ideal for decision-making. It is applied to a simple structure that defines a collection of rules and regulations and used for decision making to assign the text into its category on the basis of its content.[8]
5) *Multinomial Bayes:* The Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on the naive assumption of conditional independence between each pair of features. The Bayes theorem calculates the probability of P(c|x) where c is the class of possible results and x is the instance to be identified, reflecting some unique characteristics. Naive Bayes are often used in problems with natural language processing. The tag of a text is predicted by Naive Bayes and the likelihood of each tag for a given text is determined and then the tag with the highest one is output.

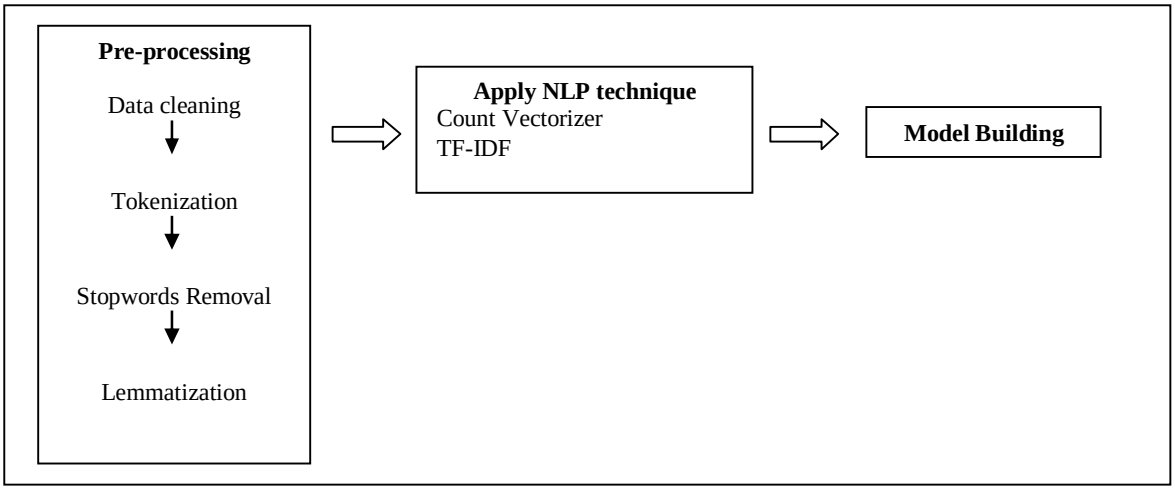


Fig. 1: Proposed Framework

IV. RESULTS

A. Metrics

The classifiers are compared based on: Accuracy, Precision, Recall and F-Measure. The true positives separated by the expected positives i.e., the sum of actual positives plus the false positives are precision. Recall is the rate of the true positive and is also called the sensitivity, which is the real positive value divided by sum of the real positive and false negative. In a model, two versions with low accuracy and high recall or vice versa are hard to compare. So, we use F-Score in order to make them equivalent. The F-score helps to simultaneously assess recall and precision. By punishing the extreme values further, it uses harmonic mean in place of arithmetic mean.

B. Classifiers Result

SVM classifier has the highest precision, 94.49% and therefore the best classification quality as shown in Table 1. Logistic and Multinomial Bayes classifiers had the best recall that is best sensitivity of 91.29% and 91.04% respectively. The F-measure combines precision and recall, the SVM and Logistic classifiers outperformed others at 92.53% and 92.05% respectively. In terms of accuracy SVM was the best with an accuracy of 92.34%.

TABLE I
Classifiers result

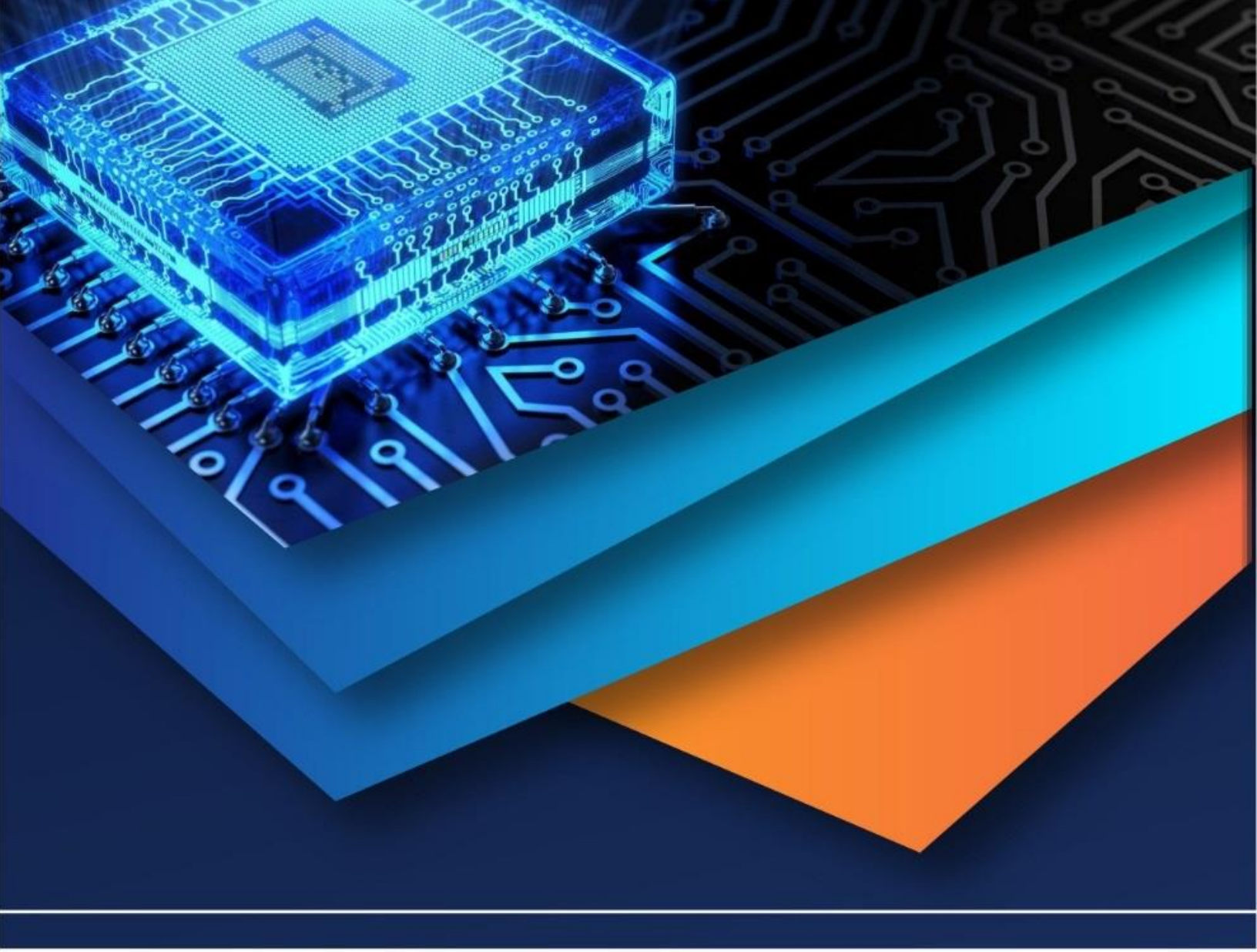
	Accuracy	Precision	Recall	F-measure
SVM	0.92344	0.94497	0.90649	0.92533
Logistic Regression	0.91713	0.92824	0.91291	0.92051
Random	0.83662	0.85309	0.81111	0.83157
Decision Tree	0.80584	0.89189	0.75913	0.82018
Multinomial Bayes	0.88161	0.85822	0.91040	0.88354

V. CONCLUSION

We have presented a method to detect fake news with the ability of a user to discern useful information from the internet services especially when news becomes critical for decision making. The issue of fake news has become more than just a marketing challenge, given the evolving landscape of the modern business world, as it warrants serious efforts by security researchers. We have proposed a simple but effective approach with the help of supervised machine learning algorithms to detect and classify real and fake news. This discussion draws up a simple typology of available methods for further refinement and evaluation and provides a framework for the creation of a systematic tool for detecting fake news.

REFERENCES

[1] N. Brien, “Machine Learning for Detection of Fake News.,” M. Eng. thesis, Massachusetts Institute of Technology, Cambridge, Jun. 2018.
[2] Zhang, J., Dong, B., & Yu, P. S., “Fake Detector: Effective fake news detection with deep diffusive neural network.” Proceedings - International Conference on Data Engineering, 1826–1829, Apr. 2020.
[3] Granik, M., & Mesyura, V. (2017). “Fake news detection using naive Bayes classifier.” 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings,2017, 900–903.
[4] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L., “Some like it Hoax: Automated fake news detection in social networks.” CEUR Workshop Proceedings, 1960, 1–12, 2017
[5] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R., “Automatic Detection of Fake News”. Sept. 2017.
[6] Zi Qiang Wang, X. S. “An Optimal SVM based Text based classification algorithm”. Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, 1378-1381, 2006
[7] Xu, B., Guo, X., Ye, Y., & Cheng, J. “An Improved Random Forest Classifier for Text Categorization”. Journal of Computers, 2913-2920, Dec. 2020.
[8] Pranckevicius, T., & Marccinkevicius, V. “Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification”. Baltic J. Modern Computing, V, 221-232, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY