IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake News Detection using various Machine Learning Techniques

Akanksha Akulwar
Department of Computer Science
College Of Engineering Pune
Pune, India

Anish Khobragade
Department of Computer Science
College Of Engineering Pune
Pune, India

Abstract—Online platforms are being used for outspreading malicious talk, which creates an impact on the minds of millions. Many distinct approaches have been brought forward to detect this fake news, but very few have been carried out in the actual world. We address this problem of estimating the rumor authentication in a real-world in less time with significantly high accuracy. We design and implement an approach addressing the above issue. We accessed whether the news is fake or not using various Machine learning techniques. We evaluate this algorithm on a set of data set scrapped from random online sites. The result shows that the performance of this improved algorithm is better than the original classification method. And finally, we consider various sizes of data to view and compare the accuracy.

Keywords—Support Vector Machine, Decision Tree, Fake Content, Naive Bayes, Machine Learning Models

I. Introduction

Fake news is referred to as misinformation either shared intentionally or sometimes by mistakes. This type of misinformation may influence the mass population and can lead to diverse effects on society. In this era of fast living lives, people prefer news consumption from online mode through social media platforms. The less time-consuming and cheaper rates also attract many to follow this trend. Furthermore, sharing, commenting, and discussing with friends becomes much easier through social media platforms. There is a vast range of examples that occurred due to the spreading of news. Some of such examples are tweets regarding 2016 us presidential elections, 2013 Boston marathon blast; these are popular fake news traversed through social media platforms. Such false news spread rate is much higher or can say faster than traditional news channels. This fake news though it travels at higher rates but has quality and authenticity issues. Consumers persuade to accept biased and false beliefs due to the intentionally spreading of fake news.

Detecting fake news and preventing it from spreading imposes lots of several new research problems. Firstly, Fake news is deliberately spread to deceive users, which makes it certainly impossible to expose them simply out from the news content as they all are diverse in topics, styling, and social platforms. Sometimes the sources spreading this fake news try to mock the Traditional news system to make people believe in their content. Thus Data specific feature-based detection is not sufficient, but other auxiliary information like social engagement and other knowledge bases are also needed. Secondly, this auxiliary information which is knowledge-based and obtained through social interaction-, fails for time-critical events newly emerging events. Also, such information may be big, unstructured, and noisy, and Hence defining appropriate methods and detecting at a higher accuracy has become open areas for research purposes.

II. Related Work

This article describes an easy fake news detection method supported by one among the synthetic intelligence algorithms – na¨ıve Bayes classifier. The research aims to look at how this particular method works for this particular problem given a manually labeled news dataset and to support (or not) the idea of using AI for fake news detection. The difference between these articles and articles on similar topics is that in this paper, the naive Bayes classifier was specifically used for fake news detection; the developed system was also tested on a comparatively new data set, which gave a chance to gauge its performance on a recent data. [1]

Their primary purpose is to manipulate the information that can make the public believe in it. There are lots of examples of such websites all over the world. Therefore, fake news affects the minds of the people. According to a study, Scientists believe that many artificial intelligence algorithms can help in exposing false news. This is because artificial intelligence is now a day becoming very popular, and many devices are available to check it partially. In this, the deep learning and machine learning concepts are used to detect fake news using na¨ıve Bayes classifier. [2]

III. Materials and Methods

A. Materials

The core of data implemented in this project had around 7818 articles of data. These articles mainly constituted news about U.S.politics. The Dataset is obtained from a random online site that has a lot of noisy data and thus required cleaning. The main features included in each row of the data were title, text, label of being fake or true. The Dataset has a mixture of true and fake news with a total 3154 number of fake entries and 3161 true entries

Table 1 Data Description

TITLE OF NEWS
TEXT OR ACTUAL NEWS
LABEL (EITHER FAKE OR REAL)

B. Data Pre-processing

The data thus obtained from the site needs to be cleaned before actual implementation. Feature Extraction, stemming, tokenizing, and then Classification is some of the preprocessing techniques that are followed during this stage. Details of preprocessing of data are as follows:

- To Find out missing entries in each column and filling them with an empty string.
- Then Merging the columns labeled as Title and Text to form a new column termed as title-text for further convenience.
- It is Then sorted for Unique values that are removing the null entries.
- Parametrize all label values under single nomenclature; for example, all ”fake” and ”Fake” are labeled with ”FAKE” similarly with the Real news.
- Further processing involves removing stop-words, removal of numbers, URLs, and special characters.
- Later, theDataset is split into testing and training datasets. The training dataset is used to train the Classifiers, and the testing dataset is used to get an unbiased estimation of unknown data.

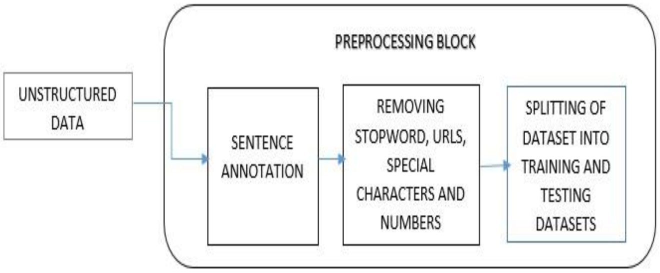
PREPROCESSING BLOCK
UNSTRUCTURED DATA
SENTENCE ANNOTATION
REMOVING STOPWORD, URLS, SPECIAL CHARACTERS AND NUMBERS
SPLITTING OF DATASET INTO TRAINING AND TESTING DATASETS

Fig. 1. Data Preprocessing flow

C. Models

1) Na¨ıve Bayes: Na¨ıve Bayes is a conditional probability model which can be used for labeling. The goal is to find a way to predict the class variable (B) using a vector of independent variables (A), i.e., finding the function f: A→B. . In probability terms, the goal is to find P(B|A), i.e., the probability of B belonging to a certain class A. ‘B’ is generally assumed to be a categorical variable with two or more discrete values. It is a mathematically simple way to include contributions of many factors in predicting the class of the next data instance in the testing set. The limitation of Na¨ıve Bayes is that they assume that all features are not dependent on each other. The Na¨ıve Bayes rule is based on the theorem formulated by Bayes:

$$P(r|s) = \frac{P(s|r)P(r)}{P(s)} \tag{1}$$

2) Logistic Regression: The logistic approach in machine learning is used for predictive analysis and for classification problems. It uses the probability concept for predicting the class of the variable. Logistic regression is a complex version of Linear regression. It uses the Sigmoid function termed as Logistic Function, which limits the hypothesis range between 0 and 1, i.e., 0 ≤h(x)≥1. Below is the equation for logistic regression: Here, Y is output, whereas x is the input variable,

$$h(Y) = \frac{1}{1+e^{-(b0+b1x)}} \tag{2}$$

b0 is a biased variable, and b1 is the coefficient of x. both b0 and b1 (beta values) are estimated from training data through the maximum likelihood estimator. The main logic behind this algorithm is, a certain threshold value is set by the user. Depending on the threshold value, the variables are categories as class A or class B (i.e., FAKE or REAL)

3) Support Vector Machine: A support vector machine (SVM), which may be used interchangeably with a support vector network (SVN), is additionally considered to be a supervised learning algorithm. SVMs work by being trained with specific data already organized into two different categories. Hence, the model is made after it’s already been trained. Furthermore, the goal of the SVM method is to distinguish which category any new data falls under. In addition, it must also maximize the margin between the two classes. The optimal goal is that the SVM will find a hyperplane that divides the Dataset into two groups. The kernel used in this application is RBF, as it is best suited for large applications like a corpus of news articles. The Radial Basis function on two samples x and x’ is given by:

$$K(x,x') = exp(-\frac{||x-x'||^2}{2\sigma^2}) \tag{3}$$

Where numerator represents the squared Euclidean distance and sigma is a free parameter.

4) Decision Tree: Decision Trees are a kind of Machine learning algorithm where we need to specify inputs with corresponding outputs within the training data(termed as Supervised learning) where the data is repeatedly split consistent with a particular parameter. The tree is often described by two units, namely decision nodes and leaves. The leaves are eventually the end results. And the decision nodes are where the info is split. It calculates the comparative change in entropy with reference to the independent variables. Alternatively,

$$IG(S,A) = H(S) - H(S,A) \quad (4)$$

$$IG(S,A) = H(S) - H(S,A) \quad (5)$$

where IG(S, A) is that the information gain by putting in feature A. H(S) is that the entropy of the whole set, while H(S, A) is the entropy after putting in the feature A, and P(x) is that the probability of event x

5) K- Nearest Neighbour: K-nearest neighbor makes predictions by operating on the training dataset directly. Predictions are made for a replacement instance (x) by rummaging down the whole training set for the K most alike instances (the neighbors) and outlining the output variable for those K instances. For Classification, this could be the mode class value; In regression, this could be the mean of the output variable. A distance measure is employed to find out which K instances in the training dataset are almost like replacement inputs. For real-valued input variables, the foremost accepted

$$d = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2} \quad (6)$$

distance standard is Euclidean distance. The above equation defines Euclidean distance between two points A(X1, Y1) B(X2, Y2).

IV. Results

Table 2 shows the results achieved after evaluating the accuracies of all the above-mentioned Machine learning models. The accuracy, precisions, and f1 score can be computed with the help of confusion matrices. A single confusion matrix was created for each model. The values shown are the averaged values over successive trials.

Table 2 Results of Algorithms

Models	Accuracy
Naive Bayes	87.91%
Logistic Regression	86.00%
Support vector Machine	67.00%
Decision Tree	72.85%
K-Nearest Neighbor	67.15%

Based on the results in Table 1, the graph in Fig 2 is constructed by taking different algorithms on X-axis and accuracy on the Y-axis. It is inferred that the Na¨ıve Bayes algorithm provides us with the highest accuracy, followed by Logistic Regression, followed by Decision Tree, then by KNearest Neighbour, and finally SVM. The Na¨ıve Bayes gives the best result due to its simple approach of finding conditional probabilities.

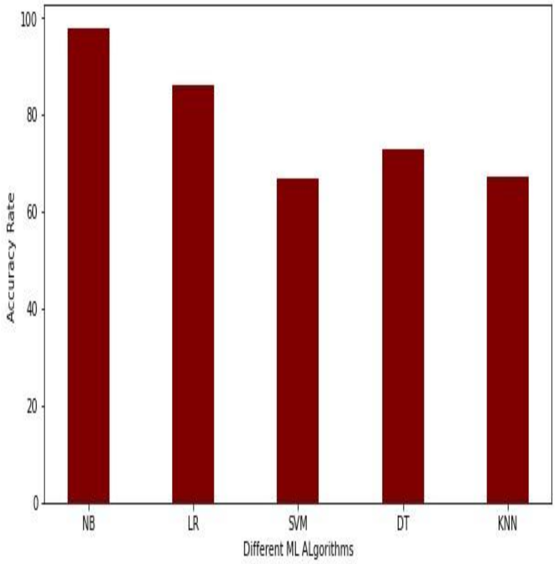

Fig. 2. Comparison of different Machine learning algorithm

V. Conclusion

This paper has given out a model for fake news revelation through different machine learning techniques. Furthermore, the paper investigated the five methods and compared their accuracies. The model that achieves the highest accuracy is Na¨ıve Bayes, and the highest accuracy score is 87.91%. Fake news detection is an evolving research area that has a scarce number of datasets. There are no data on real-time news or regarding current affairs. The current model is run against the existing Dataset, showing that the model performs well against it. In our future work, news story data are often considered associated with recent incidents within the corpus of knowledge. The next step would be to train the model and analyze how the accuracy varies with the new data to further improve it.

Acknowledgment

We are profoundly grateful to Prof. Vahida Attar, Head of Department of Computer Engineering and Information Technology, COEP, for her continuous support and Mr. Anishraj Khobragade for her expert guidance throughout to see this project reach its target from its commencement to its completion

References

[1] Mykhailo Granik, Volodymyr Mesyura, “Fake News Detection Using Naive Bayes Classifier,” *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*

[2] R.J. Poovaraghan, M.V. Keerti Priya, P.V. Sai Surya Vamsi, Mansi Mewara, Sowmya Loganathan,“Fake News Accuracy using Naive Bayes Classifier”, *International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8, Issue- 1C2, May 2019.*

[3] Hunt Allcott and Matthew Gentzkow, “Social media and fake news in the 2016 election,” *Technical report, National Bureau of Economic Research, 2017.*

[4] A. Bessie and E. Ferrara, “Social bots distort the 2016 U.S. Presidential election online discussion,” *FM, vol. 21, no. 11, Nov. 2016.*

[5] Sadia Afroz, Michael Brennan, and Rachel Greenstadt, “Detecting hoaxes, frauds, and deception in writing style online,” *2012.IEEE Symposium on Security and Privacy, San Francisco, CA, 2012.*

[6] Balmas, Meital. “When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism,” *Communication Research, vol. 41, no. 3, Apr. 2014.*

[7] Conroy, N., Rubin, V. and Chen, Y,” Automatic deception detection: Methods for finding fake news,” *Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.*

[8] Hunt Allcott and Matthew Gentzkow, “Social media and fake news in the 2016 election,” *Technical report, National Bureau of Economic Research, 2017.*

[9] Fake news websites. (n.d.) Wikipedia. [Online]. Available: *https://en.wikipedia.org/wiki/Fake$_n$ews$_w$ebsite.AccessedFeb.*6,2017.

[10] Naive Bayes classifier. (n.d.) Wikipedia. [Online]. Available:*https://en.wikipedia.org/wiki/Naive$_B$ayes$_c$lassifier.AccessedFeb.*6, 2017.

[11] Kelly Stahl., “Fake news detection in social media”, *B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 2018.*

[12] Saranya Krishnan, Min Chen.,” Identifying tweets with fake news”, *2018 IEEE Conference on Information Reuse Integration for Data Science.*

[13] Saxena, R., “How the Naive Bayes Classifier works in Machine Learning”, *http://dataaspirant.com /2017/02/06/naive-bayes-classifiermachine-learning/, Retrieved: April 15, 2019.*

[14] Akshay Jain, “Fake News Detection”, *IEEE 2018 International Students Conference on Electrical, Electronics and Computer Sciences.*

[15] Brambrick, Aylien, “Support Vector Machines: A Simple Explanation”, *https://www.kdnuggets.com/2016/07/support-vector-machines-simpleexplanation.html, Retrieved: April 20, 2019.*