Fake News Detection using Machine Learning

Sneha Sanghpriy Moon[1], Akshata Lonare[2], Aditi Chandekar[3] Mayuri Khandre[4], Prof. Mahesh Dumbere[5]

Student, Department of Computer Science & Engineering[1,2,3,4]

Associate Professor, Department of Computer Science and Enginering[5]

Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, India

moonsnehasanghpriy@gmail.com

Abstract: *Most of the smart phone users prefer to read the news via social media over internet. The news websites are publishing the news and provide the source of authentication. The question is how to authenticate the news and articles which are circulated among social media like WhatsApp groups, Facebook Pages, Twitter and other micro blogs & social networking sites. It is harmful for the society to believe on the rumors and pretend to be a news. The need of an hour is to stop the rumors especially in the developing countries like India, and focus on the correct, authenticated news articles. This paper demonstrates a model and the methodology for fake news detection. With the help of Machine learning and natural language processing, it is tried to aggregate the news and later determine whether the news is real or fake using Support Vector Machine. The results of the proposed model is compared with existing models. The proposed model is working well and defining the correctness of results upto 93.6% of accuracy.*

Keywords: Fake news , Machine Learning,, News, Online Fake news, Recommendation, text classification, social media

I. INTRODUCTION

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited be This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

Fake news detection is made to stop the rumors that are being spread through the various platforms whether it be social media or messaging platforms, this is done to stop spreading fake news which leads to activities like mob lynching, this has been a great reason motivating us to work on this project. We have been continuously seeing various news of mob lynching that leads to the murder of an individual; fake news detection works on the objective of detecting this fake news and stopping activities like this thereby protecting

the society from these unwanted acts of violence.[1][3][5].

The main objective is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news. This leads to consequences in social networking sites like Facebook, Instagram, microblogging sites like Twitter.

ISSN 2581-9429 IJARSCT

II. RELATED WORK

Overview

Several approaches have been taken to detect the fake news after massive widespread fake news in recent times. There are three types of fake news contributors: social bots, trolls, and cyborg users [3][4].Social Bots says, if a social mediaaccount is being controlled by a computer algorithm, then it is referred to as a social bot. The social bot can automatically generate.

Machine Learning (ML) Classification

Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data [11]. There are six algorithms that are adopted in this paper for classifying the fake news.

Social Media and Fake News

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking and wikis [1][2]. On the other side, some researchers consider the fake news as a result of accidental issues such as educational shock or unwitting actions like what happened in Nepal Earthquake case [3][4]. In 2020, there was widespread fake news concerning health that had exposed global health at risk. The WHO released a warning during early February 2020 that the COVID-19 outbreak has caused massive 'infodemic', or a spurt of real and fake news—which included lots of misinformation.

Decision Tree

The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a "test" on an attribute and the branching is done on the basis of the test conditions and result. Finally the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation

Random Forest

Random Forest are built on the concept of building many decision tree algorithms, after which the decision trees get a separate result. The results, which are predicted by large number of decision tree, are taken up by the random forest. To ensure a variation of the decision trees, the random forest randomly selects a subcategory of properties from each group [16][17]

The applicability of Random forest is best when used on uncorrelated decision trees. If applied on similartrees, the overall result will be more or less similar to a single decision tree. Uncorrelated decision trees can be obtained by bootstrapping and feature randomness.

Support Vector Machine (SVM)

The SVM algorithm is based on the layout of each data item in the form of a point in a range of dimensions n (the number of available properties), and the value of a given property is the number of specified coordinates [13]. Given a set of n features, SVM algorithm uses n dimensional space to plot the data item with the coordinates representing the value of each feature. The hyper-plane obtained to separate the two classes is used for classifying the data.

WhatsApp Work for Fake News Detection

To stop the spread of misinformation, WhatsApp has implemented some security measures and also fake news detection, though these are under alpha phase and are yet to be rolled out to the beta users. WhatsApp testing

Suspicious Link Detection" feature: This feature will alert uses by putting a red label on links that it knows to lead to a fake or alternative website/news. Additionally, if a message has been forwarded from a device more than 25 times, the message could be blocked.

III. PROPOSED MODEL

There are four main modules in the proposed model, named as

Data use

So, in this project we are using different packages and to load and read the data set we are using pandas. By using pandas, we can read the .csv file and then we can display the shape of the dataset with that we can also display the dataset in the correct form. We will be training and testing the data, when we use supervised learning it means we are labeling the data. By getting the testing and training data and labels we can perform different machine learning algorithms but before performing the predictions and accuracies, the data is need to be preprocessing i.e. the null values which are not readable are required to be removed from the data set and the data is required to be converted into vectors by normalizing and tokening the data so that it could be understood by the machine. Next step is by using this data, getting the visual reports, which we will get by using the Mat Plot Library of Python and Sickit Learn. This library helps us in getting the results in the form of histograms, pie charts or bar charts.

Preprocessing

The data set used is split into a training set and a testing set containing in Dataset 1-3256 training data and 814 testing data and in Dataset II- 1882 training data and 471 testing data respectively. Cleaning the data is always the first step. In this, those words are removed from the dataset. That helps in mining the useful information. Whenever we collect data online, it sometimes contains the undesirable characters like stop words, digits etc. which creates hindrance while spam detection. It helps in removing the texts which are language independent entities and integrate the logic which can improve the accuracy of the identification task.

Feature Extraction

Feature extraction is the process of selecting a subset of relevant features for use in model construction. Feature extraction methods helps in to create an accurate predictive model. They help in selecting features that will give better accuracy. When the input data to an algorithm is too large to be handled and i ts supposed to be redundant then the input data will be transformed into a reduced illustration set of features also named feature vectors. Altering the input data to perform the desired task using this reduced representation instead of the full-size input. Feature extraction is performed on raw data prior to applying any machine learning algorithm, on the transformed data in feature space.

Training the Classifier

As In this project I am using Scikit-Learn Machine leanring ibrary for implementing the architecture. Scikit Learn is an open source python Machine Learning library which comes bundled in 3rd distribution anaconda. This just needs importing the packages and you can compile the command as soon as you write it. If the command doesn"t run, we can get the error at the same time. I am using 4 different algorithms and I have trained these 4 models i.e. Naïve Bayes, Support Vector Machine, K Nearest Neighbors and Logistic Regression wic are very popular methods for document classification problem. Once the classifiers are traned, we can c heck the performance of the models on test-set. We can extract the word count vector for each mail in test- set and predict it class with the trained models

IV. METHODOLOGY

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. Figure [1] describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random forest, SVM and Naïve Bayes, Linear Regression majority voting and

other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision
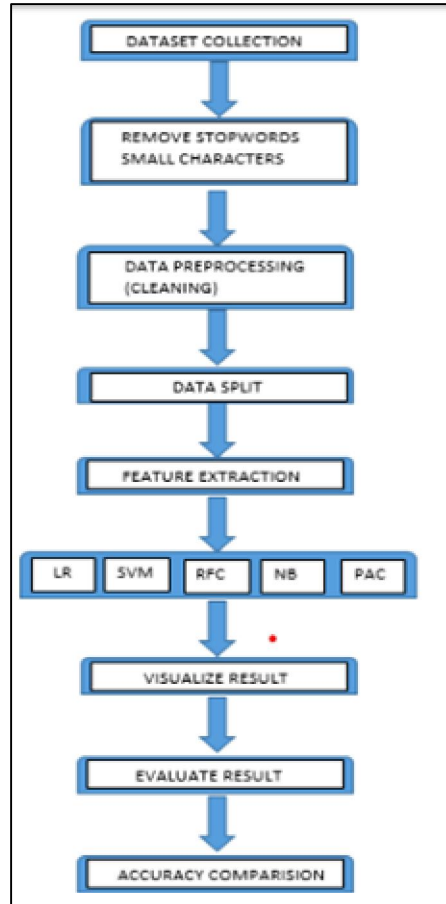

Fig 1. Describes the Proposed System Methodology

In this paper a model is build based on the count vectorizer or a tfidf matrix (i.e) word tallies relatives to how often they are used in other artices in your dataset) can help . Since this problem is a kind of text classification, Implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for countvectorizer or tfidf-vectorizer, this is done by using a n-number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as "the", "when", and "there" and only using those words that appear at least a given number of times in a given text dataset

System architect
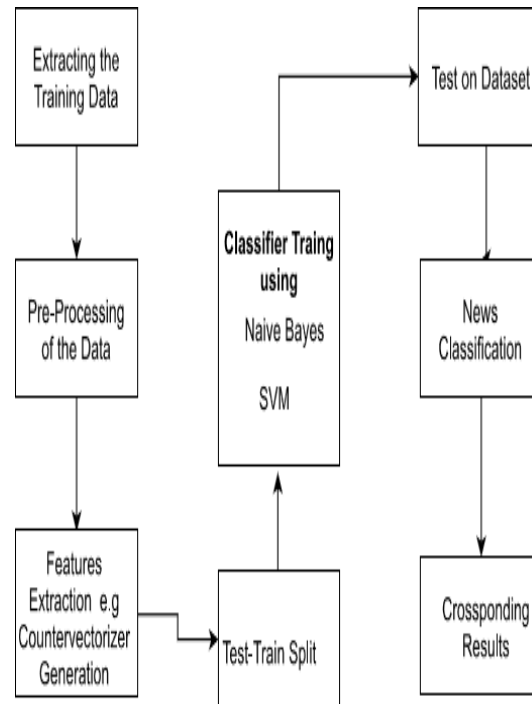
V. CONCLUSION

It is significant to find the accuracy of news which is available on internet. In the paper, the components for recognizing Fake news are discussed. A mindfulness that not all, the fake news will propagate via web-based networking media. Currently, to test out the proposed method of Naïve Bayes classifier, SVM, andNLP are used. In future, ensuing algorithm may provide better results with hybrid approaches for the same purpose fulfilment. The mentioned system detects the fake news on the based on the models applied. Also it had provided some suggested news on that topic which is very useful for any user. In the future, the efficiency and accuracy of the prototype can be enhanced to a certain level, and also enhance the user interface of the proposed model.

VI. FUTURE WORK

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

REFERENCES

[1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc., pp. 900–903, 2017.

[2] https://indianexpress.com/article/technology/ social/whatsapp-fight-against-fake-news-top- features-to-curb-spread-of-misinformation- 5256782/

[3] P. R. Humanante-Ramos, F. J. Garcia- Penalvo, and M. A. Conde-Gonzalez, "PLEs in Mobile Contexts: New Ways to Personalize Learning," Rev. Iberoam. Tecnol. del Aprendiz., vol. 11, no. 4, pp. 220–226, 2016.

[4] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," CEUR Workshop Proc., vol. 2041, no. 1, pp. 59–66, 2017.

[5] S. Gilda, "Evaluating machine learning algorithms for fake news detection," IEEE Student Conf. Res. Dev. Inspiring Technol. Humanit. SCOReD 2017 - Proc., vol. 2018– January, pp. 110–115, 2018.

[6] S. Das Bhattacharjee, A. Talukder, and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018– January, pp. 556–565, 2018.

[7] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on Twitter," Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018, pp. 274–277, 2018.

[8] P. R. Humanante-Ramos, F. J. Garcia- Penalvo, and M. A. Conde-Gonzalez, "PLEs in Mobile Contexts: New Ways to Personalize Learning," Rev. Iberoam. Tecnol. del Aprendiz., vol. 11, no. 4, pp. 220–226, 2016.

[9] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," CEUR Workshop Proc., vol. 2041, no. 1, pp. 59–66, 2017.

[10] S. Gilda, "Evaluating machine learning algorithms for fake news detection," IEEE Student Conf. Res. Dev. Inspiring Technol. Humanit. SCOReD 2017 - Proc., vol. 2018– January, pp. 110–115, 2018.

[11] S. Das Bhattacharjee, A. Talukder, and B. V. Balantrapu, "Active learning based news veracity detection with feature weighting and deep-shallow fusion," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018– January, pp. 556–565, 2018.

[12] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on Twitter," Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018, pp. 274–277, 2018.