Fake News Detection System

Neel Shah
University of North Carolina Charlotte
Department of Computer Science
nshah70@uncc.edu

Abstract*—The rise of the Internet and use of social media platforms has made the way for information distribution . With the increasing usage of social media, individuals are sharing more information, some of which are deceptive. Automated classification of a text article as deceptive or not is a difficult task. Even an expert has to explore multiple aspects before giving a verdict on the correctness of an article. In this work, I propose a machine learning system for automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, I train a combination of different machine learning algorithms using various ensemble methods and evaluate their performance on real world datasets.

Keywords—Ensemble Learners,Naive Bayes theorem, Logistic Regression Classifier, Random Forest.

I INTRODUCTION

With the advancement in the field of technology and use of multimedia platforms a large number of information is created and shared.Besides other use cases, news outlets benefitted from the widespread use of social media platforms by providing updated news in near real time. Media has evolved from newspapers, tabloids, and magazines to a digital form such as online news platforms, blogs, social media feeds, and other digital media formats. However, such platforms are also used with a negative perspective by certain entities commonly for monetary gain and in other cases for creating biased opinions, manipulating mindsets, and spreading satire or absurdity. The phenomenon is commonly known as fake news.

There has been a rapid increase in the spread of fake news over a period of time. Such fabricated articles are shared online that do not conform to facts has led to many problems.One such area affected by fake news is the financial markets, where a rumor can have disastrous consequences and may bring the market to a halt. A recent example would be rumours of the price of cryptocurrency Dogecoin is about to go up which has led to loss of many and gain of few.

Our ability to take a decision relies mostly on the type of information we accept. Our worldview is shaped on the basis of information we digest. There is increasing evidence that consumers have reacted absurdly to news that later proved to be fake. One recent case is the spread of novel coronavirus, where fake reports spread over the Internet about the origin, nature, and behavior of the virus. The situation worsened as more people read about the fake contents online. Identifying such news online is a daunting task.

Advancement in the field of natural language and processing has led to the availability of various computational techniques that can be used to test the correctness of the article.This techniques are generally based on textual content. There are a number of repositories maintained by researchers that contain lists of websites that are identified as ambiguous and fake. However, the problem with these resources is that human expertise is required to identify articles/websites as fake. More importantly, the fact checking websites contain articles from particular domains such as politics and are not generalized to identify fake news articles from multiple domains such as entertainment, sports, and technology.

News which is published online can be of different categories such as articles, post videos and it is relatively difficult to detect and classify as this strictly requires human expertise. However, computational techniques such as natural language processing (NLP) can be used to detect anomalies that separate a text article that is deceptive in nature from articles that are based on facts . Other techniques involve the analysis of propagation of fake news in contrast with real news. More specifically, the approach analyzes how a fake news article propagates differently on a network relative to a true article. The response that an article gets can be differentiated at a theoretical level to classify the article as real or fake. A more hybrid approach can also be used to analyze the social response of an article along with exploring the textual features to examine whether an article is deceptive in nature or not

In the current fake news corpus, there have been multiple instances where both supervised and unsupervised learning algorithms are used to classify text. However, most of the literature focuses on specific datasets or domains, most prominently the politics domain. Therefore, the algorithm trained works best on a particular type of article's domain and does not achieve optimal results when exposed to articles from other domains.

So I have proposed a system which is a combination of steps in which the first step is data preprocessing then we will count features/tf-idf features after that the features. I have built all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. I have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient descent and Random forest classifiers from sklearn. Each of the extracted features were used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, 2 best performing models were selected as candidate models for fake news classification. Then performed parameter tuning by implementing GridSearchCV methods on these candidate models and chose best performing parameters for these classifiers. Finally the selected model was used for fake news detection with the probability of truth. In Addition to this, We have also extracted the top 50 features from our term-frequency tf idf vectorizer to see what words are most and important in each of the classes. We have also used Precision-Recall and learning curves to see how training and test sets perform when we increase the amount of data in our classifiers.

I.I. LITERATURE SURVEY

Iftikhar Ahmad et al. have compared various computational methods that have used Ensemble Methods.Textual properties were used by them By using those properties, we train a combination of different machine learning algorithms using various ensemble methods.Furthermore, they have outlined the challenges and how such approaches can be useful in solving those [1].

Sawinder Kaur et al. have identified news articles as fake or real by using supervised machine learning classifiers such as Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Linear models, Neural Networks (NN) and Ensemble models. To get effective results, three different corpora (News Trends, Kaggle and Reuters) have been collected with similar characteristics. The feature extraction techniques, Term Frequency–Inverse Document Frequency (TF–IDF), Count-Vectorizer (CV), Hashing-Vectorizer (HV), are used to extract feature vectors from the textual content of articles. [2].

Kai Shu et al. studied the novel problem of exploiting social context for fake news detection. They propose a tri-relationship embedding framework TriFN, which models publisher-news relations and user-news interactions simultaneously for fake news classification. they conduct experiments on two real-world datasets, which demonstrate that the proposed approach significantly outperforms other baseline methods for fake news detection. [3].

Nadia K Conroy et al. have provided a typology of several varieties of veracity assessment methods emerging from two major categories – linguistic cue approaches (with machine learning), and network analysis approaches. We see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network‐based behavioral data. Although designing a fake news detector is not a straightforward problem, we propose operational guidelines for a feasible fake news detection system[4].

II. PROBLEM DEFINITION & ALGORITHM

During the recent years there has been lots and lots of false news spreading on the internet or through different mediums as a result of which this has become a major concern for identification of the false news.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 society to believe the rumors and pretend to be news. The need of an hour is to stop the rumors and focus on the correct, authenticated news articles.If there is a system which can detect the fake news system used it will solve our problems.The algorithms which are going to study and use it to solve the problem are Ensemble Models, Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient descent and Random forest classifiers

II.I. TASK DEFINITION

The model is divided into four tasks which are as follows:

1] Data Preparation

The data source used for this project is a LIAR dataset which contains three files: test,train and validation.LIAR: A benchmark dataset for fake news Detection system.The original dataset comprises of thirteen variables in order to keep things simple only two variables are used other variables can be added later to increase the efficiency of the model.The two columns of the datasets are statement and label. Statement represents the news headline whereas Label represents the class. First we read all the three files then perform some basic text preprocessing like tokenizing,stemming etc.

2] Feature Selection

In this feature extraction and selection methods from sci-kit learn python libraries will be used for feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-idf weighting.

3] Classification

In this step various classifiers would be used and identify the best performing classifier. The extracted features will be fed into different classifiers. I will be using

Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient descent and Random forest classifiers from sklearn. Each of the extracted features were used in all of the classifiers. Then I will also use Precision-Recall and learning curves to see how training and test sets perform when I increase the amount of data in our classifiers.

4] Prediction:

In this stage all things would be done the user has to enter the input in the system and the system will generate the output for the user which will be a class label. The news article headline is either false or true along with the truth probability

II.II ALGORITHMS & MODELS

1] Ensemble models

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and *independent*, the prediction error of the model decreases when the ensemble approach is used. The approach seeks the wisdom of crowds in making a prediction.

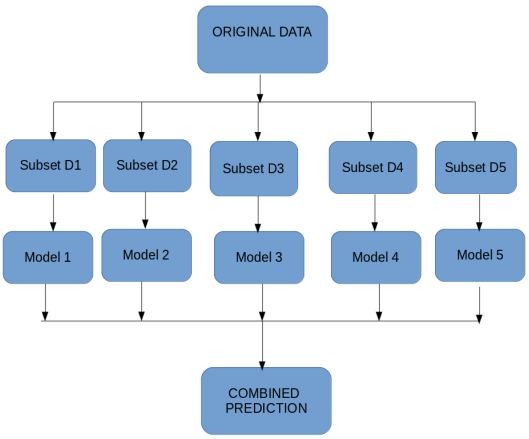

fig 1: Ensemble Modelling example

2]Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

3]Logistic Regression

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.

4]Linear SVM

Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is more preferred for classification but is sometimes very useful for regression as well. Basically, SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line.

In SVM, each data item is plotted in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to use for multi-class problems.

5] Stochastic gradient descent

The word '*stochastic*' means a system or a process that is linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In Gradient Descent, there is a term called "batch" which denotes the total number of samples from a dataset that is used for calculating the gradient for each iteration. In typical Gradient Descent optimization, like Batch Gradient Descent, the batch is taken to be the whole dataset. Although, using the whole dataset is really useful for getting to the minima in a less noisy and less random manner, the problem arises when our datasets get big.

6] Random forest

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.Random Forest has multiple decision trees as base learning models. Randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

III. EXPERIMENTAL EVALUATION

The aim of proposed methodology is accurate prediction of fake news.The methodology of the proposed model is divided into phases as shown in fig 2.

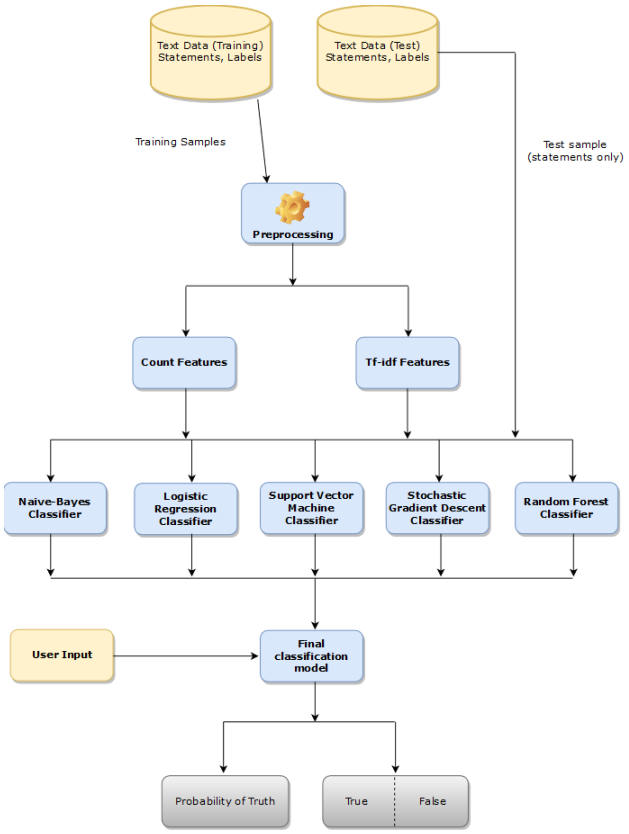

fig 2: Methodology of the model

III.I METHODOLOGY

The model is divided into four steps which are as follows:

1]Data Preprocessing

In this phase the dataset is converted into the desirable data which is needed. In this case thirteen column dataset was converted into two column dataset along with that simple text processing techniques are used and dataset is filtered

2]Feature extraction

At this stage the features which are needed are kept and features which are not needed by the model are discarded. It is a very useful step in reduction of the access data .After the extraction these features are fed to the classifier models.

3]Classification model Selection

The extracted features are used to make the model complete and predict the class of the input. In this all the algorithms are trained with the features and the model which performs the best on the training data is selected and stored.

4]Prediction

Once the model is selected then the prediction takes place in the form of the use case diagram given below fig3.

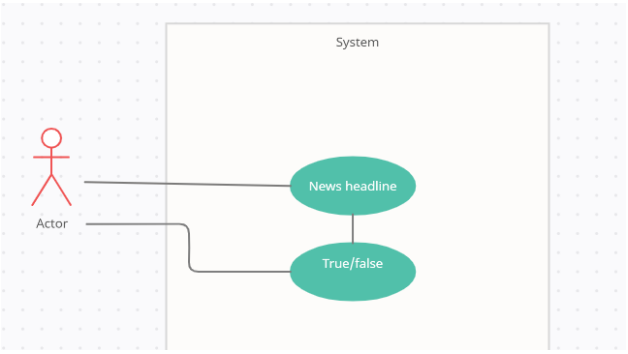

fig3:Use Case diagram

III.II. RESULTS

The results obtained from the machine learning algorithms are listed in table 1.

Table 1: Comparison of performance of Machine learning algorithms Bag of words confusion matrix and F1 scores

Sr No	Algorithm used	Confusion Matrix	F1 score
1.	SVM	[2260 2228] [2246 3506]	0.61
2.	Naive Bayes	[2118 2370] [1664 4088]	0.66
3.	SGD Classifier	[2414 2074] [2042 3710]	0.64
4.	LR	[2252 2236] [1933 3819]	0.645
5	Random Forest	[1821 2667] [1192 4560]	0.70

SVM classifier was implemented and 61.% accuracy is obtained. Naive Bayes algorithm was also implemented on python with an 66% accuracy along with SGD and LR 64% and 64.5% accuracy.Random forest had got the accuracy of 70% In terms of predicting, Random forest tree and LR both were the most accurate. Also for all five algorithms, F1 scores were calculated.

Table 2: Comparison of performance of Machine learning algorithms-n grams & tfidf confusion matrix and F1 scores

Sr No	Algorithm used	Confusion Matrix	F1 score
1.	SVM	[2016 2472] [1524 4228]	0.68
2.	Naive Bayes	[841 3647] [427 5325]	0.72
3.	SGD Classifier	[10 4478] [13 5739]	0.71
4.	LR	[1617 2871] [1097 4655]	0.72
5	Random Forest	[1979 2509] [1630 4122]	0.67

III.III DISCUSSION

The final selected model which had performed the best among all the classifier algorithms is our Logistic Regression classifier. Random forest Tree has also been close in accurately identifying the class labels of the input dataset. The performance of both the models can be seen in the following figures with respect to the learning curve

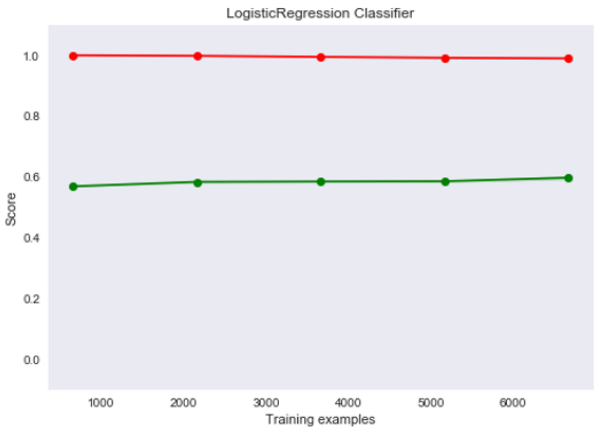

fig 3: Learning curve of LR

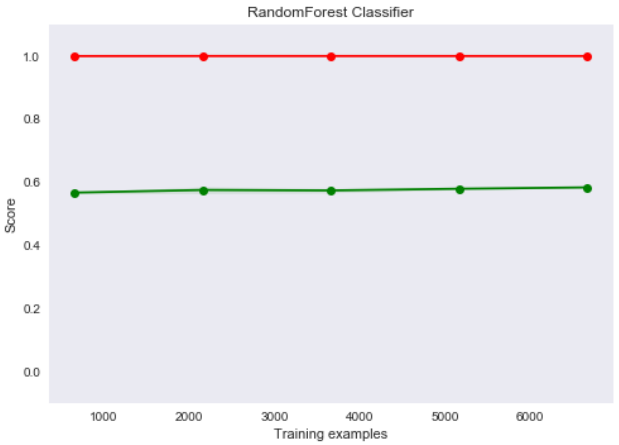

fig 4: Learning curve of Random Forest

The result has less accuracy then as expected it can be improved by increasing the dataset and adding more features into the dataset since I have reduced the dataset inorder to reduce the complexity which has affected the result. Ensemble Modelling can be very helpful when using multiple models as it can be used to improve the accuracy of the model since it uses all the models all together to predict the class of the data.

IV RELATED WORK

The papers which were surveyed found that most of the systems which were initially developed only used a single model to predict the probability of the news being fake whereas the model which is developed in this case uses various algorithms and selects the one with the highest accuracy which helps to predict betterly. Ensemble modeling was also observed while reviewing other papers. It is a better method than the method used by me in this case because it runs all the models simultaneously and then predicts instead of selecting one model which has proved to be of great value but it increases the overall complexity of the model.However, model ensembles are not always better. New observations can still be confusing. That is, ensembles cannot help unknown differences between sample and population. Ensembles should be used carefully.

V. FUTURE WORK

From the above study, it is clear that detecting fake news is possible but in order to correctly identifying is difficult As observed that our best performing models had an f1 score in the range of 70's. This is due to less data being used for training purposes and simplicity of our models. For the future implementations,Introduction of some more feature selection methods such as POS tagging, word2vec and topic modeling. In addition, increase the training data size.The void of better prediction can also be removed by doing ensemble modelling which will help weak models to rightly classify the class label. Real time fake news identification in videos can be another possible future direction.

VI CONCLUSION

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research,I have discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles and used the feature set as an input to the models. The learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others. We used multiple performance metrics to compare the results for each algorithm..Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in the spread of fake news.

VII. REFERENCES

[1] Iftikhar Ahmad, Muhammad Yousaf,[1]Suhail Yousaf and Muhammad Ovais Ahmad “Fake News Detection Using Machine Learning Ensemble Methods” Published online 2020 Oct 17. doi: 10.1155.

[2] Sawinder Kaur,Parteek Kumar,Ponnurangam Kumaraguru “ Automating fake news detection system using multi-level voting model” Published online 2019 Nov 2.

[3] Kai Shu, Suhang Wang, Huan Liu “ Beyond News Contents: The Role of Social Context for Fake News Detection ” Published online doi.org/10.1145/3289600.3290994 January 2019.

[4] Nadia K Conroy, Victoria L Rubin, Yimin Chen “Automatic deception detection: Methods for finding fake news” BMC Bioinformatics. 2019; 20(Suppl 7): 194. Published online doi.org/10.1002/pra2.2015.145052010082 24 February 2016.

[5] Fake News Detection System. Accesses on: May 9, 2021. Available:https://github.com/nishitpatel01/Fake_News_Detection.

[6] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers.

[7] Geeksforgeeks Accessed on May 10 2021. Available:https://www.geeksforgeeks.org/random-forest-regression-in-python/.