Fake News Detection using Machine Learning and Natural Language Processing

Deepankar Kansal (MT20007)
Vineet Maheshwari (MT20020)
Palak Tiwari (MT20103)

Abstract

Fake news are misleading news stories that come from non-reputable sources which is having a significant impact on our social life. Fake news detection is an emerging research area which is gaining interest but involved some challenges due to the limited amount of resources (i.e., datasets, published literature) available. We propose in this paper, a fake news detection model that use *'nltk'* and *'re'*, *glove embedding* and *bag of words* feature extraction techniques analysis and machine learning techniques like *Logistic regression, Support Vector Machines, Naive Bayes, Gradient Boosting and Bidirectional LSTM (a variant of RNN)*. Experimental evaluation yields the best performance using *'bag of words'* as feature extraction technique, and Gradient Boosting as a classifier, with an accuracy of 96% on appending the title of the news with the text part of the news for creating the corpus. We have performed tests on three variants of datasets and observe that Gradient Boosting classifier performs better for all the three datasets and gives the best accuracy.

1 Introduction

In this digital era there are multiple news that are floating on the internet and there is no substantial proof to claim whether a particular news is *fake/unreliable* or *real/reliable*. Fake news verification aims to employ technology to identify intentionally deceptive news content online. Also Social media have the capacity of news delivery mechanisms on a mass scale yet much of the information is of questionable veracity. In this paper, we seek to produce a model that can accurately predict the likelihood that a given article is fake or not.

Today anyone can publish content which is credible or not, that can be consumed by the world wide web. Sadly, fake news accumulates a great deal of attention over the internet, especially on social media. People get deceived and don't think twice before circulating such mis-informative pieces to the far end of the world. This kind of news vanishes but causes the harm it is intended for. The given social media sites that play a major role in supplying counterfeit news include Facebook, Twitter, Whatsapp etc.

Even though the problem of fake news is not a new issue, detecting fake news is believed to be a complex task given that humans tend to believe misleading information and the lack of control of the spread of fake content. Fake news has been getting more attention in the last couple of years, especially since the US election in 2016. It is tough for humans to detect fake news. It can be argued that the only way for a person to manually identify fake news is to have a vast knowledge of the covered topic. Even with the knowledge, it is considerably hard to successfully identify if the information in the article is real or fake.

We present in this paper 'bag of words' and glove embedding feature extraction based approach to detect fake news, which consists of using text analysis based on these library's features and machine learning classification techniques. We study and compare different classification techniques, namely, Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes's variant Multinomial Naive Bayes, Gradient Boosting and Bidirectional LSTM (a variant of RNN). Experimental evaluation is conducted using the datasets compiled from 'kaggle' website, yielding very encouraging results.

2 Importance of the project

Detection of fake news online is important in today's society as fresh news content is rapidly being produced as a result of the abundance of technology that is present. The piece of fake news content

can be *visual or linguistic-based*. In order to detect fake news, both linguistic and non-linguistic cues can be analyzed using several methods.

Linguistic Approaches in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception. Linguistic-based type of fake news is in the form of text or string content and generally analysed by text linguistics. Its content largely focuses on text as a communication system and includes characteristics like tone, grammar, and pragmatics that allows discourse analysis. Examples of linguistic-based platforms are blog sites, emails and news sites. Blog sites are managed by users and the content produced is unsupervised which considers it easy to receive wrong information. Email is another medium where its users can receive news and this poses as a challenge to detect and validate their authenticity. Popular news websites too can generate their own content and attract users with their authentic presence.

Most liars use their language strategically to avoid being caught. In spite of the attempt to control what they are saying, language 'leakage' occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage. The goal in the linguistic approach is to look for such instances of leakage or, so called 'predictive deception cues' found in the content of a message. To detect such fake news following strategies will help a lot:

- **Perform an extensive feature set evaluation study** with the purpose to lead in an effective feature set to detect fake news articles.

- **Perform an extensive Machine Learning (ML) classification algorithm** benchmarking study, such classification study contains Logistic Regression (LR), Support Vector Machine (SVM), Boosting techniques like Gradient Boosting and many more algorithms. Also the neural networks like RNN and its variations.

- **Unbiased dataset for Fake News detection,** creating the Unbiased dataset containing a balanced number of fake and real news articles. We shall call a dataset unbiased if it contains a balanced distribution of articles from different categories and from different news sources.

- **Quality results in fake news detection.** Our experimentation has shown that our approach can achieve accuracy up to 96%.

3 Literature Survey

[1] Predicted fake news using Naïve Bayes classifier, Support Vector Machine and Logistic Regression, got maximum accuracy of 83 percent on using Naïve Bayes classifier with lidstone smoothing, this model which will judge the counterfeit news articles. This shows that Naïve Bayes classifier is both simple and powerful for Natural Language Processing tasks such as text classification problems. Used NLTK for preprocessing on kaggle dataset. Performed experiments on different feature set that is body, body with headline and headline, got best accuracy with body with headline.

[2] Works on fake news detection using ensemble machine learning algorithms such as AdaBoost and Bagging. There experimental results show that the use of an enhanced linguistic feature set with word embeddings along with ensemble algorithms and Support Vector Machines (SVMs) is capable to classify fake news with high accuracy of 95%. They mainly have used SVMs, Naive Bayes, Decision Tree, KNN, AdaBoost and Bagging ensemble machine learning algorithms, but the best results and accuracies achieved with SVMs and ensemble machine learning algorithms. For preprocessing they have used word embeddings and linguistic feature sets that helps in better feature extraction. They have obtained accuracy up to 95% on over all datasets used with the AdaBoost to be first in rank and SVM Bagging algorithms to be next in ranking but without statistically significant difference.

[3] Shlok Gilda outline several approaches that seem promising toward the aim of correctly classifying misleading articles. The paper implements two preprocessing steps namely term frequency-inverse document frequency(TF-IDF) on bigrams and probabilistic context free grammar(PCFG) on a corpus of 11000 dataset obtained from Signal Media and tests it on various classification algorithms-Support Vector Machine, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees, and Random Forests. They obtain that the TF-IDF of bigrams fed into Stochastic Gradient Descent model identifies the non-credible sources of news(i.e., fake news) with an accuracy of 77.2%, with PCFGs having slight effect on the recall.

[4] Conroy, Rubin, and Chen's paper provides

a typology of several varieties of veracity assessment methods emerging from two major categories – linguistic cue approaches (with machine learning), and network analysis approaches. They see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network-based behavioral data. This paper basically proposes operational guidelines to build a feasible fake news detecting system.

[5] N-gram features based approach to detect fake news, which consists of using text analysis based on n-gram features and machine learning classification techniques. Experimented and compared six different supervised classification techniques, namely, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Linear Support Vector Machine (LSVM), Decision tree(DT) and Stochastic Gradient Descent (SGD). Experimental evaluation is conducted using a dataset compiled from real and fake news websites, yielding very encouraging results, with an accuracy of 92% as the best performance using Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction technique, and Linear Support Vector Machine (LSVM) as a classifier.

[6] Presented different algorithms for classifying statements made by public figures, were implemented. In proposed system Logistic regression (LR), Naïve bayes (NB), Support vector machine (SVM), Random forest (RF) and deep neural network (DNN) classification techniques are utilized that will help to detect fake news. Classification techniques like LR, RF, SVM NB and DNN for feature selection and extraction utilized, DNN will work fine in execution time and accuracy cases but it needs large memory than other. Then compared Logistic regression (LR), Naïve bayes (NB), Support vector machine (SVM), Random forest (RF) and deep neural network (DNN) on basis in terms of time and memory and accuracy, according to comparison results it exhibit that DNN Algorithm is improved than rest algorithm in accuracy and time kind because rest classifiers requires more time and gives less accuracy hence DNN is more crucial to detect the fake news.

4 Data Analysis and Feature Generation

4.1 Datasets Description

One of the major challenge is to find the dataset which contains real and fake news proportionally (that is *Unbiased dataset*) as dataset plays a very important role in making the model more accuracte by creating a corpus which has a variety of words in it. Few of the major points that we have to look which searching for dataset is as follows:-

- Availability of both truthfulness and deceptiveness instances.
- Verifiability of 'ground truth'.
- Homogeneity in lengths.
- Homogeneity in writing matter.
- The manner of delivery (i.e., how the article has been written).

Getting the correct information implies assembling or distinguishing the information that relates with the results which needs to be foreseen; for example information that contains a flag about occasions which needs to be taken care about. The datasets should be lined up with the issue which is being attempted to explain. Keeping these challenges in mind we have taken three datasets from Kaggle which was released for some competition and contains mostly political news of United States. These news articles were collected from various social media platforms and was then flagged accordingly based on its authenticity. Dataset1 which is unbiased contains 20800 samples with 10413 Fake and 10387 Real news samples. It has 5 columns id,title,text,author and label. Similarly Dataset2 is also unbiased with a total of 44000 data samples(approx) and Dataset3 is biased with 12000 data samples(approx) and will give low accuracy. The 3rd dataset that we are using contains more number of fake news instances than real news.

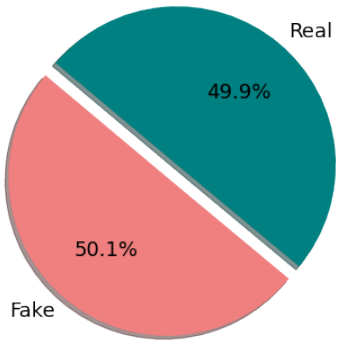

Figure 1: Fake and Real news article count for 1st dataset

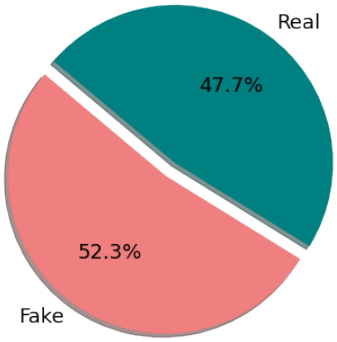

Figure 2: Fake and Real news article count for 2nd dataset

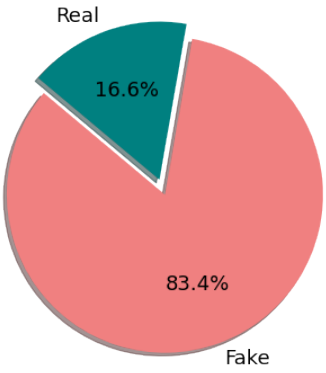

Figure 3: Fake and Real news article count for 3rd dataset

4.2 Preprocessing

Machine Learning project requires information in legitimate way which has to be extracted through a valid dataset. For better outcomes through the Machine Learning model preprocessing is essentially required on the dataset.For applying the models the given data needs to be vectorized before which we refine the data by tokenization, removing stopwords, removing punctuations and converting to lower case. This preprocessing helps in reducing the data to relevant data by removing irrelevant information.

Stopwords are words that are the commonly used words in language which helps in structuring the sentence, mostly include articles, pronouns, prepositions and conjunctions. Example is, are, the, on, by, for, not, of, how. Here 'not' is also a stopword but for Fake news detection we require the use of 'not' so it has to be excluded from the stopwords.

The given dataset was split into 70%:30% for train and test, which was in csv format thus required preprocessing. Each sample corresponds to a news article body. We used stopwords from NLTK. Firstly using 're' package we find all the tokens preprocess them then using PorterStemmer we perform stemming, stems are created by removing the suffixes or prefixes used with a word.

After stemming the strings we applied *bag of words feature extraction technique* using CountVectorizer, in which we have given the input to create a Unigram and Bigram combinations on which, this class has applied bag of words technique later. For Bi-LSTM we have used Glove embedding for converting the words into their embeddings and after that using these as embedding matrix we trained the Bi-LSTM model.

4.3 Train and Test Splitting

Machine learning regularly works with two informational collections: *training and test*. The principal set which is being used is the training set, the biggest of the two. Running a training set through a machine learning algorithm shows how to weigh diverse highlights and changing them into coefficients as per their probability, to predict better on the test set. These coefficients, also called parameters,will be contained in tensors and together they are known as the model, since it encodes a model of the information on which it is being trained.

Other dataset on which we check how our trained model is working and what is the efficiency of our model, is called Test set. Or we can say test set is like sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

We have achieved best accuracies by splitting train:test to ratio 70%:30%.

5 Evaluating Model

Classification models applied are:

- Gradient Boosting (GBC)
- Logistic Regression (LR)
- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Bidirectional LSTM

Gradient Boosting Classifier (GBC)

Gradient boosting belongs to the family of powerful boosting machine learning algorithms that have demonstrated tremendous success in the form of accuracy. It is a sequential ensemble learning

technique where the performance of the model improves over iterations. This method creates the model stage-wise, it tries to reduce the loss by adding decision trees. Gradient boosting is a highly robust technique for developing predictive models.

Logistic Regression (LR)

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. The LR model uses gradient descent to converge onto the optimal set of weights (θ) for the training set.

Naive Bayes (NB)

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. Naive Bayes is a simple but powerful algorithm for predictive modeling. It is mostly used in natural language processing (NLP) problems. Naive Bayes predict the tag of a text. They calculate the probability of each tag for a given text and then output the tag with the highest one.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression purposes. SVMs are mostly used in classification problems. SVMs are founded on the idea of finding a hyperplane that best divides a dataset into two classes. Sometimes linear SVM would work fine, but sometimes we might need *kernel tricks* to improve the accuracy by adding more dimensions to the providing dataset features.

Bidirectional LSTM

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as-is and the second on a reversed copy of the input sequence. This can provide additional context to the network and result in faster and even fuller learning on the problem.

We have applied various techniques and variation for training our model with different corpus of data. We have basically trained our model on based on the corpus created only using the title of the news, then only using the text of the news and finally on both the title and the text of the news for all the three datasets that we are using. We see that the title and text combination used for creating the corpus and training the model results in the best accuracy for all the three datasets. Below is a plot in which depicts the three variation of creating the corpus and the accuracies obtained on training a Logistic Regression model on it for the three datasets:

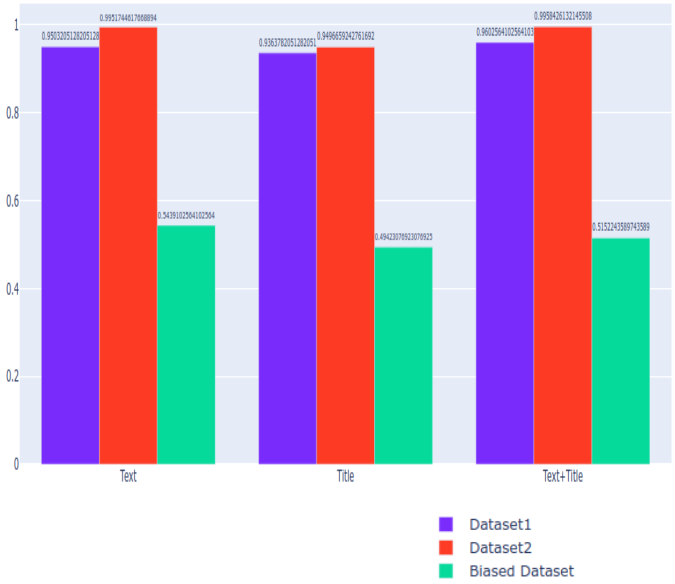

Figure 4: Comparison of accuracies based on creating the corpus using the title, the text and both title and text

Now let's compare the accuracies achieved by the above mentioned classification models on all the three datasets by building the corpus for each of the model using both the title and the text of the news in Figure 5, 6 and 7. From the figures we can clearly see that gradient boosting performs very well and has the best accuracy for all the three datasets that we have used with the train and test split ratio of 70:30.

6 Results and Conclusion

Using the above-mentioned algorithms, i.e. Naïve Bayes classifier, Support Vector Machine and Logistic Regression, Gradient Boosting and Bidirectional LSTMs, we attained the accuracies as shown in tables below. The maximum accuracy obtained is using Text+Title with Gradient Boosting Classifier. LR and Gradient Boosting were giving almost similar accuracy, SVM and Bidirectional LSTM were little less, however accuracy obtained from Naive Bayes was relatively lesser than other models. LR and Gradient

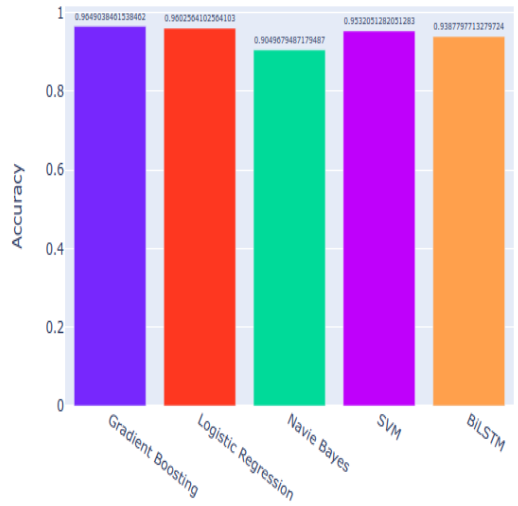

Figure 5: Classification model's accuracy comparison for dataset 1

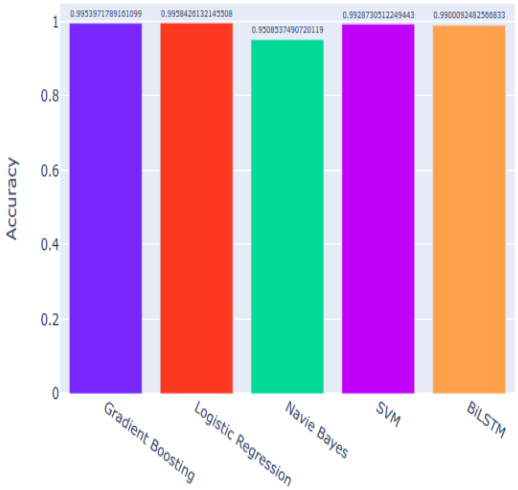

Figure 6: Classification model's accuracy comparison for dataset 2

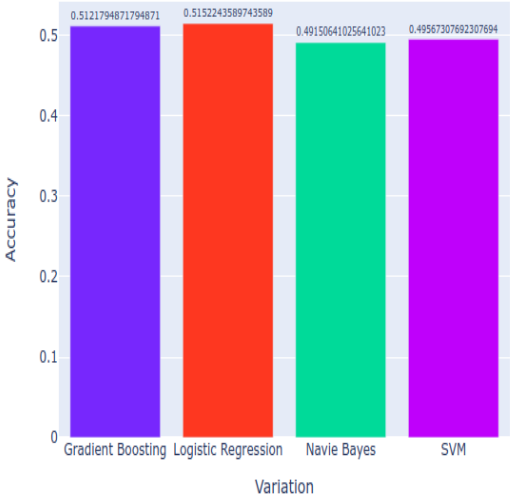

Figure 7: Classification model's accuracy comparison for dataset 3

Boosting were performing better because features that we are extracting from the corpus are well separated.

Table 1: Accuracy comparison based on training the models on various corpus

Dataset on LR	**Text**	**Title**	**Text+Title**
Dataset1	95.03%	93.63%	96.02%
Dataset2	99.51%	94.96%	99.58%
Dataset3	54.39%	49.42%	51.52%

Table 2: Accuracy comparisons on different model trained on corpus of Text+Title for Dataset1

Models on **Dataset1**	**Accuracy using text+title**
Logistic Regression	96.02%
SVM	95.32%
Naive Bayes	90.5%
Gradient Boosting	96.5%
Bidirectional LSTM	93.87%

Table 3: Accuracy comparisons on different model trained on corpus of Text+Title for Dataset2

Models on **Dataset2**	**Accuracy using text+title**
Logistic Regression	99.6%
SVM	99.28%
Naive Bayes	95.1%
Gradient Boosting	99.6%
Bidirectional LSTM	99%

Table 4: Accuracy comparisons on different model trained on corpus of Text+Title for Biased Dataset

On Biased Dataset	**Accuracy using text+title**
Logistic Regression	51.5%
SVM	49.6%
Naive Bayes	49.15%
Gradient Boosting	51.22%

7 Future Work

In the future, a possible improvement would be implementing TF-IDF on bi-grams and n-grams as one of the feature extraction method and also we can use various embedding techniques such as word2vec and BERT so that we can improve the performance of the model even more and we will be able to build a more robust model which can

also take care of the outliers in the dataset. We can also apply other machine learning algorithms like *KNN, ensemble learning methods like Bagging* with larger datasets. In that way, the fake news detection task would not only be content-based but would also improve the prevention of their spread in social networks. We can also look forward to build a robust model which can also work on the native launguages of a particular state and flag those news as fake or real.

8 References

1. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing", International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-7, Issue-6, March 2019.
2. Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, Panagiotis Karadais, "Behind the cues: A benchmarking study for fake news detection", Expert Systems with Applications Volume 128, 15 August 2019, Pages 201-213.
3. Shlok Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection", IEEE 15th Student Conference on Research and Development (SCOReD), December 2017, pp.110-115.
4. N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news", Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.
5. Hadeer Ahmed, Issa Traore, and Sherif Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques", ECE Department, University of Victoria, Victoria, BC, Canada meresger.hs@gmail.com, itraore@ece.uvic.ca 2 School of Computer Science, University of Windsor, Windsor, ON, Canada Sherif.SaadAhmed@uwindsor.ca
6. Chaitra K Hiramath, Prof. G.C Deshpande, "Fake News Detection Using Deep Learning Techniques", CSE Department, KLS Gogte Institute of Technology, Belagavi,Karnataka,India,

9 Link to the Drive

Link to the Codes, Report and PPT uploaded in drive