Fake News Detection from Online media using Machine learning Classifiers

Shalini Pandey¹,⁴, Sankeerthi Prabhakaran¹,⁵, N V Subba Reddy² and Dinesh Acharya³

¹MTech CSIS, Department of Computer Science and Engineering, MIT, MAHE, Manipal
²Department of Computer Science and Engineering, MIT, MAHE, Manipal
³Department of Computer Science and Engineering, MIT, MAHE, Manipal
⁴shalini.pandey@learner.manipal.edu, ⁵sankeerthiprabhakaran.19@gmail.com
²nv.reddy@manipal.edu, ³dinesh.acharya@manipal.edu

Abstract: With the advancement in technology, the consumption of news has shifted from Print media to social media. The convenience and accessibility are major factors that have contributed to this shift in consumption of the news. However, this change has bought upon a new challenge in the form of “Fake news” being spread with not much supervision available on the net. In this paper, this challenge has been addressed through a Machine learning concept. The algorithms such as K-Nearest Neighbor, Support Vector Machine, Decision Tree, Naïve Bayes and Logistic regression Classifiers to identify the fake news from real ones in a given dataset and also have increased the efficiency of these algorithms by pre-processing the data to handle the imbalanced data more appropriately. Additionally, comparison of the working of these classifiers is presented along with the results. The model proposed has achieved an accuracy of 89.98% for KNN, 90.46% for Logistic Regression, 86.89% for Naïve Bayes, 73.33% for Decision Tree and 89.33% for SVM in our experiment.

1. Introduction

In today’s world various developments in the technology have led to nuance that “Data is the oil” of the 21st century. There has been a drastic shift in the source of News consumption from Print media to Social media. As a support to this statement, it can be seen that in the year 2013, News was consumed at 63% on Print media and 18% on social media and the same statistics in April 2020 have resulted in, Print media’s contribution declining to a rate of 26% whereas the Social media’s has risen up to 39%. With increase in Social media news consumption, the proliferation of Fake news is becoming an issue.

At its simplest fake news can be described as false stories that are fabricated in order to influence public opinion or defame a Person. It has also been recorded that fake news receives more views than real one’s on social media and supporting this claim on the famous social networking platform “Facebook” 20 fake news showed more involvement of its user’s compared to the top 20 real news stories. It’s observed that the features such as sharing, commenting and tagging a friend in a post have aided in spreading of these news largely in Social media.

Various steps have been taken to control this issue and one way is to distinguish these and stop their spreading. There have been studies proposed earlier which make use if the Machine learning concepts to take down these news articles, such as in paper [1] KNN classifier has been proposed to label the news as fake or real, however due to the nature of the text data available on net this technique has not resulted with credible accuracy.
Hence in the paper, this issue of handling various text formats available on net, so as to increase the efficiency of the Classifiers. A few Data pre-processing steps such as Stemming, stop words and lemmatization are discussed which shall help to refine the data before it’s fed to the Classifiers.

The Classifiers considered for now are KNN and Logistic Regression as they are easy to interpret and can handle noise in data in a better way. The Dataset considered for this project consists of 6335 articles, 30% used for training and 70% for testing the model. The next sections are divided as Methodology, Observation and conclusion.

2. Related Works
In paper [3] the authors have explored ways to increase the efficiency of KNN algorithm so that it can give out better results. The evolutionary Genetic Algorithm is used to select the finest parameters of the nonlinear functions that are suitable for each feature, and the results are better comparatively and on similar lines in paper [4], Preeti Nair and Indu Kashyap have made implored that by introducing resample technique and Inter quartile range technique (IQR) in the pre-processing steps the data fed to classifiers are normalized which gives out better working of the algorithm.

In paper [5], authors created a fake news detection model based on headlines, as well as data on user social site traffic.

In paper [6], K. Nagashri and J. Sangeetha in order to identify fake news have used the count vector techniques and made use of several Machine learning concepts and evaluated them on the basis of accuracy, precision, recall, and F1 score and concluded that TFIDF is a better text preprocessing technique.

In paper [7], The authors attempted to discover the relationship between the words and the context in which they appear within the text, as well as how it could be used to classify texts as genuine (negative cases) or fictitious (positive cases). They made use of models such as Count Vectorize to convert character-based texts into numeric representations and investigated which model is capable in determining the text as real or fake.

In paper [8], Shlok Gilda has made use of term frequency-inverse document frequency (TF-IDF) of bi-grams and probabilistic context free grammar (PCFG) detection and applied to a collection of around 11,000 articles. Machine learning classifier algorithms such as Random Forests, Gradient Boosting, Stochastic Gradient Descent are used to identify the fake news. They have received an accuracy of 77.2%.

3. Models
3.1 KNN classifier
KNN is a supervised machine learning algorithm used for the classification and regression problems. It’s fed a group of labeled input data so that it can develop a function that shall determine and give the label for the given unclassified input. It works on principal of “Nearest neighbors”. K is the value of nearest neighbor’s a group can have. For example, let’s consider a group of data points plotted on a graph with co-ordinates (x, y) values in each instance space; On receiving a new input the algorithm will determine its label according to the nearest neighbor. Since the value of ‘K’ plays a huge role in classification, the optimum value for this is decided by performing various trial and error methods, the most popular method being ‘elbow method’.
3.2 Logistic Regression
Logistic regression is another Supervised Machine learning algorithm which is used for data that is co-dependent on each other, such as heads or tails. It is used to capture the relation between the binary valued data and convert that into a function based on one dependent variable and one or more independent variables.

3.3 Support Vector Machine
Furthermore, Support vector machine algorithm that gives out accurate classification of linear data. In case the given data is of non-linear form we can make use of kernel trick to avoid complex transformations of dimensions [5] into a linear model. This algorithm develops a hyper plane in N-dimension space (dependent on the dimension of input fed) during the training period of the model. This hyper plane is determined as the boundary line between different groups labeled in the given corpus. This line also selected in such a way that it has the maximum distance from the data points of each group and helps in categorizing the new input.

3.4 Naïve Bayes
This classifier is mainly suitable for contextual data and hence has been used, as it suits the nature of the dataset used. It is built on the principle of Bayes theorem. The multinomial naïve bayes classifier has been used to determine the category of document and make the prediction based on the regularity of words in the file.

3.5 Decision Tree
In our everyday life each task consists of various decisions taken on availability of resources. This concept is incorporated by this algorithm. A tree with every internal node representing a decision taken and the leaf nodes representing the outcome of decision is built and the output predicted by traversing through the tree for most likely outcome.

3.6 Natural Language Processing
The pre-processed data is converted into numeric value in the form of vectors. There are pre-training algorithms (Word2Vec) available in the NLP toolkit, which has been utilized in the project. All the alphabet and special characters are converted into vectors for better performance of the above discussed classifiers.

4. Methodology
The NLTK toolkit which contains libraries set and many programs oriented to NLP is utilized. Even the algorithms of machine learning for clustering of data, regression and its classification i.e., Scikit learn have been imported. These three libraries are important factors in the program which is designed in combination with others libraries such as SciPy and NumPy.

The dataset has been collected from GitHub repository. After getting the dataset, methodology is built in three phases: the first phase is of data pre-processing, this elaborates the changing of datasets from .csv file to a python object that belongs to Pandas to define data frames which shall help in handling the data more proficiently.
In the subsequent phase the data is divided into two data frames, one being labeled as false and other one as true based on the information known beforehand. In the later phase, tokenization algorithms have been performed on these data frames to get clean data which is further divided into training and test datasets and fed to supervised algorithms belonging to the Scikit Learn package to achieve an array which helps us to analyze the accuracy of the classifiers.

Figure 1: Methodology Process

In this project, the usage of Natural language Processing (NLP) has been done for computational tool; For natural language processing and analysis PANDAS library has been used.

4.1 Datasets
The dataset used to implement our model has been taken from the Github’s public repository https://github.com/GeorgeMcIntire/fake_real_news_dataset which consists of news articles (6553- English language). Each article’s features include title, content along with it being labeled as true or false news. Most of the data references are from the American news i.e., from New York Times.

4.2 Data Pre-processing
By the nature of the dataset used, it contains a lot of noise being a natural language. To make the data apt for the algorithms to work on, it is undergoes various computations. Data normalization is a necessary pre-processing step. To begin with clearing of the data, the identification of punctuation marks and stop-words is done followed by their removal. Then the data is tokenized and converted into lower case by calling a function [9] to remove imbalance among them. This process shall shorten the dataset by removing the unnecessary data.

4.2.1 Removal Stop words. Stops words are basically words that add value to other words or define a relation between words. They can widely include adjectives, adverbs, prepositions, conjunctions and determiners. Since our dataset consists of various article, it is imminent to remove these stop words before the data is fed as input to the classifiers. For instance these words include a, an, another, nor, but, or, towards, yet, in etc. After eliminating them from our data corpus, we get reduced distinct words are the output [9].

Figure 2: Pre-processing the dataset
4.2.2 Stemming. This is the next process in normalization of text which is to convert the tokens to their equivalent basic/root words. This process is referred to as Stemming. It is used to reduce the forms of words in data. Stemming does this by changing the fix of words. Snowball Stemmer Algorithm has been adapted in this model as it works better than portal stemmer. It converts words like extreme, extremely to extreme, minister changes to minist in the data set. In the dataset the word ‘secretory’ was most commonly used and hence this algorithm was applied mostly on this word.

4.3 Word2vec
Later the cleaned, tokenized data has been converted into vector form using the word2vec technique. This technique was introduced by Mikolov et al. in 2013 and has demonstrated to be quiet efficient. It is a neural network structure that is used in supervised learning for word embedding. The model is trained on a set of data to develop function so as to identify similar words. This is done by repeatedly updating weights by forward and backward propagation, post which it becomes capable to detect synonyms or suggest additional words. A distinct vector shall be assigned to each word in the corpus and these vectors are decided by performing simple mathematical functions indicating the level of semantic similarity between the words represented by those vectors.

The training data consisted of news articles where in each word had its own contextual meaning which has been embedded using word2vector to its numeric equivalent.

The function is minimum number of times a word has been repeated in the text and their mean is calculated. Since it is better to have the array list of similar words to be mapped to similar vectors, the model is trained on pre-existing Google models so that the word2vec algorithm can give better results. The sentences less than mean length was eliminated on an assumption that they don’t have much reference in the article [7].

In the given dataset 300 features have been considered. Every word in the sentence is transformed to a vector and those vectors belonging to word2vec model are summed up. Then the data is normalized by dividing the obtained value in previous step by the number of words present in that particular sentence.

4.4 Visualizing using TSNE
T-Distributed stochastic neighbor embedding (t-SNE) reduces the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding. In this way, t-SNE maps the multi-dimensional data to a lower dimensional space and attempts to find patterns in the data by identifying observed clusters based on similarity of data points with multiple features. However, after this process, the input features are no longer identifiable, and you cannot make any inference based only on the output of t-SNE. Hence it is mainly a data exploration and visualization technique.

Perplexity (default: 30): In the given graph there are two dimensions mentioned as Dim1 and Dim2 where the perplexity is related to the number of nearest neighbors that are used in other manifold learning algorithms. Consider selecting a value between 5 and 50. Here the value is 50.
5. Experiment Analysis

As five classifiers have been implemented, their performance on how well they were able to classify the given article set is compared. For this purpose we have made use of Confusion matrix. A confusion matrix displays the number of misclassification and correct classification made by the model. The result observed in terms of confusion metrics.

Considering the fake news being classified as positive by the classifier there are 4 possible sections which are discussed below:

- The top left section labels the articles that have been correctly classified as fake, referred to as True Positives.
- The bottom left section labels the articles that have been incorrectly classified as fake news, referred to as false positive.
- The bottom right section labels the articles that have been correctly classified as true news, referred to as true Negative.
- The top right section labels the articles that have been incorrectly classified as real news, referred to as True Positives.

In dataset, though the average accuracy of logistic regression is higher than any other classifier used.
Figure 4: Confusion matrices of different Classifiers
As observed in the given confusion metrices for respective classifiers, the number of misclassified data is low which makes it good to be implemented practically on large datasets.

6. Results
After implementing the machine learning algorithm, the accuracy of each classifier is estimated. It can be observed that all the classifiers have accuracy above 80% [2] except Decision Tree. The following matrix shows the fake news detection without normalization. Depending on the Classifiers or techniques used to change the data into vectors, varied results are obtained.

- In Matrix 1, for the optimal solution of K, elbow method has been used. An odd list of the values of k for KNN till the range of 0–50 is created and null list for cv score. Here the defined K fold value is 10 to reach the optimal value. In the given graph, the point where there’s a drastic drop among all other points considered to be best for value of K is chosen. So here, for the value of k=5 there are least misclassified articles, as shown in figure:

![Image](image.png)

**Figure 5:** Misclassification error for different values of k

- In Matrix 2, SVM algorithm has been used. As the 70% of data has been used for training, the accuracy of remaining 30% of the test data is estimated. Firstly the accuracy is estimated on the basis of hyperparameter than later on pipelining approach which is implemented by grid search whose motive is to reduce the overfitting of the data. Then the accuracy is found out by standardizing the column and same accuracy result is achieved as we got in default hyperparameter. Below the classification report by use of grid search is displayed, which is elaborating about the datasets which comes in the hyper plane.
• In Matrix 3, Logistic Regression algorithm has been used. As the 70% of data has been used for training, the accuracy of remaining 30% of the test data is estimated. In this approach, we tried to find the accuracy by default hyperparameter. It was nearly approximate after applying the regression method over it. Hence the factors which define the optimal work of algorithm have reached 93% as shown .Below is the classification report with accuracy of LR classifier.

![Classification Report](image)

**Figure 6**: Classification Report of SVM Classifier

![Classification Report](image)

**Figure 7**: Classification report of Logistic Regression classifier

• In Matrix 4, Naïve Bayes algorithm has been used. As 70% of data has been used for training, the accuracy of remaining 30% of the test data is estimated. The given data is already vectorized. To achieve the best results, the negative vectors have been scalarized and then the classifier was
implemented on the given set of \((1877, 100)\). As per the given classification report precision value of naïve bayes is lower than other learning models.

![Classification Report](image)

**Figure 8:** Classification Matrix of Naïve Bayes

- In Matrix 5, Decision Tree algorithm has been used. As 70% of data has been used for training, the accuracy of remaining 30% of the test data is estimated. In this approach the accuracy was predicted by using default method. On comparison with previously used classifier’s this one had the lowest accuracy as shown below.

![Classification Report](image)

**Figure 9:** Classification Matrix of Decision Tree

The positive predictive value (precision) of the model represents the appropriate text among the reposessed text documents, whereas sensitivity (recall) is the fraction of total amount of related text documents that were actually retrieved. Hence there is also graph which defines the comparison between these supervised learning algorithms. On the basis of the accuracy it can be estimated which classifier will work efficiently on detection of the news.
7. Conclusion
Talking about the objective, such as the classification of news is a complex task even with using the techniques of classifiers since the input data is in text format and the news has a large number of characteristics that need to be considered. In our paper this complex issue has been addressed with the help of the classifiers that have achieved an accuracy of 89.98% for KNN, 90.46% for Logistic Regression, 86.89% for Naïve Bayes, 73.33% for Decision Tree and 89.33% for SVM.

By using Word2vec it is observed that processing of text for computation is time consuming. Apart from this it’s easier to execute the the classifiers with a good accuracy report. Because of high consumption of RAM and disk, usually Word2Vector is not recommended however it gives semantic relation for processing data into vectors. This project can be further extended as a practical application that would be ready to take any input irrespective of language and determine if it’s fake or real.

8. References
[1] Agudelo, G.E.R., Parra, O.J.S. and Velandia, J.B., 2018, October. Raising a model for fake news detection using machine learning in Python. In Conference on e-Business, e-Services and e-Society (pp. 596-604). Springer, Cham.
[2] Choudhary, P., Pandey, S., Tripathi, S. and Chaurasiya, S., 2021. Fake News Detection Based on Machine Learning. In Advances in Smart Communication and Imaging Systems (pp. 67-75). Springer, Singapore.
[3] F. Sanei, A. Harifi and S. Golzari, “Improving the precision of KNN classifier using nonlinear weighting method based on the spline interpolation,” 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), 2017, pp. 289-292.
[4] P. Nair and I. Kashyap, "Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 460-464.
[5] Kesarwani, A., Chauhan, S.S., Nair, A.R. and Verma, G., 2021. Supervised Machine Learning Algorithms for Fake News Detection. In Advances in Communication and Computational Technology (pp. 767-778). Springer, Singapore.

[6] A. Kesarwani, S. S. Chauhan and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier," 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, pp. 1-4,10.1109/ICACCE49060.2020.9154997.

[7] Nagashri, K. and Sangeetha, J., 2021. Fake News Detection Using Passive-Aggressive Classifier and Other Machine Learning Algorithms. In Advances in Computing and Network Communications (pp. 221-233). Springer, Singapore.

[8] Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A., Cai, J., Li, L., Vuong, K. and Wadhwa, E., 2020. Fake news detection with different models. arXiv preprint arXiv:2003.04978.

[9] S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOREd), 2017, pp. 110-115, doi: 10.1109/SCORED.2017.8305411.

[10] I. Kareem and S. M. Awan, "Pakistani Media Fake News Classification using Machine Learning Classifiers," 2019 International Conference on Innovative Computing (ICIC), 2019, pp. 1-6, doi: 10.1109/ICIC48496.2019.8966734.