PalArch's Journal of Archaeology of Egypt / Egyptology

FAKE NEWS AND HATE SPEECH DETECTION WITH MACHINE LEARNING AND NLP

Samujjwal Goswami, Manoj Hudnurkar, Suhas Ambekar

Symbiosis Centre for Management and Human Resource Development ,

Symbiosis International (Deemed University), Pune, India

Email: suhas_ambekar@scmhrd.edu

Samujjwal Goswami, Manoj Hudnurkar, Suhas Ambekar: Fake News and Hate Speech Detection with Machine Learning and NLP -- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(6). ISSN 1567-214x

Keywords: Fake news, Hate speech, TD-IDF, machine learning, NLP.

ABSTRACT

With the increase in the ease of publishing and distributing news over the years, the fake news and hate speech propaganda has taken up a huge chunk in our daily routine, whether we can identify them or not. We must act to tackle this scenario as these can have contributed to the increase in political or communal hatred, which can cause severe damage to society. In the proposed research, we are going to extract features of the language and content by collecting examples of both real and fake news. We are going to train a model to classify fake news articles based on the NLP technique called TD-IDF (term frequency-inverse document frequency) vectorization which gives us the importance of each keyword within the news article or the speech. Then with the application of logistic regression we can classify the article/speech that can help us to identify such articles and deal with them.

1. Introduction

With the advent of Social Media and ever increasing publishing work it has become important to identify fake news and hate speeches embedded in various articles or posts. These fake news and hate speeches can severely impact our society spreading hatred into various community, religion or different groups of people. The flexibility and anonymity offered by the internet has made things very important to us to detect and discard such articles and posts. (Ankesh Anand, 2019)The government of different countries are also trying to

deal with this situation but with the increasing volumes of such articles and posts, it has become difficult to tackle this scenario manually. That is why we need applications of Machine Learning and Natural Language Processing (NLP). With the application of Machine Learning and NLP, we can classify articles and posts into different categories and apply the same logic with to them. Even with the application of these, it is not easy to identify hate speeches and fake news as these articles may have various forms of information which can spread hatred and miscommunication to the people and create a chaotic state in the society. One article can have hate speeches, profane and abusive languages and even cyberbullying, which can be all classified as toxic languages and they can have negative impact on our society as they target ones religion, ethnicity, race, gender, disability or sexual orientation. (Aditya Gaydhani, 2018) To handle the scenario better, we must first understand what are hate speeches and fake news.

2. Literature review

The broad questions that we have on the topic are:

- What are fake news and hate speeches and how they are impacting our society?
- What are the effective methods that can be applied to detect fake news and hate speeches?
- What is the method that can be most relied upon?

The basic objective of our research is to find out the answers of the questions mentioned above based on the past research that have been carried out already on the topic. Since 2015, researchers have invariably tried to carry out their study in this aspect and we have tried to do a cumulative study of their approaches

2.1. Hate Speech

Hate speech can be derived as any sort of communication which derogates any person or a community based on nationality, race, caste, gender, ethnicity etc. Some examples are (Joni Salminen, 2020):

- Tell those idiots to go hell!
- That women should mind her own business.
- Wipe out the Muslims.

These hate messages can be classified as abusive, hostile and threatening , but for research purpose, we will take all of these classes under one roof and treat them all as hate speeches. The goals of these hate speeches are mainly twofold: to intimidate a particular targeted person or community and secondly to let the bigots know that they are not alone. So building a counter measure to tackle these, first of all we need to identify if a post or an article belongs to the category of hate speech or not.

2.2. Fake News

In simple terms, Fake news are the news which are not based on truths and facts. News plays an important role in our day to day life and that is why fake

news can create a significant problem to our society. Fake news is an increasingly common feature of today's social and political landscape as we have seen it can lead to some hideous crimes like murder. It is not uncommon in India that someone has done some sort of damage to someone's life or property based on totally baseless and fake news. With the increasing number of WhatsApp and Facebook users, the spreading of such news has become really easy and it can really do some serious harm to our daily life. Source verification, fact checking and accountability are some of the basic journalistic approaches which can be easily bypassed or simply ignored by individuals or organizations publishing content on different social networks. This is one of the reason because of which terminologies like troll, alternative media facts, fake news, post truth media etc. comes out. (Marina Danchovsky Ibrishimova, 2020)

The traditional way to deal with such fake news is to do proper fact checking. But with the increasing number of such news has created problem in manual checking and that is where Machine Learning and NLP has come to our aid (Balmas, 2014).

In this paper, we are going to implement hate speech and fake news detection by devising a machine learning model. We first collect our required datasets, we train our model in TD-IDF (term frequency-inverse document frequency) basically for feature extraction and metric scores. Then we will be using classification algorithms like Support Vector Machine (SVM), Naive Bayes, logistic regression, XGBoost and compare the algorithms to find out which one works the best with TDIDF features. (J, 2019)

2.3. Related Work

Various independent methods and approaches have been implemented to detect and handle fake news and hate speeches over the years. Most of the approaches deals with extracting the lexical features of the data, pack them into a bag of words and do the differentiation (Pinkesh Badjatiya S. G., 2017) (Aditya Gaydhani, 2018) (Ziqi Zhang, 2018). On the other hand some approaches has been taken to do the sentiment analysis of the data and then classify the data as hate speech or not (Fazal Masud Kundi, 2014). But sentiment analysis necessarily do not give us the whole crux of the data. This is because in a particular speech or article, only a small percentage of words can represent the hatred hidden inside the speech and the rest of the article may show positive influence. To tackle Fake news, in (Ankesh Anand, 2019), the concept of Clickbait detection is mentioned. Clickbaits are nothing but catchy headlines that attacks a person's curiosity and makes him/her to click that particular link which may spread fake news among the readers. To detect Clickbait, the author has proposed both NLP and Deep Learning and evaluated that Deep learning is more efficient in this case. As a part of deep learning algorithm, the author has used Recurrent Neural Network (RNN), long short term memory (LSTM) and Gated Recurrent Unit.

In the research paper (Ziqi Zhang, 2018), the author has performed DNN (Deep Neural Network) on data extracted by twitter API to classify the text as hate

and no hate. The author has also used Convolutional Neural Network (CNN) to create a machine learning model and have calculated precision, recall and F1 metrics for accuracy prediction.

For fake news detection, in (Ray Oshikawa, 2020), the author has used classification algorithms like regression and sentiment analysis. The proposal of a hybrid approach is mentioned in (Zhixuan Zhou, 2018) where the author proposes to have a model combined with the text and user responses that the article receives and then calculate the sentiment score for it. In (Aditya Gaydhani, 2018), the author has used an N-gram method instead of mono or bi gram method and has shown that this method has higher proficiency in text classification.

There are basically two types of classification methods available to perform in case of creating a model. They are:

- Non- neural Methods
- Neural Methods

In non-neural methods, the methods like SVM, Naive Bayes, Logistic regression etc. falls into. On the other hand neural methods include RNN, LSTM, and CNN etc. For our research, we will focus mainly on the non-neural methods do our analysis.

There is a clear debate that we can observe among the researches that has been going on between the two mentioned methods. In (Ray Oshikawa, 2020), the author has said that neural methods are clearly better in terms of detecting any foul language or misinformation and says that for complicated data it should be used more often. In the same tone, (Pinkesh Badjatiya S. G., 2017) in their research said that deep learning methods are more superior to the other related methods. On the contrary, in papers like (Thomas Davidson, 2017) and (Joni Salminen, 2020), they have found that Non neural methods are the basics of doing classifications and only after performing them we can move ahead towards the modern methods. That is the primary reason, we are focusing more on the non-neural classification methods for our purpose.

In order to detect fake news, in (Marina Danchovsky Ibrishimova, 2020), the author has mentioned that in primary research, it is found that in a fake news, the use of proper nouns are more compared to other nouns. So to check the credibility of one post or article, we can evaluate the ration of proper noun to other nouns and evaluate the model according to that. But the problem with fake news detection algorithm will always be there as without proper fact checking, absolute surety is difficult to obtain.

There are various problems in online hate detection (Joni Salminen, 2020). The first one is the number of false positives and false negatives. False positive occurs when the model detects a non-threatening sentence as a hateful content and false negative occurs when the model cannot identify a hateful message and pass them as non- threatening. Subjectivity of the hateful comments in an article can also create problem. Apart from them, polysemy, meaning one word having different meanings within the article can also create confusion and the model may not perform as per the expected standards. The methods used in (Joni Salminen, 2020) basically involves keyword- based classification,

distributional semantics and deep learning classifiers. The author has also performed Feed- forward neural network (FFNN) for the classification machine learning model. The author compares the ROC AUC scores of all the classification model used and then finally comes up with the result given by the best model.

In (Peter Bourgonje, 2018), the author presents a system for stance detection of headlines to detect fake news with regard to their respective articles. The system is based on simple n-gram matching for the binary classification of "unrelated" vs. "related" headline/article pairs and then they have used Logistic regression to come up with the results by measuring the accuracy in terms of precision, recall and F measure.

The text classification methods can be further subdivided into two different categories (Shanita Biere, 2018):

Classical methods basically deals with manual feature extraction and combining that with different statistical algorithms. Feature engineering process mainly involves TF-IDF, bag of words, n-gram etc. whereas statistical methods like random forest, SVM, Naive Bayes, logistic regression etc.

Deep learning methods uses neural networks to automatically create features in a given dataset. Previously, the NLP techniques are mostly influenced by classical machine learning approaches, but as per the author, lately it is seen that nonlinear neural networks have been showing an impressive result in feature extraction and text classification. The most popular deep learning methods to do so are Recurrent Neural Networks (RNN), Long Short-Term Memory network (LSTM) and Convolutional Neural Network (CNN). (Pinkesh Badjatiya S. G., 2017)

In the research paper (Bashar Al Asaad, 2018), to tackle fake news the author has used a dataset of real and fake news to train and test a model in python using Scikit learn library. By extracting features from the data using bi- gram method and bag of words, the author has used two approaches namely linear classification and probabilistic classification. The detection of fake news is to be done mainly through Clickbait detection. In the paper, the author talks about four approaches of machine learning:

- Supervised Learning
- Unsupervised Learning
- Reinforced Learning
- Evolutionary learning

The machine learning method for supervised learning involves these basic steps: (Bashar Al Asaad, 2018)

- Data collection and preparation: This process involves looking for proper data set, preprocessing and cleaning the data for feature extraction. This process is very important to get the optimum result from the model.
- Feature selection: It involves identifying the features that will be most important to our analysis.
- Choice of algorithm: The choice of algorithm depends on the dataset that we have. After the feature selection, we need to choose a proper algorithm to extract the features from the data.

- Model and parameter selection: It involves choosing the proper machine learning model and parameters for best performance.
- Train and Test data for the model: We need to have separate data for both training the model and testing it on a different set of data. Only after training the model with the train data, we can proceed further to test the model with the test data and come to a particular result.

The approach taken by the author (Bashar Al Asaad, 2018) to detect fake news is that she has first taken the step of parsing the HTML source code of the web page. Through this parsing the author is able to extract the required information from the web page for fake news detection. Then the author has used Machine Learning to see if the title has some sort of clickbaits or if the main body of the text is fake or not. Then the author uses cosine similarity method to check the similarity between the lists of news titles and the news title they have. With such comparisons, the author has created model to check the authenticity of a news and check for presence of any clickbaits in the title.

Thus, we see that there has been a significant amount of work in the field of detecting hate speeches and fake news. Apart from the journal articles that we have taken into consideration for our literature review, there are plenty more research papers that are available in this particular area. By studying the papers we have come to know that although there are different independent research papers available for fake news and hate speech detection, there are very limited research that has been done taken these two together. That is why in this research paper, our aim is to do the research on hate speech and fake news detection by taking them under one umbrella. As from the analysis above, we get to know that it is almost impossible to detect if a news is fake or not without proper fact checking, so we are going to club this feature of our research with the detection of hate speeches. There is a more probabilistic chance that if a news or article is fake, it has to have some sort of negativity or hatred associated with it. With this, if we find that there is some sort of hatred associated with a particular article, we can find out if the hatred is genuine or it indulges in any sort of fake news and we will have to do it with the help of proper fact checking. So, this research paper will help us especially in detecting hate speeches and also we can check if any hatred is associated with an article or post and then then we can check their authenticity with proper fact checking.

One another research gap that we have found is that for non-neural methods, the researchers have applied different classification algorithms individually, but have not compared the results of different algorithms to one another. Even if the authors have done some sort of comparison, the numbers of such algorithms are very less. So, in our research we have planned to carry out classification using different non neural algorithms like SVM, Naive Bayes, Logistic Regression and XGBoost and compare the result of all of them together to detect fake news and hate speeches.

3. Research methodology

We have implemented a methodology of checking a piece of text or post and classify them as hate message or not. If we have to put it in brief, we have to first tokenize the words and from the review of the related work, we got to

know that TF-IDF is much more efficient in this regard rather than going with the bag of words. (Funk., 2018) After the feature extraction, we are going to use various NLP classification techniques like SVM, Naive Bayes and even logistic regression to check which algorithm gives us the most accurate results. The coding part will be done in Python using different built in Python libraries like Scikit Learn, Numpy, Pandas etc. To identify which classification score gives us the more accurate result, we will evaluate their F1 scores and evaluate according to that.

3.1. Natural Language Processing

Natural Language processing or NLP is also called computational linguistics. (Shanita Biere, 2018) It is basically used to have an automatic processing of human defined languages. NLP is comparatively new and is a multidisciplinary and large field and can be defined as interaction between human languages and computer. It is a part of machine learning which is also driven by an integral part of Artificial Intelligence. (de Gibert, 2018)

3.2. Data Collection and Preprocessing

The first step towards using NLP to detect hate speeches and fake news is to get the data. (de Gibert, 2018)We collected the data from github which is a collection of mostly one liner comments collected from different domains. The data is prepared in a way that it covers most of the aspects of hatred that are mostly used in literature or in colloquial form. The data collected is in the text form and we have collected it in a way that we can segregate the data into train and test data.

- Data Sampling: We have to create our train and test file from the file and we have to sample it accordingly. We split the data as 70-30% for train and test respectively for our classification purpose. The sampled data contained classified texts of hate and non-hate speeches.

So, after segregation, we have three files, one containing all the data, one sampled train data and one sampled test data. After the data collection is done, we need to clean the data so that it is applicable to the classification algorithms. In order to preprocess the data, different steps are taken:

- Removal of Space: Unnecessary spaces and signs are removed from the data. Only the data that has alphabetic value (a-z, A-Z) are taken into account.
- Tokenization: we need to do stemming and lemmatization to the whole set of data with the help of Python libraries. After the tokenization of text into features, it is important to do the conversion of texts into numbers. Machine cannot identify texts, so to for the machine to understand the text and the weight of each and every word, we will have to vectorize the text. This process we will do with the help of TF-IDF vectorization, as discussed earlier because it has proven to much more efficient than conventional methods like Bag of Words

3.3. TF-IDF

TF-IDF stands for Term frequency and Inverse Term Frequency. It is primarily used to vectorize a particular set of texts into vectors so that computer can understand them. TF-IDF vocabulary is used when we begin to train the model and thereafter reuse in the test data also. This process gives us the importance of the words in a sentence and weighs its importance within that particular sentence. TF-IDF is a proven method for text classification as instead of counting the words in a sentence, it can give us the importance or frequency of that word within the whole text.

TF-IDF calculates the frequency of a term in a text along with its count of frequency in the whole set to characterize the input text. With the help of these frequencies, one can calculate the value of that word in the document. In TD-IDF, we calculate the weight of a word by multiplying two metrics that are term frequency (how many times the word is used in a text) and Inverse document frequency (How many sentences are there in the whole text containing that particular word). So both Term Frequency and Inverse Document Frequency can be defined as:

TF= (Number of repetition of words in a sentence) / (Number of words on a sentence)

And

IDF= Log ((Number of sentences in the text) / (Number of sentences containing the word))

Logarithm is used in order to calculate the IDF as it is a monotonically increasing function. The weight score of a word in the text is determined by multiplying both the values of TF and IDF. So,

Weight of a word in a text = TF * IDF

After TD-IDF vectorization in Python, each and every word will have a designated TD-IDF vector value (which is a number) giving their importance in the whole text. When we evaluate the TD IDF values for each and every word, we will require to normalize the value which we will do with the help of Python programming.

3.4. Classification

After the vectorization of the texts are done, we are going to move to the most important part of our research, which is to apply classification algorithms and find out the accuracy of each of them. We will train the model with our train data and do the testing on out test data to evaluate their accuracy. The accuracy of these algorithms will give us the idea of how these algorithms fare with the vectorize value of TF-IDF and we can compare these algorithms for the most accurate result.

In our research, we have decided to use four prominent machine learning classification algorithms namely Support Vector Machines, Logistic Regression, Naive Bayes and XGBoost. But before applying any of these algorithms, it is important to understand how these classification algorithms work to give us the best result in hate speech and fake news detection.

3.5. Support Vector Machines

Support vector machines or SVM is a prominent machine learning algorithm for text classification and regression analysis. It does not create any assumptions towards the data. Rather than it creates a hyper plane that segregates the input data into two segments. These two sets create a plane and the point on these planes are termed as support vectors. The advantages of using SVM is that it is less complex compared to the deep neural network methods and the interpretation is also not that complex.

3.6. Naive Bayes

Naive Bayes is one another machine learning classification algorithm that can be extensively used for text classification. The algorithm provides a probabilistic approach which is based on Bayes' theorem. It uses the assumption of conditional independence and the total probability theorem. Naive Bayes basically calculates the probabilities of values in a particular dataset by counting the frequencies of them. One of the major advantages of Naive Bayes is that it is considered to be the fastest when compared with other machine learning classification algorithms. Also it is very useful while dealing with a large data sets.

3.7. Logistic Regression

Logistic Regression is a technique from the field of statistics which is widely used in machine learning for text classification. It is mostly used in binary classification problems and that is why it is very popular as a text classification algorithm. Logistic regression uses logistic function or sigmoid function and its expression is given as,

$$1 / (1 + e^{-(t)})$$

Logistic regression is typically used when the dependent variable is categorical, as it gives us a binary output through which we can do out classification process.

3.8. XGBoost

XGBoost is the final machine learning algorithm that we are going to use for our classification purpose. XGBoost stands for Extreme Gradient Boosted Decision Trees and is an ensemble algorithm which extensively uses decision trees for the classifications. These decision trees are first combined with gradient boosting in order to create successive models to learn from the previous classification model's error. This means that the model first observes the result of the first decision tree that it uses and based on the result the algorithm moves to the subsequent decision trees. (Joni Salminen, 2020)

3.9. Other classification methods

Apart from the four classification techniques mentioned above that we are going to use for our analysis, there are some other traditional classification algorithms are these such as Random Forrest and Decision Tree. But we are going to limit our research to the algorithms mentioned above as we see from

out literature review that these four are the most efficient in text classification compared to others.

3.10. Evaluation of algorithms:

After the classification algorithms are applied, we need to calculate a parameter through which we can evaluate which classification algorithm performs the best along with TD-IDF vectorization. In order to do so, we will use the F1 score for each of the algorithms that we have applied. But, to calculate the F1 score, we first need to understand two basic terms

- Precision or positive prediction value
- Recall or true positive rate

Both Precision and Recall can be derived from the confusion matrix of the algorithms. Their formula is given as:

Precision (p) = True Positive / (True Positive + False Positive)

Recall (r) = True Positive / (True Positive + False Negative)

We can now evaluate the F1 scores for all the algorithms as F1 score is nothing but the harmonic mean of the two. So the formula to calculate F1 score can be given as:

$$F1 = 2 \times (p \times r)/ (p + r)$$

So, by evaluating F1 score for each of the algorithms, we can find out which algorithm gives us the highest value and accordingly we can conclude which algorithm performs the best along with TD-IDF vectorization to classify hate speech and fake news in the given dataset.

4. Results and findings

As mentioned in the research methodology, first of all we calculate the TD-IDF scores with 10 cross validation and the values obtained for all the classification algorithms are mentioned in the below table:

Classification Algorithm	TD-IDF scores
Logistic Regression	0.723
Support Vector Machine	0.69
Naive Bayes	0.65
XGBoost	0.762

Table 1: TD-IDF Scores

After the TD-IDF evaluation with 10 cross validation, we move to apply the algorithms using different built in libraries in Python. After their successful execution, we need to find out the F1 scores for all the algorithms. Using the formula mentioned in the methodology, we first evaluate the confusion matrix and then calculate the precision and recall values for all the algorithms. The table below represents the precision and recall values for the classification algorithms:

Classification Algorithm	Precision	Recall
Logistic Regression	0.77	0.75
Support Vector Machine	0.73	0.75
Naive Bayes	0.69	0.72
XGBoost	0.82	0.83

Table 2: Precision and Recall Values

After this, we can easily calculate the F1 values with the harmonic mean of precision and recall for each algorithm. The Table below gives us the required F1 scores:

Classification Algorithm	F1 Scores
Logistic Regression	0.76
Support Vector Machine	0.74
Naive Bayes	0.7
XGBoost	0.82

Table 3: F1 Scores

5. Discussions

The result above shows that XGBoost outperforms the other classification algorithms and so it can be derived that it is the highest performing algorithm. After XGBoost, Logistic Regression performs better than the remaining two while Naive Bayes is the least performing among them all. But it is to be taken into consideration that the result shown here is true only in case of the given dataset. If we perform the same procedure with the same algorithms but on a different dataset, there is a probability that the results might differ from the current ones. Also there are various parameters that can be used to check the performance of these algorithms. For example, along with F1 scores, we can also use the area under the ROC-AUC curve to evaluate their performance. But for our paper, we have concentrated on measuring performances of the classification algorithms based on F1 scores only. Also it is to be taken into account that, in order to detect hate speech and fake news, which is the primary motive of our research, we need to use only one of the classification algorithms mentioned above. We have calculated the performance results of all the algorithms only to find out which one works best with our present set of data.

Although we have created different models for different classification models, none of them can give us the 100% correct result. XGBoost gives us the most accurate results, but it has a recall of 0.83, which means 17% of the data that is actually hateful, the model has misclassified them. Also, the precision for XGBoost is 0.82. So, the model has classified 18% data as hateful or offensive which should have been classified as clean.

In one of the past research, (Aditya Gaydhani, 2018), the authors have carried out a similar approach and compared the results of Naive Bayes and Logistic Regression together. Just like our research, it was found that for that given data also, Logistic Regression outperformed Naive Bayes in terms of accuracy. The author then evaluates the F1 score for only logistic regression to calculate how the model is performing. In some other work, (Joni Salminen, 2020), (Y, 2012)

authors have done classification for hate speech detection using BERT and XGBoost and XGBoost seems to show better results between them. So, we can see that our result falls in the line of some of the other studies that have been conducted in the past. One unique advantage that our analysis have is that it uses as many as four classification algorithms have been used here for comparison of better results whereas it was seen that previous studies have limited their research to two or three at most.

6. Limitations

We have tried our best to create an efficient method and a model to detect hate speech and fake news in our research, but certainly there are some limitations in our approach. We have listed our limitations below:

- The data we have taken is in English language and so the model will only not work for any other languages apart from English. There can be hate speeches and fake news in other languages also which needs to be detected, but our model will not be able to do so.
- We have taken a comparatively smaller dataset for hate speech and fake news detection and getting the results according to that. The truth is now a days, the volume of hate speeches and fake news are growing exponentially and we need to take all of them into account.
- In the present era, new hate speeches are evolving every day. Some terms that can spread hatred among people might not have existed few years back. So for such new words that are coming day by day, we need to update the train data frequently. So, our model might not work as expected when we have to deal with such words.
- As we have said, to detect fake news, we need proper fact checking. In our model, we first need to check if there is any hatred associated with the speech and then we are going for the fact checking for detecting fake news. So, our model is more efficient in detecting hate speeches rather than fake news.
- We have limited our research only to the traditional classification algorithms like SVM, Naive Bayes, XGBoost and Logistic regression. There are various deep learning and neural network methods that can also be applied for classification purpose and they might give us a better result.

7. Conclusions and recommendations

There are different sorts of toxicity that hate speeches and fake news can spread in our society and measures should be taken in order to tackle them. The aim of our research paper is to detect fake news and hate speeches with the help of Natural Language Processing (NLP). For successful detection, we first needed to understand what fake news and hate speeches are and that is why we gave an overview of the topics. We also got to know what NLP is and how it can be used to detect hate speeches and fake news with the help of different classification algorithms. We have reviewed further literature to understand what the methods are that has been or can be taken to successfully carry out our objective. Through the literature review we got the understanding of the application of various neural and non-neural methods to detect hate speech and fake news and how we can conduct the study for our purpose.

For our research purpose, we have applied 10 cross fold TD-IDF vectorization for feature extraction from out dataset. We have broken down out data into two datasets: train dataset to apply classification algorithms namely SVM, Naive Bayes, XGBoost and Logistic Regression and test on the test dataset. After completing our analysis, we found that, for our dataset, Along with TD-IDF vectorization, XGBoost performs the best and Naive Bayes performs the worst in classification. We have used F1 scores for all the algorithms to measure their performance in our research.

The study is mostly going to be useful to curb the spread of hatred and fake news among the people in society so that we can get rid of various unwanted topics and instances. Also one can use the model to create a business plan which works in identifying and dealing with such offensive articles and fake news. As a future study, we can find the different combinations of methods of feature extractions and classification algorithms. There are many such techniques which can give us a better accuracy to detect fake news and hate speeches. Also, to have a more sophisticated method, we should use features of data mining which will work on big data. This is because, so far we have applied our research methods in a smaller dataset and to achieve higher accuracy score we can use bigger dataset that has more types hate speeches and fake news of different linguistic features.

References

Aditya Gaydhani, V. D. (2018). Detecting Hate Speech and Offensive Language on Twitter with ML. arXiv:1809.08651v1 [cs.CL].

Ankesh Anand, T. C. (2019). We used Neural Networks to Detect Clickbaits. arXiv:1612.01340v2, 7.

Balmas, M. (2014). When fake news becomes real:. Communication Research, 41(3), 430.

Bashar Al Asaad, M. E. (2018). A Tool for Fake News Detection. ResearchGate Conference Paper • September 2018.

Bourgonje, P. S. (2017). From clickbait to fake news detection: . proceedings of the 2017 EMNLP., 36.

de Gibert, O. a. (2018). Hate Speech Dataset from a White Supremacy Forum. Association for Computational Linguistics, 11-20.

Fazal Masud Kundi, S. A. (2014). Detection and Scoring of Internet Slangs for Sentiment Analysis. Life Science Journal.

Funk., M. H. (2018). Machine Learning for the Quantified Self. Springer International Publishing AG.

J, Z. (2019). Identifying Offensive Tweets Using BERT and SVMs. 1904.03450 [cs].

Aditya Gaydhani, V. D. (2018). Detecting Hate Speech and Offensive Language on Twitter with ML. arXiv:1809.08651v1 [cs.CL].

Ankesh Anand, T. C. (2019). We used Neural Networks to Detect Clickbaits. arXiv:1612.01340v2, 7.

Balmas, M. (2014). When fake news becomes real:. Communication Research, 41(3), 430.

Bashar Al Asaad, M. E. (2018). A Tool for Fake News Detection. ResearchGate Conference Paper • September 2018.

Bourgonje, P. S. (2017). From clickbait to fake news detection: . proceedings of the 2017 EMNLP., 36.

de Gibert, O. a. (2018). Hate Speech Dataset from a White Supremacy Forum. Association for Computational Linguistics, 11-20.

Fazal Masud Kundi, S. A. (2014). Detection and Scoring of Internet Slangs for Sentiment Analysis. Life Science Journal.

Funk., M. H. (2018). Machine Learning for the Quantified Self. Springer International Publishing AG.

J, Z. (2019). Identifying Offensive Tweets Using BERT and SVMs. 1904.03450 [cs].

Joni Salminen, M. H. g. (2020). Developing an online hate classifier for multiple social media platforms. Huma centric computing and information sciences.

Marina Danchovsky Ibrishimova, K. F. (2020). A Machine Learning Approach to Fake News Detection Using NLP. researchgate.

Peter Bourgonje, J. M. (2018). From Clickbait to Fake News Detection:. DFKI GmbH, Language Technology Lab.

Pinkesh Badjatiya, S. G. (2017). Deep learning for hate speech detection in. 26th International Conference on, 1(International World Wide Web Conferences Steering Committee.), 760.

Pinkesh Badjatiya, S. G. (2017). Deep learning for hate speech detection in tweets.

Ray Oshikawa, J. Q. (2020). A Survey on Natural Language Processing for Fake News Detection. arXiv:1811.00770v2 [cs.CL], 8.

Shanita Biere, P. d. (2018). Hate Speech Detection Using Natural Language Processing Techniques. Vrije University.

Thomas Davidson, D. W. (2017). Automated hate speech detection and the problem of offensive.

Y, C. (2012). Detecting offensive language in social media to protect adolescent online safety. international conference on social computing, 80.

Zhixuan Zhou, ,. H. (2018). Fake News Detection via NLP is Vulnerable to Adversarial Attacks. 21.

Ziqi Zhang, L. L. (2018). Hate Speech Detection: A Solved Problem? IOS Press arXiv:1803.03662v2 [cs.CL], 21.