IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

# Analyzing And Detecting The Fake News Using Machine Learning

**Anant Kumar[1], Satwinder Singh[2*], Gurpreet Kaur[3]**

[1,2]Department of Computer Science and Technology, Central University of Punjab, Bathinda, Punjab, India
[3]Department of Law, Bathinda College of Law, Bathinda, Punjab, India

[*]*Corresponding Author: satwinder.singh@cup.edu.in, Tel.: +91-93000-64064*

DOI: https://doi.org/10.26438/ijcse/v7i5.10441050 | Available online at: www.ijcseonline.org

Accepted: 15/May/2019, Published: 31/May/2019

***Abstract*—** In recent years, as our social media has become more and more prevalent. The news websites and blogs have become into the limelight, there are number of web pages and social media that have come into the state that claim to report on upcoming events, but whose reliability has been brought up into question. Now the debate over such websites and news agencies has become so prevalent that the issue of 'fake news' is itself an vital part of the news world. What establishes 'fake news,' in any case, has just turned out to be less clear as the topic has turned out to be increasingly normal, with standard news sources. Nowadays' fake news is making various issues from mocking articles to a created news and plan government publicity in certain outlets. fake news and absence of trust in the media are developing issues with immense consequences in our general public. It is needed to look into how the techniques in the fields of computer science using machine learning, natural language processing helps us to detect fake news. Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. In this research a comprehensive way of detecting fake news using machine learning model has been presented that is trained by two different data which is based on US election fake news and recent Indian political fake news respectively.

***Keywords*—**Machine Learning, Fake News, Data Cleaning, Classification Model, Text processing, Natural Language Toolkit.

## I. INTRODUCTION

World of advanced digital media is expanding continuously thus does the inclination of organizations to grow it all the more picking up them greatest financial advantages. This particular urge calls for an ever increasing number of progressions concerning making and growing crisp substance whether as websites that goes for marking organizations or as online papers and magazines. Since from last few decades' medium of communication had changed. Now a day people are using social networks very extensively for news updates. In terms of data our research is completely based on the data's and information that possibly can be gathered from internet, open source database repositories [1].

The word Fake news came into public domain during 2016 US(United State) election and many research has been conducted regarding it [2]. So data scientist have been working on data collection which consist of the every information regarding news like title, text, news source, time when it was published. So many aspects have been studied and are still left to study to what level fake news can spread propaganda and can election result could be influenced in any part of the country. Many rumors, short stories in media and study to some extent have shown that the fake news has the ability to penetrate deep into the mas and social media which can reach both urban and rural social media consumer, since we don't have much amount of Indian political fake news based data that can be used for machine learning. Different corporate and political parties have started rooting up their IT cell which tries to touch as many people possible using fake pages and fake accounts on Facebook, twitter, Instagram WhatsApp and other apps. Now even the commercial media have started coming up with biased news and started a trend of adopting the news that benefits them the most or the people running the media house for some political or financial gain.

We manually added up some fake and real news after going through News articles of some fake busting news websites and to use those data into Machine Learning Model created for classification. Hence, the research covers both the political and social media of US and India respectively. The method has been implemented successfully to classify Fake and Real news.

Rest of this paper is organized as follows: social media statistics is discussed in section 2. Dataset description is mentioned in section 3. Section 4 contains the machine learning model used for the classification. Section 5 include the graphical representation of the actual result along with the result parameters. Finally conclude the paper at the end.

 

## II. SOCIAL MEDIA STATISTICS

If just take example of Facebook, the amount of fake news have increased exponentially. Corporates, political parties and big agencies have understood the power of Facebook and its reach. Agencies have been smartly building up Facebook pages and accounts which tries to spread important news at first and as the fan base increases the amount of propaganda news gets started shared within the followers. Since fake pages gets huge amount of views, the mainstream media start seeing low counts of page view on their news post, comments and page engagements. As the fake news are spiced up, fantasied, added up with cooked stories and made interesting using wrong facts, figures and photoshopped images people start sharing them on all other social media chat applications [3]. If we go by the Fig 1 we can clearly see a drastic downfall of mainstream news shared by verified accounts with respect to the fake news as the election day approaches.

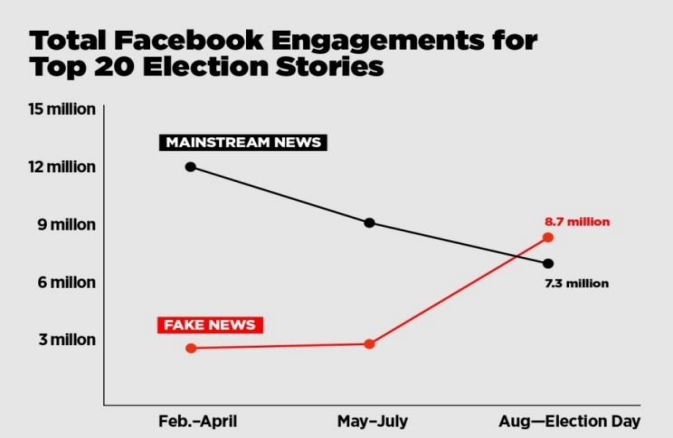


**Fig 1. Describes the Engagements for top 20 Election stories (Baron, statista, 2019)**

## III. DATASET DISCRIPTION

Relevant Collecting the distinct data with labels on fake news is quite a complex step in Machine Learning. since there were no definite measure or parameter that test on what ground a particular news is fake or real. Two dataset was collected. The first data has been collected from kaggle repository which focuses on US-based presidential election. The second dataset has been collected manually from news and articles sources which focuses on Indian politics related fake and real news. For convenience we shall use Dataset-1 for Kaggle website dataset and Dataset-2 for the manually collected data respectfully. Dataset-1 in this research include (6335 x 3), which means 6335 rows and 3 columns while Dataset-2 is of size (335 x 3).

Table 1. Dataset Description

| Dataset | Source | Target | Size of Data | Column Name |
|---|---|---|---|---|
| **Dataset-1** | Kaggle | US politics based | 6335 X 3 | Text, Title, Label |
| **Dataset-2** | Manually from Different News Source | Indian Politics Based | 335 X 3 | Text, Title, Label |

Each Fake news have labels in the 3rd column. So, in this research we used data set of news that were collected from US election 2016. Other information includes no. of view count, no. of like count, no. of comments made count. Dataset-2 includes (335 X 3), which means 335 rows and 3 columns, same format that was used in Dataset-1. So the data is collected in xlsx format as per the convenience or looking at the rate of data importing errors.

| title | text | label |
|---|---|---|
| You Can Smell Hillary's Fear | Daniel Greenfield, a Shillman Journalism Fello... | FAKE |
| Watch The Exact Moment Paul Ryan Committed Pol... | Google Pinterest Digg Linkedin Reddit Stumbleu... | FAKE |
| Kerry to go to Paris in gesture of sympathy | U.S. Secretary of State John F. Kerry said Mon... | REAL |
| Bernie supporters on Twitter erupt in anger ag... | — Kaydee King (@KaydeeKing) November 9, 2016 T... | FAKE |
| The Battle of New York: Why This Primary Matters | It's primary day in New York and front-runners... | REAL |

**Fig 2. First five columns of Dataset-1**

The figure 1 shows the first five rows of dataset-1 whereas first five rows of dataset-2 has been shown in fig 2. The Columns states the characteristics of the data with target label as Fake and Real stating to what category the data belongs.

| text | title | label |
|---|---|---|
| report pakistan tri take action daqood dawood ... | Dawood Ibrahim s assets worth 15 000 crores se... | Fake |
| sexual harass charg level sever women journali... | MeToo movement Union Minister MJ Akbar resigns... | Real |
| a video gone viral social media abhisar sharma... | Journalist Abhisar Sharma Bribing A Villager T... | Fake |
| union minist bharatiya janata parti leader smr... | If PM Modiji Loses I Will Commit Suicide Did S... | Fake |
| pakistan spread terrorist activ name islam | pakistan spreads terrorist activities in the n... | Real |

**Fig 3. First five columns of Dataset-2**

The figure 4 shows the composition of the fake news of kaggle dataset.we have 50.43% of Real news and 49.57% of Fake news in this dataset. In this we can notice that both the labels are quite balanced. The dataset needs to be balanced inorder to have better learning. Unbalanced data refers [1] to a situation where the number of comments for all classes in a classification dataset is not same. So we try to have equal amounts of label in the dataset sot that the model doesn't show a biased result. In some cases if one of the label increases then the model prediction sometimes shows a biased result.

 

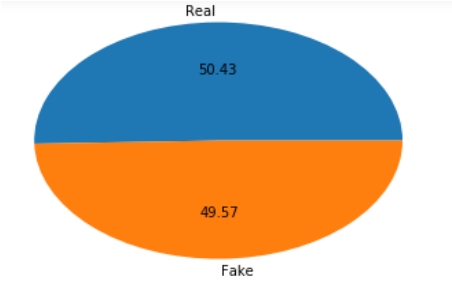


**Fig 4. Pie chart of Dataset**

The Fig 4 shows amount of Fake and Real news in Bar chart format. In this we can notice that both the labels are balanced. The dataset needs to be balanced inorder to have better learning. Unbalanced data refers to a situation where the number of comments for all classes in a classification dataset is not same. So we try to have equal amounts of label in the dataset set that the model doesn't show a biased result. In some cases if one of the label increases then the model prediction sometimes shows a biased result.

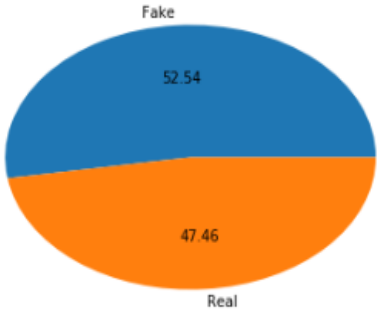


**Fig 5. Bar chart of First Dataset**

The Pie chart figure 5 shows the composition of the fake news of Indian Fake news dataset. we have 52.54% of Fake news and 47.46% of Real news in this dataset.

## IV. MACHINE LEARNING MODEL

Four clssification algorithm were used for the machine learning which are Logistic Regression, Naive Bayes Classifier, Nearest Neighbour and Neural Network. Four clssification algorithm were used for the machine learning which are Logistic Regression, Naive Bayes Classifier, Nearest Neighbour and Neural Network. Logistic Regression is a statistical technique for analyzing a data set in which there are one or more independent variables that determine an outcome [4]. The outcome is calculated with a dichotomous variable (in which there are only two possible outcomes). Nearest Neighbours algorithm simply stores instances of the data that is being trained [5]. Classification is computed from a simple majority poll of the nearest neighbours at each point [6]. This classification algorithm is made simple to implement, quite robust to train data that are noisy, and constructive if size of the data to train is huge [7]. Neural Network consists of units (neurons), which are organized in layers, which translate an input vector into some output [8]. Each node or unit takes an input, applies a function to it and then passes the output on to the next layer. Usually the networks are defined to be feed-forward in which a unit feeds its output to all the nodes or units present on the next layer, but there is no feedback to the previous layer. Thereafter weightings are applied to the signals passing from one node to another, and it is these weightings that are tuned in the training phase to adjust a neural network to the particular problem at hand [9].

## V. RESULTS AND DISCUSSION

Based-on different algorithms the accuracy and confusion-matrix has been discussed. Here we discuss the results of Linear SVC performed on Dataset 1. In the Fig.10 it can be seen that the accuracy score is 92.7%. The Precision, Recall, F1-score is as shown below in the figure.

```
----------------------LinearSVC------------------
              precision    recall  f1-score   support

        FAKE       0.91      0.94      0.93       608
        REAL       0.94      0.92      0.93       659

accuracy score is 0.9273875295974744
```

**Fig 10. Result of LinearSVC on Dataset-1**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. So here in Fig 11 the confusion matrix is shown that is obtained after applying LinearSVC on Dataset-1. The value 2037 denotes True Negative while 2014 denotes True positive. We want our model to have maximum value for these result parameters. We then have 177 as false negative and 154 as false positive. We want these values to be as low as possible as denotes the types of error made by model in classification. The Table 2 describes the labelling and explanation of confusion matrix.

```
[2037  154]
[ 177 2041]
```

**Fig 11. Confusion matrix of LinearSVC on Dataset-1**

 

The figure 11 demonstrates the confusion matrix of the classification algorithm LinearSVC on the Dataset-1. The Figure contains True Positive value of 2037, 2041 as True Negative, 154 as False Negative and False Positive as 177 respectively.

Table 2 Confusion matrix

| | | Actual class | |
|---|---|---|---|
| | | **True** | **False** |
| **Predicted Class** | **True** | True Positive(TP) | False Negative(FN) |
| | **False** | False Positive(FP) | True Negative(TN) |

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Precision can be defined as when a positive value is predicted, how often is the prediction correct can be termed as precision. The formula for precision has been shown in equation (1).

While Recall can be defined as When the actual value is positive, how often is the prediction correct can be termed as precision [10].

The formula for Precision has been shown in Equation (2).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

F1-Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1-Score is usually more useful than accuracy, especially if you have an uneven class distribution [11]. The Formula for F1 has been shown in equation (3).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

The figure 8 describes the most common words in the dataset-1. The graph describes what sort of words are mostly used in the news article. The words can be name of any politician, country, feeling, an adjective, related to social political, issue, economical issue or sensitive topic that are mostly been discussed in the news circle. The most common words give an idea what sort of words can attract or change the sentiments of a reader. The figure 9 describes the most common words in the dataset-2.

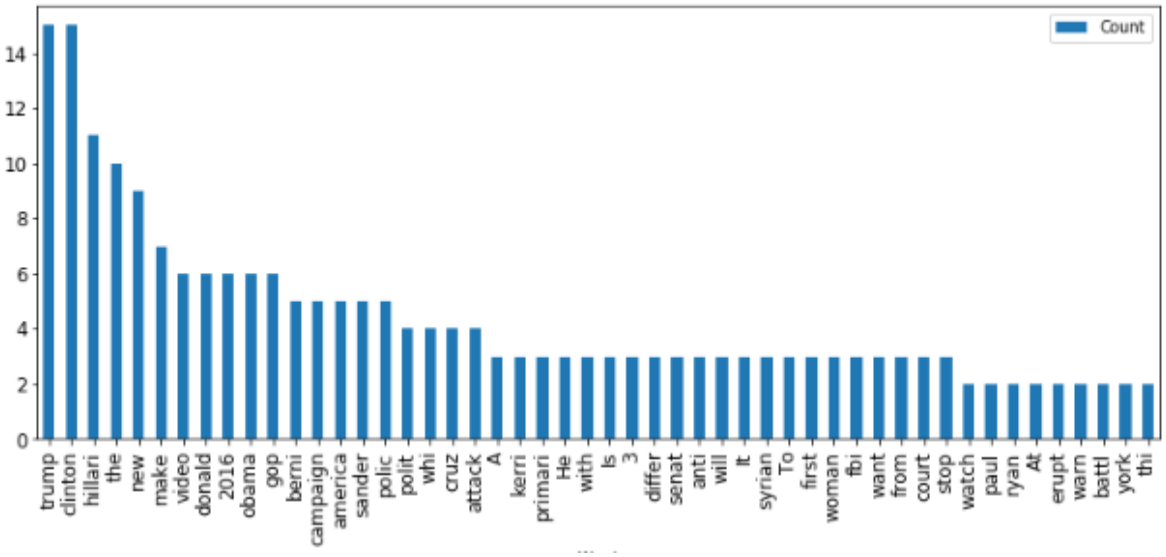


**Fig 8 Bar chart for most occurring words in Dataset-1**

 

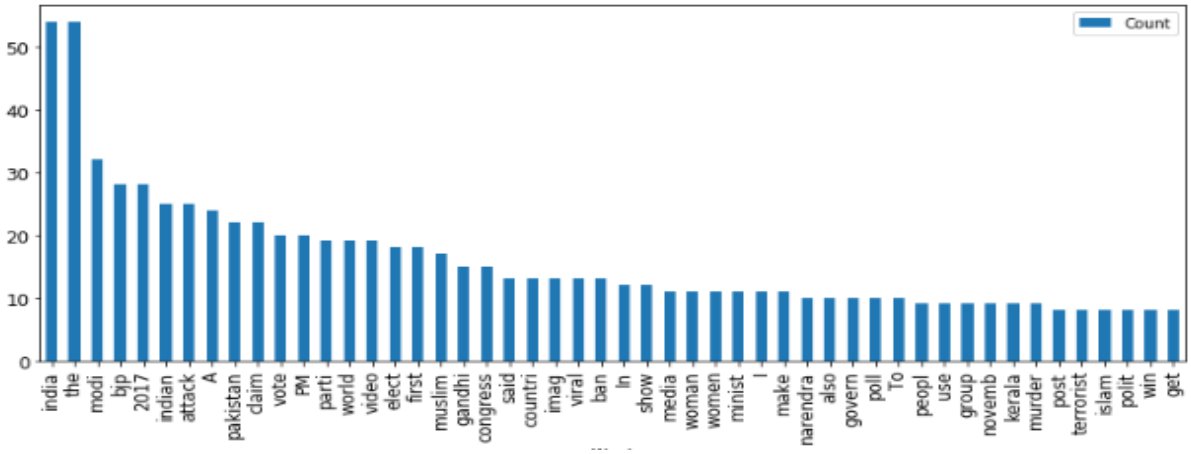


**Fig 9 Bar chart for most occurring words in Dataset-2**

The fig 8 shows the most common words which is calculated by Frequency Distribution function present in the library of natural language toolkit for natural language. Firstly, all the word is tokenized and passed through Frequency Distribution which calculates the words occurring throughout the text with the frequency of their appearance in the text. Then most common words are calculated with the number of appearance shown in the Figure 8 and Figure 9 for both the Datasets.

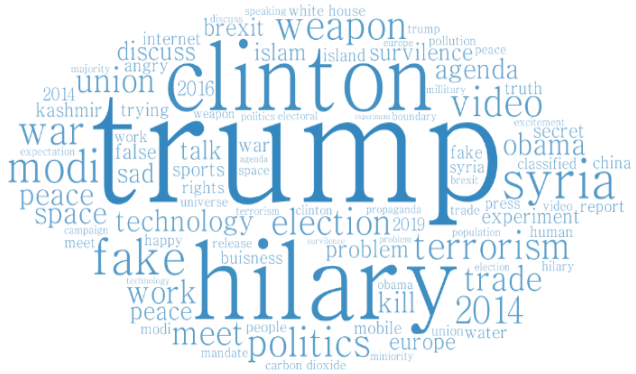


**Fig 10 Word cloud representation of most common 30 words in Dataset-2**


**Fig 11 Word cloud representation of most common 30 word before on Dataset-2**

From the most common words word cloud are drawn which is show in the figure 10 and figure 11 for both the datasets respectively. The word cloud is the collection of words in the form of cloud to represent the characteristics of the words as a visualization.

Now we discuss the Results that we got from Indian Fake News Dataset. The amount of entries or news in the Dataset-2 where comparatively less with respect to Dataset-1. Dataset-1 had 6335 news entry while Dataset-2 has 335 News entry. Thus the amount of value from which the model can learn and generalize were concentrated and hence the accuracy could reach up to 100 percentiles.

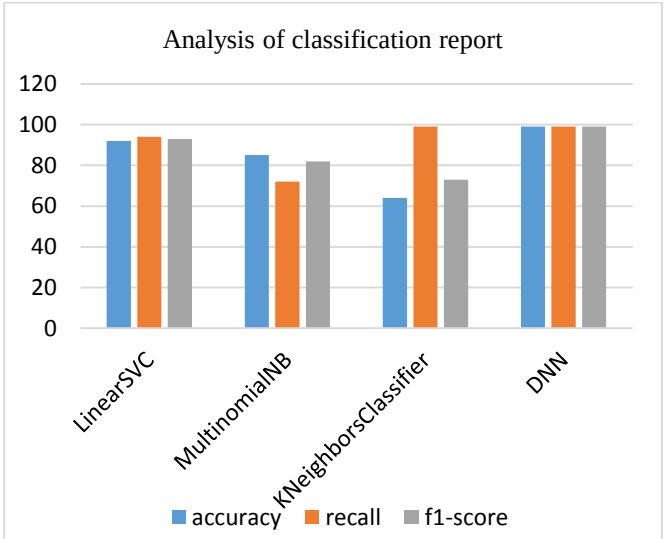


**Fig 12 Analysis of Classification report for Dataset-1**

The Result that we get from applying other classification on Dataset1 is mentioned in then Fig 12. As we can see in the

 

Figure 13 the accuracy is 85%. The reason for the score of not closer to 100 can be many. For E.g. Lack of surplus amount of fake news data, Language that the news is followed up like Mix of English, Hindi and other Regional Languages. Accuracy of 85% is quite good when it comes to applying machine learning algorithm on Indian Politics news with less number of datasets consist mix of both English and Hindi Words.

```
In [197]: print(confusion_matrix(y_test,predictions))
          print(" \n accuracy :",
                metrics.accuracy_score(y_test,predictions))
          print("\n",classification_report(y_test,predictions))

          [[47  8]
           [ 7 39]]

           accuracy : 0.8514851485148515

                        precision    recall  f1-score   support

                  Fake       0.87      0.85      0.86        55
                  Real       0.83      0.85      0.84        46
```

**Fig 13 Accuracy and Confusion matrix for Indian Fake News Dataset**

As the availability of applying text processing on different regional Indian languages are available for data science and as the dataset size increases, the accuracy will certainly increase. The Precision, Recall and F1-score for the classifier in the above figure 13 is shown as well which describes the confusion matrix obtained. The result for three parameters have been described for both the fake and real news labels.

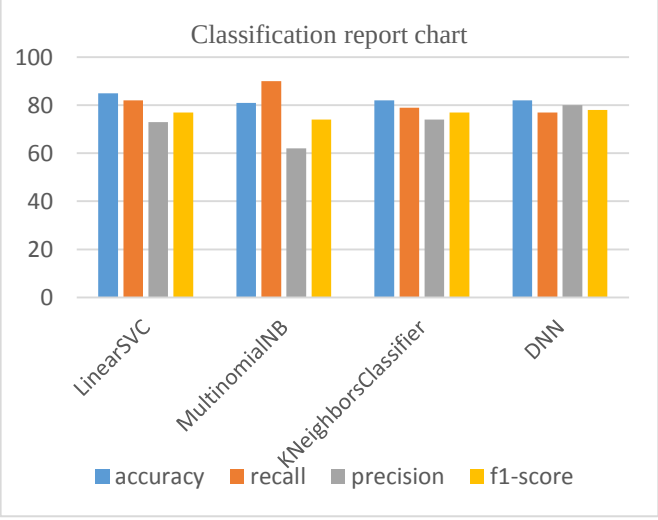


Fig 14 Analysis of Classification report for Dataset-2

The other classification algorithm results are shown in the Fig 14. So we use our trained model for prediction. We take some sample of news statement and then use prediction as per the learning. So the results predicted by trained model has been encircled with red in the Fig 15 below.

```
text_clf.predict([" china wants to distroy world trade "])
array(['Fake'], dtype='<U4')

text_clf.predict([" Priyanka Gandhi tutored kids to raise obscene slogans against PM Modi"])
array(['Fake'], dtype='<U4')

text_clf.predict([" delhi merto gets a tag of worlds green metro"])
array(['Real'], dtype=object)
```

**Fig 15 Prediction of a given News**

For better accuracy and reliable result the dataset needs to be updated on regular interval. Since the news base and current issues regularly add up to the content from conspiracy and manipulating stories can be made up. The combination of the both fake and real news stories need to be added in the database.

## VI. Conclusion and Future Scope

The purpose of the thesis is to detect whether a given unknown or unclarified news is fake or Real using Different Supervised Machine Learning Technique. We had a textual dataset with entries of News with Labels. In this Machine Learning Model, we used different classification algorithm. By using different algorithm, we got different but persistence results which suits our objectives. We used two different databases for our learning and understanding. One dataset was on US politics which was collected during presidential election and the second database was based on Indian politics. Since we had quite arranged and huge amount of data in Dataset-1 we reached an accuracy close to 99 percent. Since availability of data related to fake Indian news are less in number so data was arranged annually. Since the Dataset-2 had very less amount of data respect to Dataset-1 hence the accuracy could reach up to 85 percent.

On the basis of result, we can say that Linear model and DNN are the best performing algorithm for our Machine Learning Model which is based on text data. As the Amount of dataset increases we expect the performance of the classifier model to increase considerably. We need natural language processing tool for other Indian Language to cover the wide spread of fake news in groups of people speaking other languages. Implementing the model with other regional languages along with Updated dataset will improve the accuracy in future works.

## VII. References

[1] E. Alpaydin, "*Introduction to machine learning*", MIT press, 2009.

[2] Kotsiantis,B.Sotiris, I. Zaharakis,P.Pintelas, "*Fake news detection*

 

*on social media: A data mining perspective*," ACM SIGKDD Explorations Newslette, vol. 19, no. 1, pp. 22-36, 2017.

[3] Conroy,J. Niall,L. Victoria L. Rubin,Y. Chen, "*Automatic deception detection: Methods for finding fake news*," In the Proceedings of the 2015 Association for Information Science and Technology, vol. 52, no. 1, pp. 1--4, 2015.

[4] S. B. I. Z. a. P. P. Kotsiantis, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering , vol. 160, pp. 3-24, 2007.

[5] M. a. J. P. Stevanovic, "*An efficient flow-based botnet detection using supervised machine learning*," In the Proceedings of the 2014 International Conference on Computing, Networking and Communications, pp. 797-801, 2014.

[6] S. Koley, E. Koley,S. Ghosh,S. G. Shukla, "*Detection and Classification of Open Conductor Faults in Six-Phase Transmission Line Using Wavelet Transform and Naive Bayes Classifier*," International Conference on Computational Intelligence and Computing Research, pp. 1-6, 2017.

[7] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint* , vol. arXiv:1705.00648, pp. 12-13, 2017.

[8] Ruchansky, Natali, S.Seo,Y. Liu, "*Csi: A hybrid deep model for fake news detection*" In the Proceedings of the 2017 Information and Knowledge Management. ACM, 2017.

[9] M. M. V. Y. a. A. Granik, "*Determining Fake Statements Made by Public Figures by Means of Artificial Intelligence*" International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, pp. 424-427, 2018.

[10] A. P. Flach, "*The geometry of ROC space: understanding machine learning metrics through ROC isometrics*" In the Proceedings of the *2003 20th* International Conference on Machine Learning (ICML-03), pp. 194-201, 2003.

[11] H. a. A. G. Bhavsar, "A comparative study of training algorithms for supervised machine learning," *nternational Journal of Soft Computing and Engineering (IJSCE)*, vol. 2.4, pp. 2231-2307, 2012.

## Authors Profile

Anant Kumar pursed Bachelor of Engineering from Birla Institute of Technology, Mesra, India in the year 2016. He is currently pursuing Masters of Technology (cyber security) from Central University of Punjab, Bathinda, Punjab, India. He is currently working in the area of Data Science and software security.

Dr. Satwinder Singh had completed his Ph.D in 2014 from Guru Nanak Dev University, Amritsar. He is currently working as an Assistant Professor at Department of Computer Science and Technology, Central University of Punjab, Bathinda, Punjab, India, He has 15 years teaching experience. He has published research papers in reputed journals and conferences. His research interests include Re-engineering of Software System , Maintenance and Fault prediction of Object Oriented Systems, Big data analytics and Text Data Analytics.

Mrs. Gurpreet Kaur had completed her Ph.D in 2016 Punjabi University, Patiala. She is currently working in Department of Law, Bathinda College of Law, Bathinda, Punjab, India. She has 11 years of teaching experience. She has expertise in Criminal Law, International Law and Fake News.

 