Fake News Detection using Machine Learning

Pranita P. Deshmukh[1*], Sakshi A. Dulhani[2], Parmita C. Adkane[3], Priyanka Y. Belekar[4], Isha J. Raja[5]

[1]*Assistant Professor, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera, India*
[2,3,4,5]*B.E. Student, Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera, India*

***Abstract*: Most smart phone users prefer to read news stories through online forums. The news websites are publishing the news and provide the source of validation. The question is how the stories and articles that are distributed on social media such as what's App groups, Facebook pages, Twitter and other small blogs and social media sites are authorized. It is dangerous for the society to believe on the rumors and pretend to be news. The necessity for an hour to prevent rumors especially in developing countries like India, and to specialize in fair, proven issues. This paper deals with the revision of existing machine learning algorithms like Naïve Bayes, Logistic Regression, Support Vector Machine proposed to detect and reduce false information from various social media platforms. This paper provides a comparison of the results of existing fake news detection methods using different algorithms of machine learning.**

***Keywords*: Artificial Intelligence, Machine Learning, Naive Bayes, SVM, NLP, LR, Fake news detection.**

1. Introduction

Fake News is one of the most controversial stories that has attracted attention over the past year. The media reports that social media has played a key role in the outcome of the 2016 US elections. Propaganda, conspiracy theories and other myths have been widely used in the media for the second time as political gain and ideological fraud. Clearly, social media is a powerful tool for spreading lies. Modern life has become much more relevant and people around the world should appreciate the great contribution of internet technology to the transmission and sharing of information. There is no doubt that the internet has made our lives easier and access to more information has worked. This is an evolutionary process in human history, but at the same time it does not focus on the line between the real media and the brutally constructed media. Internet users can often follow their favorite events in online mode, and the proliferation of mobile devices makes this process even easier. But great opportunities come with great challenges. Many media outlets have a great influence on society, and as is often the case, there are those who want to take advantage of this fact. Sometimes to achieve certain goals, the mainstream media can use information in a variety of ways. This leads to the production of completely untrue or completely false headlines. Many scientists believe that counterfeit stories can be solved by means of machine learning and artificial intelligence. This is because recently the artificial intelligence algorithms have begun to improve the function in many classification problems (image recognition, voice detection and so on) because hardware is cheaper and larger databases are available.

Although many false stories are flawless, they are used for entertainment purposes only but readers do not understand the truth of these stories and change them according to the context. It is therefore very difficult for students to understand the motto of stories whether they are issued for entertainment purposes or for any other purpose. That is why it is so important to create such a model that can easily reflect the theme of the story so that readers are not distracted.

2. Related Work

[1] It shows a simple technique for fake news detection using Naive Bayes classifier. This approach was executed as a software system and tested against a dataset of Facebook news posts. The authors achieved classification accuracy of roughly 74% on the test set which a decent result is considering the relative simplicity of the model.

[2] It provides an ideal overview of satire and humor, developing and illustrating the unique features of satirical news, which copies the format and style of journalistic reporting. Satirical news stories were carefully matched and checked in contrast with their legitimate news counterparts in 12 contemporary news topics in 4 domains (civics, science, business, and "soft" news). The authors proposed an SVM-based algorithm, developed with 5 predictive features (Absurdity, Comedy, Grammar, Negative Affect, and punctuation) and tested their combinations in 360 news articles. Their best predicting feature combination (Absurdity, Grammar and Punctuation) detected satirical news stories with 84% recall and 90% precision (F-score=87%).

In [3], the authors proposed a unified model, i.e., TI-CNN, which can combine the text and image information with the corresponding explicit and latent features. The proposed model has a strong elasticity, which can easily absorb other aspects of the news. Alternatively, the convolutional neural network enables the model to detect all inputs at once, and can be trained much faster than LSTM and many other RNN models. Test results have shown that TI-CNN can effectively detect false stories based on explicit and implicit traits learned in convolutional neurons. The dataset in [3] focuses on the news about American presidential election.

In [13], the authors have used Logistic Regression classifier for classification of fake news. They have presented a detection

*Corresponding author: ppadeshmukh@gmail.com model for fake news using TF-IDF analysis through the lenses of different feature extraction techniques. They have investigated different feature extraction and machine learning techniques. The model achieved accuracy of approximately 72% when using TF-IDF features and logistic regression classifier. In [4], Jain et al. (2019) have demonstrated the model with support for ML and NLP techniques to collect articles using a Support Vector Machine (SVM) and resolved whether the stories were true or false. They have used a support vector machine algorithm for binary classification to organize the articles and based on that model works to categorize the articles either fake or real. The authors have used three main modules to purify their articles in their proposed models as an aggregator, authenticator and recommendation system. In [4], they have also used the Naïve Bayes algorithm to check whether the articles were fake or real and for obtained with 93.50% accuracy achieved by the combining these three algorithms i.e., Naive Bayes, SVM, and NLP.

3. Fake News Detection Techniques

How to detect fake news using different ML Algorithms?

There are many existing approaches for detection of fake news. But here we would see three types of existing

ML Algorithms which are used for detection of fake news.

1. Naïve bayes
2. Logistic regression
3. Support vector machine

1) *Naïve Bayes*

A Naive Bayes classifier is a supervised machine learning algorithm which uses Bayes theorem. The variables that are used to create the model are independent of each other. It is demonstrated that this classifier itself provides pretty good results.[5][7].

$$P((X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

The classification is managed by obtaining the maximum posterior, which is the maximal P(Ci |X) with the above consideration applying to Bayes theorem. This consideration greatly reduces the computational cost by only counting the class distribution. Naive Bayes algorithm can be used to detect whether a news article is fake or real.

2) *Logistic Regression*

Logistic regression is Supervised Machine Learning algorithm. It is used for identifying the categorical dependent variable using a given set of independent variables. An example of logistic regression could be applying machine learning to determine whether a given news article is fake or not. Since we have two possible outcomes to this question - yes the news is fake, or no the news is not fake - this is called binary classification. Logistic regression predicts the output in the form of either Yes or No, 0 or 1, true or false, etc. but rather than giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, "S" shaped logistic function is fitted, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something like whether the cells are cancerous or not, news is real or fake, etc. Logistic Regression is a significant supervised machine learning algorithm because it has the capability to provide probabilities and classify new data using continuous and discrete datasets. The below image is showing the logistic function:

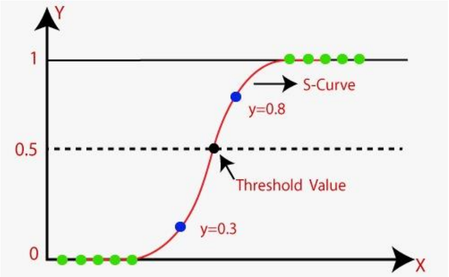


Fig. 1. Logistic regression

3) *Support Vector Machine*

Support Vector Machine is a Supervised Machine Learning algorithm, which can be used for Classification as well as Regression problems. Though, primarily, it is used for Classification problems in Machine Learning. The objective of the SVM algorithm is to create the best line or decision boundary that can separate n-dimensional space into classes so that we can easily put the new data point in the accurate class in the future. This best decision boundary is termed as hyperplane. SVM chooses the extreme points/vectors which helps in creating the hyperplane. These extreme cases are known as support vectors, and therefore algorithm is named as Support Vector Machine. Consider the below diagram in which there are two different classes that are classified using a decision boundary or hyperplane:

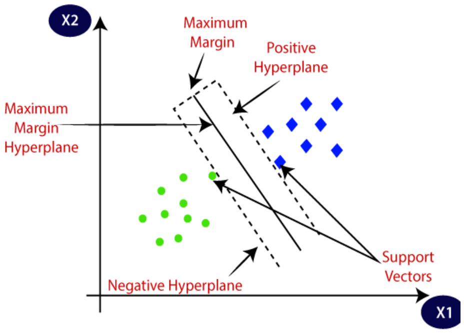


Fig. 2. Support Vector Machine

4. Comparative Study of Different Approaches

Table 1 shows a comprehensive summary of ML approaches for fake news detection. All these results shown in table 1 are obtained by training the model using different ML algorithms. Using Naïve Bayes, accuracy achieved on test set is 88.37%. Using Logistic Regression, accuracy achieved on test set is 92.18%. And using SVM, accuracy achieved on test set is 93.27%. Among all the approaches highest accuracy is obtained by using Support Vector Machine Algorithm.

Table 2
Results

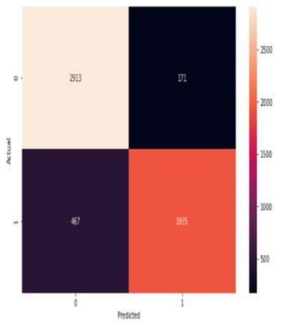
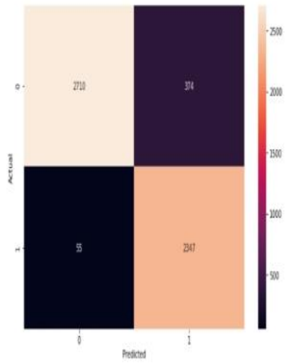
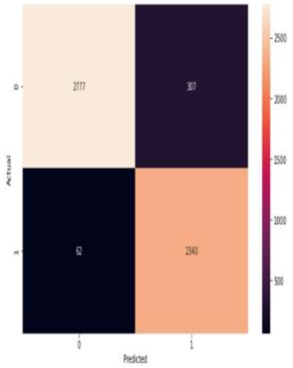
Algorithm and Results	Confusion Matrix
Multinomial NB Algorithm Accuracy of NB classifier on training set: 91.68 Error Rate of NB classifier on training set: 0.08 Accuracy of NB classifier on test set: 88.37 Error Rate of NB classifier on test set: 0.12 Precision: 0.8867777713062213 Recall: 0.8837039737513671 F1_Score: 0.8825486327498272 True positive = 2913 False positive = 171 False negative = 467 True negative = 1935 [[2913 ,171] [467, 1935]]	
Logistic Regression Accuracy of LR classifier on training set: 94.58 Error Rate of LR classifier on training set: 0.05 Accuracy of LR classifier on test set: 92.18 Error Rate of LR classifier on test set: 0.08 Precision: 0.9286367100289689 Recall: 0.9218009478672986 F1_Score: 0.9221031467259588 True positive = 2710 False positive = 374 False negative = 55 True negative = 2347 [[2710, 374] [55, 2347]]	
Support Vector Machine Accuracy of SVM classifier on training set: 96.43 Error Rate of SVM classifier on training set: 0.04 Accuracy of SVM classifier on test set: 93.27 Error Rate of SVM classifier on test set: 0.07 Precision: 0.9369421639600084 Recall: 0.9327378782355086 F1_Score: 0.932978686274408 True positive = 2777 False positive = 307 False negative = 62 True negative = 2340 [[27 77, 307] [62, 2340]]	

Table 1
Result comparison

Implementation Method	Accuracy on Test Set
Naïve Bayes	88.37%
Logistic Regression	92.18%
SVM	93.27%

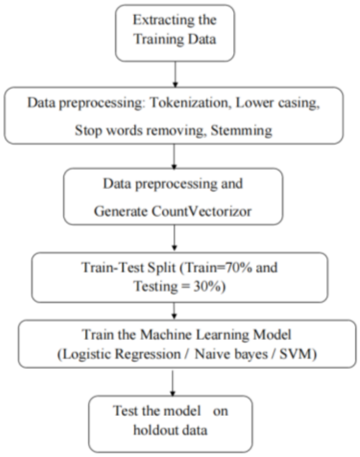

Extracting the Training Data
Data preprocessing: Tokenization, Lower casing, Stop words removing, Stemming
Data preprocessing and Generate CountVectorizor
Train-Test Split (Train=70% and Testing = 30%)
Train the Machine Learning Model (Logistic Regression / Naive bayes / SVM)
Test the model on holdout data

Fig. 3. Flow chart – Classifier Training

5. Conclusion

It is significant to identify whether a news article is fake or real. Machine learning-based classification algorithms play a very important role in the detection of fake news from social media, which is a very complicated and difficult process due to the diverse social, political and economic, and many other related factors. In the paper, different approaches for detection of fake news are discussed.

References

[1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc., pp. 900–903, 2017.

[2] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," pp. 7–17, 2016.

[3] Y. Yang, L. Zheng, J. Zhang, Q. Cui, X. Zhang, Z. Li, Philip S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection,"2018.

[4] Jain, Anjali, et al. "A smart System for Fake News Detection Using Machine Learning." 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). Vol. 1. IEEE, 2019.

[5] Y. Seo, D. Seo, and C. S. Jeong, “FaNDeR: Fake News Detection Model Using Media Reliability,” IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2018–October, no. October, pp. 1834–1838, 2019.

[6] R. V. L, C. Yimin, and C. N. J, “Deception detection for news: Three types of fakes,” Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, 2016.

[7] S. Gilda, “Evaluating machine learning algorithms for fake news detection,” IEEE Student Conf. Res. Dev. Inspiring Technol. Humanit. SCOReD 2017 - Proc., vol. 2018– January, pp. 110–115, 2018.

[8] Saxena, Deepika, and Ashutosh Kumar Singh. "Auto-adaptive learning-based workload forecasting in dynamic cloud environment." International Journal of Computers and Applications (2020): 1-11.

[9] D. Saxena, and A.K. Singh, "A proactive autoscaling and energy efficient VM allocation framework using online multi-resource neural network for cloud data center", Neurocomputing 426 (2021): 248-264.

[10] Kumar, Jitendra, Deepika Saxena, Ashutosh Kumar Singh, and Anand Mohan., "Biphase adaptive learning-based neural network model for cloud datacenter workload forecasting." Soft Computing (2020): 1-18.

[11] Saxena, Deepika, Ashutosh Kumar Singh, and Rajkumar Buyya. "OP-MLB: An Online VM Prediction based Multi-objective Load Balancing Framework for Resource Management at Cloud Datacenter." IEEE Transactions on Cloud Computing 01 (2021): 1-1.

[12] Murari Choudhary, Shashank Jha, Prashant, Deepika Saxena, Ashutosh Kumar Singh. “A Review of Fake News Detection Methods using Machine Learning,” 2021 2nd International Conference for Emerging Technology (INCET), 2021.

[13] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed, “Fake News Detection Using Logistic Regression.” International Research Journal of Engineering and Technology, Vol. 6, 2019.