International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

FAKE NEWS DETECTION USING MACHINE LEARNING

MR.R.PRAVEEN KUMAR[1], P.DINESHWARAN[2], RAVIKUMAR R[3]

[1]Assistant Professor, Department of Information Technology, Panimalar Institute of Technology, Chennai, India

[2,3]Student, Department of Information Technology, Panimalar Institute of Technology, Chennai, India.

Abstract: Fake news is one of the most serious problems in the modern age of the internet and social media. Since the internet is pervasive, everybody depends on different online resources for news. Today, social networks have become an integral part of many people's daily routines, and as a result, practises such as networking, marketing, advertising, news, and agenda development have begun to be conducted through them. With the rise in popularity of social media sites like Facebook, Instagram, and Twitter, etc.., Fake news quickly spread across millions of users in a very short period of time. To solve this problem, machine learning techniques based on Natural Language Processing, as well as other algorithms, will be used. We suggest using NLP (Natural Language Processing) methods to detect 'fake news,' that is, news reports that are unreliable and come from untrustworthy sources, in this framework. We proposed such a method based on a count vectorizer or a TFIDF matrix(Term Frequency Inverse Document Frequency). Two posts with similar word counts are very likely to have slightly different meanings. Combating fake news is a clear text recognition initiative with a basic proposition. As a result, the proposed system will compile a dataset of both false and true news, then use a machine learning algorithm based on naive bayes and support vector machines to build a model that can mark an article as fake or true based on its terms and phrases. As a result, the user has the ability to identify news as false or actual, as well as verify the authenticity of the website that published the news.

Keywords:

Internet, Social Media, Fake News, Classification, Artificial Intelligence, Machine Learning, Authenticity.

Introduction:

Anyone in today's world can publish content on the internet. Fake news, unfortunately, attracts a lot of attention on the internet, especially through social networking sites. People are misled, and they don't think twice about sending such erroneous information to the farthest reaches of the system. Such practises are harmful to society because rumours or ambiguous news may dispel negative thoughts among people or a particular group of people.Preventive measures are expected to deal with such activities at the same rate as technological advances. Broad communications plays a huge role in influencing the general population, and as is customary, a few individuals try to profit from it. There are several websites that provide false information. Under the pretence of being true news, they deliberately aim to carry out purposeful advertising, deceptions, and falsehood. Their primary responsibility is to maintain control over the data that allows others to have confidence in it. There are numerous examples of such sites all over the world. As a result, fake news has an effect on people's minds.According to a report, scientists believe that a variety of man‐made brainpower measurements will aid in the discovery of false news.In this system we use machine learning algorithms such as Natural Language Processing(NLP), Navie Bayes and Support Vector Machine (SVM) in order to avoid the spread of rumours through Twitter platforms. This is done to prevent the spread of false news, which leads to activities such as mob lynching. We've been getting reports of mob lynchings that result in a person's death on a regular basis. This framework is designed to detect fake news and stop similar events, thus shielding society from these unnecessary acts of violence.In order to detect fake news, we use machine learning algorithms such as Navie Bayes and Support Vector Machine (SVM).

Proposed work:

The key goal of this system is to identify fake news, which is a big issue in today's world. It is necessary to develop a model that can distinguish between "true" and "fake" news. This has consequences in social networking sites like Twitter, where fake news gets a major boost and spreads around the country and across the world. The proposed method assists in assessing the authenticity of news. If the news is incorrect, the consumer is directed to the appropriate news article.

Methodology:

1. Data collection: William Yang Wang's "A New Benchmark Dataset for Fake News Detection" was used to introduce and validate the proposed framework. This repository's dataset is divided into three sets: training, validation, and test. There are 12,836 short statements in the dataset that have been labelled for truthfulness, topic, context/venue, speaker, state, group, and prior history. The dataset initially classified the news into six fine-grained labels for the truthfulness ratings: pants‐fire, false, barely-real, half-true, mostly‐true, and true.

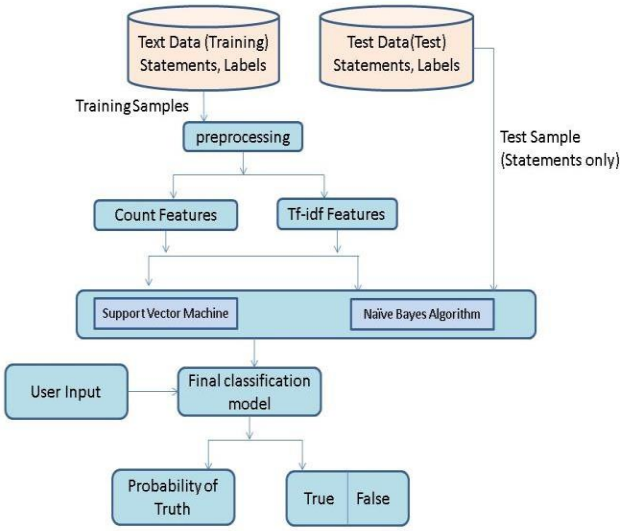
2. Data preprocessing: The six-label classification problem was translated into a binary classification problem with True and False labels for the proposed scheme. The labels were converted using the following mapping: • pants-fire --> False • false --> False • barely-true --> False • half-true --> True • mostly-true --> True • true --> True In addition, only the news headline was used as an input for classification. Thus, in the preprocessing stage, the labels were first mapped using the above-mentioned mapping, after which only the labels and news statement columns were extracted from the dataset and saved in csv format for future

use. Following the preprocessing, we were able to obtain the following three cleaned files: • train.csv • valid.csv • test.csv

3. Feature extraction: The following feature extraction method is used to help machine learning models gain insights from news headlines:

Count Vectorizer: First, the English stopwords were stripped from all of the news headlines using scikit‐learn's CountVectorizer, and then they were tokenized using spaces and punctuation marks as the delimiter. After all of the headlines had been tokenized, a sparse matrix of all of the news headlines as rows and the tokens as columns was restored. In addition to their morphological use, a number of n-grams were returned to make the tokens reflect the sense in which they were used.

4. Modelling: Logistic Regression, Random Forest Classifier, Nave Bayes, SVM Classifier, and voting classifier were the models used for training. The features extracted from the CountVectorizer are used to train the models. After that, using GridSearchCV and a 5-hold out cross validation set, all of the models were hyperparameter tuned for all of the different possible parameters. The aim of this hyperparameter tuning was to boost the models' f1-score. After the models were fine-tuned, they were evaluated on a test range, and evaluation metrics for the models were determined.



Output screenshot:

```
In [56]: train_data.sample(5)
```

Out[56]:

	label	news
3842	True	Polling shows that nearly 74 percent of Nation...
6480	False	I left the city with $43 million in the bank.
4521	False	Says she couldn't take stimulus money because ...
4026	True	The United States is the only industrialized c...
10111	False	The Health Care and Education Reconciliation A...

Train Data:

```
In [35]: show_eval_scores(svm_pipeline, test_data, 'SVM Classifier Count Vectorizer')

Report for ---> SVM Classifier Count Vectorizer
Accuracy is: 0.5666929755327546
F1 score is: 0.7211782630777045
Precision score is: 0.5657370517928287
Recall score is: 0.9943977591036415
```

Output using SVM Classifier:

Output using Nave Bayes Classifier:

```
In [28]: show_eval_scores(nb_pipeline, test_data, 'Naive Bayes Count Vectorizer')

Report for ---> Naive Bayes Count Vectorizer
Accuracy is: 0.6203630623520127
F1 score is: 0.7326292384658143
Precision score is: 0.6073732718894009
Recall score is: 0.9229691876750701
```

Conclusion:

The majority of activities in the twenty‐first century are done online. Newspapers, which were once preferred in hard copies, are now being replaced by applications such as Facebook, Twitter, and news articles. The growing issue of fake news further complicates matters by attempting to sway people's opinions and attitudes about the use of digital technology. When a person is misled by real news, one of two things can happen: people begin to believe that their assumptions about a specific topic are right. To fix this issue, we built our Fake News Detection system, which takes user feedback and classifies it as true or false. To do so, various NLP and Machine Learning algorithms must be used, as well

as a suitable dataset for training the model. We also used the TF-IDF vectorizer to vectorize the text data. Various performance metrics are often used to evaluate performance.

References:

[1] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti and M. Akbar, "Fake news detection using deep learning models: A novel approach", Transactions on Emerging Telecommunications Technologies, 2019.

[2] Ghinadya Suyanto, "Synonyms-Based Augmentation to Improve Fake News Detection using Bidirectional LSTM", 8th International Conference on Information and Communication Technology (ICoICT), 13 August 2020.

[3] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms", Physica A: Statistical Mechanics and its Applications, vol. 540, pp. 123174, 2020.

[4] Anmol Uppal; Vipul Sachdeva; Seema Sharma, "Fake news detection using discourse segment structure analysis", 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 09 April 2020.

[5] M. G. Hunt Allcott, "Social media and Fake News in the 2016 Election", Journal of Economic Perspectives, vol. 31, no. 2, pp. 211-236, 2017.

[6] P. P. J. Monu Waskale, "A Review Rumors Detection on Twitter Using MachineLearningTechniques", Internation al Journal of Scientific Research & Engineering Trends, vol. 05, no. 03, 2019.

[7] S. S. Y. L. Natali Ruchansky, "CSI: A Hybrid Deep Model for Fake News Detection", CIKM 2017 Internationall Conference on Information and Knowledge mangaement, 2017.

[8] A. A. Monther Aldwairi, "Detecting Fake News in Social Media Networks", The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, 2018.

[9] J. S. G. N. S. Terry Traylor, "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator", IEEE 13th International Conference on Semantic Computing 2019.

[10] S. S. S. S. P. W. S. H. Subhadra Gurav, "Survey on Automated System for Fake News Detection using NLP & MachineLearning Approach", International Research Journal of Engineering and Technology, vol. 06, no. 01, 2019.

[11] V. A. P. S. R. Sivasangari V, "A Modern approach to identify the fake news using machine learning", International Journal of Pure and Applied Mathematics, vol. 118, no. 20, 2018.

[12] A. S. S. W. J. T. H. L. Kai Shu, "Fake News Detection on Social Media: A Data Mining Perspective", ACM SIGKDD Explorations Newsletter, pp. 22-36, 2017.

[13] D. B. S. Z. Oluwaseun Ajao, "Fake News Identification on Twitter with Hybrid CNN and RNN Models", SMSociety '18 Proceedings of the 9th International Conference on Social Media and Society, 2018.

[14] A. A. I. Costel-Sergiu Atodiresei, "Identifying Fake News and Fake Users on Twitter", Procedia Computer Science, pp. 451-461, 2018.

[15] K. S. S. W. R. G. F. W. H. L. Shuo Yang, "Unsupervised Fake News Detection on Social Media: A Generative Approach", AAAI Conference on Artificial intelligence 2019.