International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Fake News Detection Using Machine Learning

*Ankit Kumar, Dr. Deepika Bansal**

ABSTRACT:

Nowadays people are heavily dependent on social media for news. On social media, there is a lot of news created by humans and machines. And this widely circulated news could be true or false. Fake news has a significant political and social impact on people's lives, and its dissemination creates fear in them. In today's world, detecting fake news is one of the most important tasks. There are many models and strategies developed to detect fake news. In this project, different machine learning techniques are used to detect fake news and their accuracy is compared.

Keywords: Term Frequency-Inverse Document Frequency., Bag of words, Support Vector Machine, Logistic Regression, Random Forest Classifier, K-Nearest Neighbor, Decision Tree Classifier

1. Introduction

With the arrival of the digital era, the intake of online information has grown, and with this, the problem of fake news propaganda has risen. Any misinformation that is presented as news is referred to as fake news. This information might be provided willingly or unwillingly. At times, fake news is released with the intention of influencing and changing people's opinions, evoking strong emotions in readers, and spreading false beliefs. Fake news can be given through other than fake movies, text and pictures. Fake news can be employed to incite panic and cause mass hysteria. Fake news can also be employed to undermine science, as in the instance of the anti-vax campaign. There is a great deal of misinformation on the Internet concerning different diseases and their cures, which can cause health issues or even lead to death.

Online news is gaining popularity, and its rate of consumption is extremely high due to the continuous and fast growth of social media and technology. Social media is also popular due to its ease of use and affordability. Almost every smartphone owner spends time on their phone. The low expense of opening a social media user account leads to numerous unverified accounts, social bots and cyborgs that help spread false news. Individuals post unverified news to boost public activity on social media, which aids in the spread of rumors. With all this information on social media platforms, it is becoming more challenging to distinguish between real and false news. Social bots and clickbait help the dissemination of pseudo-news around the Internet, and the comments beneath these news stories give them credence.

There is a need of the hour to find fake news on time. Several organizations waste vast amounts of time and money manually annotating news. This highlights the necessity of having a trustworthy automated system for finding fake news. Most of the available work in this area is recent work, which describes patterns that occur repeatedly in propaganda news after they have been already spread or suggests new features to train a classifier on, grounded in ideas that have not been tested together. It is therefore hard to estimate the potential of supervised models trained from features suggested in recent research to find propaganda news.

In this paper simple fake news detection methods based on different machine learning algorithms and using two different feature extraction methods – Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-words(bow) are presented as their accuracy on the ISOT dataset is compared. The purpose of the research is to analyze the impact of different combinations on the fake news detection accuracy.

2. Data Set

The ISOT dataset has been used here. The ISOT Fake News dataset is a collection of various thousand fake news and true articles, which were collected from different recognized news websites and websites rated as unreliable by Politifact.com. About the fake news articles, they were collected from different sources. The fake news articles were collected from untrusted websites that were marked by PolitiFact (a USA-based fact-checking organization) and Wikipedia. The dataset includes different types of articles on different topics, but most articles are based on political and World news topics.

The dataset includes two CSV files. The first file named “True.csv” includes more than 12,600 articles from reuter.com. The second file named “Fake.csv” has more than 12,600 articles from different fake news resources. Every article has the following information: article title, text, type and the date on which the article was published.

This dataset includes 44898 data in total, 21417 are real news (labeled as 1) and 23481 are fake news (labeled as 0). The file includes title, text (news body), subject, date and label. The news subjects have different categories like ’politics News’, ‘World news’, ‘News’, 'politics', ’Government News’, ’left news’, ’US News’, ’Middle east’. We have chosen features like title and text in the ISOT dataset to train our models.

3. Feature Extraction Techniques

3.1. TF-IDF

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It refers to the measurement of how specific a word in a series or corpus is for a text. The significance goes up in direct proportion to how many times in the text a word appears but is offset by the word frequency in the corpus (dataset).

The importance of a term is inversely proportional to its frequency in documents.TF gives us information about how often a term occurs in a document and IDF gives us information with the relative rarity of a term in the collection of documents. By multiplying these two values together we obtain our final TF-IDF value.

The higher the TF-IDF score the more important or related

$$tf\ idf\ (t, d, D) = tf\ (t, d).idf(t, D) \qquad (1)$$

The term is as a term gets less relevant, its TF-IDF score will approach 0**.**

3.2. Bag-of-words

A bow model is a way of extracting features from the text for the purpose of modelling, e.g., with machine learning algorithms.
The method is very simple and flexible and can be used infinitely in different manners for extracting features from documents.
A Bow is a text representation that captures the presence of words in a document. It involves two things:

- A vocabulary of known words.
- A measure of the presence of known words

It is called a “bag” of words because any information about the order or arrangement of words in the document is ignored. The model is only interested in whether known words appear in the document, not where in the document.
In practice, the Bow model is manually used as a tool for feature generation. Once the text has been converted into a “bag of words”, we are able to calculate different measure to describe the text. The most common type of characteristics or feature calculated from the bow model is term frequency, i.e., the number of times a term appears in the text.

4. Algorithms

4.1. Logistic Regression

One popular machine learning technique that falls under the category of supervised learning is logistic regression. It is employed to forecast the categorical dependent variable from a collection of independent variables. It forecasts a categorical dependent variable's result. The outcome must therefore be discrete or categorical value. Instead of displaying the precise values as 0 and 1, it displays the probability values that fall between 0 and 1. It can be Yes or No, 0 or 1, true or False, and so forth. It can classify new data using both continuous and discrete datasets and provide probabilities.

In its most basic form, logistic regression is a statistical model that models a binary dependent variable using a logistic function; however, there are many more intricate variations. The dependent variable in a binary logistic model, which is represented by an indicator variable with the labels "0" and "1," has two possible values, such as pass/fail. The log-odds (the logarithm of the odds) for the value "1" in the logistic model is a linear combination of one or more independent variables (also known as "predictors"); these independent variables can be either +continuous (any real value) or binary (two classes, coded by an indicator variable).

4.2. Support Vector Machine

A well-liked machine learning technique for classification, regression, and other learning tasks is the Support Vector Machine (SVM). The SVM algorithm plots each data item as a point in n-dimensional space, where n is the number of features you have. Each feature's value corresponds to a specific coordinate. Next, we classify by finding the hyper-plane that best distinguishes the two classes.

The selection of the kernel function is the central problem in SVM research. Nevertheless, there isn't a more practical or effective way to create a suitable kernel function that references issues. In practice, the more commonly used kernel functions are Linearity kernel function, polynomial kernel function, RBF (Radial Basis Function) kernel function and Sigmoid kernel function.

4.3. Decision Tree

A decision tree is a structure that resembles a flowchart, with each internal node standing for a feature test, each leaf node for a class label (a choice made after all features have been calculated), and branches for feature conjunctions that result in those class labels. Classification rules are represented by the routes from root to leaf. One predictive modelling technique used in machine learning, data mining, and statistics is the decision tree.

An algorithmic method that finds ways to divide a data set according to various conditions is used to build decision trees. It is among the most popular and useful supervised learning techniques. A non-parametric supervised learning technique for classification and regression problems is the decision tree. Decision trees try to categorize instances after learning the splits. Misclassification can happen in a variety of splits. During classification, decision trees may perform poorly at some splits and well at others. Gradient Boosting is a collection of numerous decision trees. The model's overall classification performance is improved by combining many weak learners. A collection of classification and regression trees (CART) makes up the tree ensemble model. The prediction scores at each level are summed up to get the final predicted score.

4.4. Random Forest

One popular machine learning algorithm from the supervised learning approach is Random Forest. It can be used for machine learning problems including both regression and classification. The basis of it is the idea of ensemble learning, which is the process of combining multiple classifiers to solve a challenging issue while improving the model's functionality. A classifier called Random Forest uses several decision trees on different subsets of the dataset and averages them to increase the dataset's predictive accuracy.

In the first stage, N decision trees are combined to create the random forest; in the second stage, predictions are made for each tree produced in the first phase.

The Working process can be explained in the below steps-

- Select random K data points from the training set
- Build the decision trees associated with the selected data points (Subsets).
- Choose number N for decision trees that you want to build.
- Repeat Step 1 & 2.

For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority vote.

4.5. K Neighbors Classifier

One of the most basic machine learning algorithms based on the supervised learning approach is K-Nearest Neighbor. This algorithm places the new case in the category that is most similar to the existing categories based on assuming that the new case and data are similar to the available cases. It keeps track of all the data that is accessible and uses similarity to categorizing new data points. This implies that the K-NN algorithm makes it simple to classify newly discovered data into a well-suited category. Although it can be used for both classification and regression, it is mostly used for classification problems.

Since K-NN is a non-parametric algorithm, it doesn't assume anything about its base data. Because it stores the dataset and then performs an action on it at the time of classification, it is also known as a lazy learner algorithm. This is because it does not learn from the training set right away.

During the training phase, it simply stores the dataset and classifies new data into a category that is very similar to the new data.

5. Results

Table 1 shows the accuracy of the machine learning algorithms following the use of TF-IDF and bow as feature extraction techniques. It is discovered that the accuracy obtained with bow is marginally higher than that obtained with TF-IDF for feature extraction.

Table 1 – Accuracy found with different machine learning algorithms.

Machine Learning Algorithm	Column A (*t*)	Column B (*t*)
Logistic Regression	99.65	98.63
Random Forest Classifier	99.85	99.75
Decision Tree Classifier	99.65	99.59
K Neighbors Classifier	80.54	87.7
Support Vector Machine	99.51	99.38

6. Conclusion

Manually classifying news calls for in-depth subject-matter knowledge as well as the ability to spot textual irregularities. The issue of using machine learning models to classify fake news articles was covered in this study. Most of the news is covered by news articles from a variety of domains in the ISOT dataset, which is where we gathered the data for our study. Comparing the accuracy of various algorithms combined with various extraction methods to distinguish between bogus and authentic news is the main goal of the study. Comparatively speaking, some models have reached greater accuracy than others. Based on the results, we discovered that using as the feature extraction method produced higher accuracy than using TF-IDF.

Researchers need to focus on the many unresolved problems with fake news detection. For example, finding the key factors that contribute to the spread of news is a crucial step in reducing the spread of fake news. To find the main sources of fake news, machine learning methods and graph theory can be used. Similarly, finding fake news in videos in real time may be another avenue for future research.

REFERENCES

1. Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. SN Applied Sciences, 2(4), 525.
2. Reddy, H., Raj, N., Gala, M., & Basava, A. (2020). Text-mining-based fake news detection using ensemble methods. International journal of automation and computing, 17(2), 210-221.
3. Granik, M., & Mesyura, V. (2017, May). Fake news detection using naive Bayes classifier. In 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON) (pp. 900-903). IEEE.
4. Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: a deep learning approach. SMU Data Science Review, 1(3), 10.
5. Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.
6. Kumar, S., Kumar, S., Yadav, P., & Bagri, M. (2021, March). A survey on analysis of fake news detection techniques. In 2021 International conference on artificial intelligence and smart systems (ICAIS) (pp. 894-899). IEEE.
7. Patel, A., & Meehan, K. (2021, June). Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine. In 2021 32nd Irish signals and systems conference (ISSC) (pp. 1-6). IEEE.
8. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1 (pp. 127-138). Springer International Publishing.
9. Poddar, K., & Umadevi, K. S. (2019, March). Comparison of various machine learning models for accurate detection of fake news. In 2019 Innovations in Power and Advanced Computing Technologies (i-PACT) (Vol. 1, pp. 1-5). IEEE.
10. Babu, D. J., Sushmitha, G., Lasya, D., Krishna, D. G., & Rajesh, V. (2022, March). Identifying fake news using machine learning. In 2022 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1-6). IEEE.
11. Bali, A. P. S., Fernandes, M., Choubey, S., & Goel, M. (2019). Comparative performance of machine learning algorithms for fake news detection. In Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part II 3 (pp. 420-430). Springer Singapore.
12. Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
13. Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning.
14. Sangamnerkar, S., Srinivasan, R., Christhuraj, M. R., & Sukumaran, R. (2020, June). An ensemble technique to detect fabricated news article using machine learning and natural language processing techniques. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.
15. Jiang, T. A. O., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). A novel stacking approach for accurate detection of fake news. IEEE Access, 9, 22626-22639.
16. Tiwari, V., Lennon, R. G., & Dowling, T. (2020, June). Not everything you read is true! Fake news detection using machine learning algorithms. In 2020 31st Irish signals and systems conference (ISSC) (pp. 1-4). IEEE.