International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Fake News Detection Using Machine Learning

Preeti Barla

Department of Computer Science and Engineering

Email-preetibarla6@gmail.com

Smruti Ranjan Swain

Department of Computer Science and Engineering

Email - smrutiranjan@gift.edu.in

ABSTRACT-

The rapid use of social media sites like Facebook and Twitter, along with the advent of the Internet, has allowed for the dissemination of information at a level never before seen... More people than ever before are making and sharing content on social media, and unfortunately, some of it is false or otherwise unfounded. It is difficult to automate the process of determining if a written article contains misinformation or disinformation. Prior to reaching a conclusion on an article's veracity, even a domain expert must consider several factors. Automated news article categorization is our proposed usage of a machine learning ensemble technique in this study. In this study, we examine various linguistic characteristics that can be used to distinguish between real and fake news. Taking use of these features, we evaluate the performance of a variety of machine learning algorithms trained using various ensemble methods on four real-world datasets. Results from experiments show that our suggested ensemble learner method outperforms individual learners.

Keywords: World Wide Web, Social Media platforms, Information distribution, Content Sharing Textual Features, Machine Learning, Machine Learning ensemble technique, Real-worlds dataset etc.

INTRODUCTION

Social media sites like Twitter and Facebook, made possible by the expansion of the Internet, allowed for the unprecedented dissemination of information. The ability to provide members with near-instantaneous updates on breaking news items is just one way that news organizations have profited from the social media boom. In recent years, internet news platforms, blogs, and social media feeds have largely supplanted traditional news outlets such as newspapers, magazines, and tabloids. Customers may have an easier time obtaining the latest news. Facebook recommendations drive most visitors to news websites. Users are able to argue, talk, and exchange views on subjects like democracy, education, and health on these social media platforms, which makes them very powerful and useful. Nevertheless, there are some who see these platforms negatively and utilize them for financial gain, skewed opinion creation, mind manipulation, or the dissemination of satire or absurdity. "Fake news" is a popular term for this situation. The 2016 US elections were the most visible example of the dramatic rise in the dissemination of disinformation throughout the past decade. Many issues have arisen in fields as diverse as politics, sports, health, and science as a result of the widespread distribution of misinformation online
True and fake news can be disseminated on the same network, but this approach takes that into account, making it more accurate. One possible way to gauge an article's credibility is to look at the number of responses it getsSeveral studies have investigated methods for detecting and categorising false news on social media sites like Twitter and Facebook... In theory, there are numerous forms of deceit. . Support vector machine (SVM), logistic regression (LR), decision tree (DT), and random gradient descent (SGD) are just a few of the machine learning models that may be trained using it. The most effective methods were logistic regression and support vector machines, which had an accuracy rate of 92%. Overall accuracy decreased as the number of n-grams computed for a particular article increased, according to the research.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

REVIEW OF LITERATURE

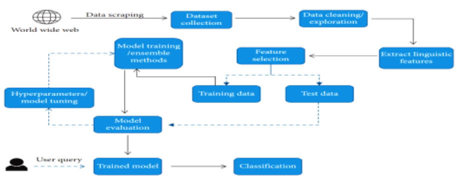
Using metadata and textual features, the author trained many ML models. The essential focus of the writer was on the application of CNNs. An additional convolutional layer is inserted following a bidirectional long short-term memory (LSTM) layer in order to record the interdependencies among the metadata vectors When doing research online, a dataset containing party declarations is used. Subject, speaker, profession, authority, party affiliation, location, and history are other considerations that fall under metadata.

Ultimately, the accuracy rate dropped to 27.7% when only speaker and text were included, and to 27.4% when all metadata components were considered in addition to text. "Agree," "disagree," "discuss," and "unrelated" are the labels assigned to each item based on how relevant the title is to the content. Authors evaluated text linguistic metrics, such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF), using a one-hidden-layer multilayer perceptron (MLP) classifier. After that, the final layer's output was processed using the SoftMax function. All of the articles in the database include titles, bodies, and labels.

Methodologies that relied on networks included things like processing and analysis of linked data and social network activities. The writers took a unique approach to studying the features of social media news. Using Twitter and other social media sites, they compared the spread of false information with that of true news. This study looks into different analytical approaches to try to figure out how fake news travels on the internet.

2.2 Algorithms

We evaluated classifiers for identifying fake news using the following learning algorithms as part of our proposed method.



"Figure: Algorithm training and news article classification workflow"

- ***"Logistic Regression":***

We apply a logistic regression (LR) model to classify texts using a wide range of features that only offer a true/false or legitimate article/fake article output. We tested several settings and adjusted the hyperparameters to get the best result for each dataset, all with the goal of getting the most accurate LR model possible.

The following are the mathematical formulations of the logistic regression hypothesis function.

$$h_\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}. \qquad (1)$$

- ***Support Vetor Machine:***

Support vector machine (SVM) is one of several kernel functions at our disposal; it is a model for binary classification issues... The objective of a support vector machine (SVM) model, Finding the plane that clearly separates two data sets with the widest feasible margin is the difficulty, since a hyperplane can be placed anywhere in an N-dimensional space... Mathematical representations of the SVM model's cost function are provided in and displayed.

$$J(\theta) = \frac{1}{2}\sum_{j=1}^{n}\theta_j^2, \qquad (3)$$

- ***Multilayer Perceptron:***

The three main components of an artificial neural network called a multilayer perceptron (MLP) are an input layer, a hidden layer (or layers), and an output layer. Even though a basic MLP just requires the three layers, we were able to get the best results from our trials by experimenting with different parameters and layer counts to fine-tune the model. Here is a function representation of a simple one-hidden-layer multilayered perceptron model:

$$f(x) = g\left(b^{(2)} + W^{(2)}\left(s\left(b^{(1)} + W^{(1)}x\right)\right)\right). \qquad (6)$$

- ***K-Nearest Neighbors (KNN):***

An MLP is a type of artificial neural network that primarily consists of three layers: an input layer, one or more hidden layers, and an output layer. We were able to optimise the model by testing with various parameters and layer counts, even though a simple MLP only requires the three layers. This allowed us to obtain the best results from our experiments. A simple one-hidden-layer multi-layer perceptron model is shown in the following flow diagram.:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}, \qquad (7)$$

$$\text{Manhattan distance} = \sum_{i=1}^{k}|x_i - y_i|, \qquad (8)$$

$$\text{Minkowski distance} = \left(\sum_{i=1}^{k}|x_i - y_i|^q\right)^{1/q}. \qquad (9)$$

2.3 Ensemble Learners:

$$G_{\text{ind}} = 1 - \sum_{i=1}^{c} (P_i)^2, \qquad (10)$$

To enhance the overall accuracy of truthfulness/falsity article categorization, our concept was to combine feature input from textual characteristics with current ensemble techniques.. Ensemble learners improve accuracies by training many models using the same technique, which decreases the overall error rate and increases the model's performance.

2.4 Benchmark Algorithms:

Here we go over the benchmark algorithms that we used to evaluate our technique.

- Linear SVM

As suggested in we employ a linear SVM strategy. Training the linear SVM on the feature set allowed us to make a valid comparison, as previously mentioned.

- Convolutional Neural Network

Automatically detecting false news was accomplished by use of a convolutional neural network (CNN). When working with our dataset, we used the identical procedure.

2.5 Datasets:

All of the datasets used for this study are available to the public and were made available as open-source. You can carefully check the assertions made in those political pieces with fact-checking websites like snopes.com and politifact.com...
"ISOT Fake News Dataset," or DS1, is the first database to compile both real and fake items retrieved from the web. While reuters.com is a reliable source for real news, PolitiFact has already identified a number of websites from where the misleading tales were culled. The collection contains 44,898 elements, out of which 21,417 are factual and 23,481 are pseudo-facts. Although the total corpora cover a wide range of topics, their main emphasis is on political news. You may find the second dataset, DS2, on Kaggle. The training set has 20,386 articles, whereas the testing set contains 5,126 articles.

2.6 Performance Metrics:

In order to determine how effective algorithms were, we used a wide variety of metricswhich has four columns for metrics like true positive, false positive, true negative, and others. **Accuracy:** A common and widely-used statistic is accuracy, which measures the proportion of true or erroneous observations that are accurately anticipated. One way to find out how accurate a model is is to use the following equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \qquad (11)$$

- **Recall**:A model's "recall" is the percentage of correct predictions relative to its entire number of forecasts. Here, it stands for the proportion of true articles projected as true relative to all true articles.

Table 1 Confusion Metrics

	Predicted true	Predicted false
Actual true	True positive (TP)	False negative (FN)
Actual false	False positive (FP)	True negative (TN)

$$\text{Recall} = \frac{TP}{TP + FN}. \qquad (12)$$

- **Precision:**A high F1-score indicates a good balance between recall and accuracy. Put simply, it determines the reciprocal of the two by utilizing harmonics. The system takes into account both the positive and negative results that are considered fake. If you want to know your F1-score, you can apply this formula. positively anticipated (true) articles that are tagged as true:

$$\text{Precision} = \frac{TP}{TP + FP}. \qquad (13)$$

- **F1 Score:**As a surrogate for memory and accuracy, the F1-score works well. It finds the happy medium, to restate it another way. When making our decisions, we consider both the true and misleading positives and negatives. Here is the formula to calculate the F1-score:

$$\text{F1} - \text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \qquad (14)$$

DISCUSSION

	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.97	0.91	0.91	0.87
Linear SVM (LSVM)	0.98	0.37	0.53	0.86
Multilayer perceptron	0.98	0.35	0.94	0.9
K-nearest neighbors (KNN)	0.88	0.28	0.82	0.77
Ensemble learners				
Random forest (RF)	**0.99**	0.35	0.95	**0.91**
Voting classifier (RF, LR, KNN)	0.97	0.88	0.94	0.88
Voting classifier (LR, LSVM, CART)	0.96	0.86	0.92	0.85
Bagging classifier (decision trees)	0.98	**0.94**	0.94	0.9
Boosting classifier (AdaBoost)	0.98	0.92	0.92	0.86
Boosting classifier (XGBoost)	0.98	**0.94**	0.94	0.89
Benchmark algorithms				
Perez-LSVM	**0.99**	0.79	**0.96**	0.9
Wang-CNN	0.87	0.66	0.58	0.73
Wang-Bi-LSTM	0.86	0.52	0.57	0.62

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

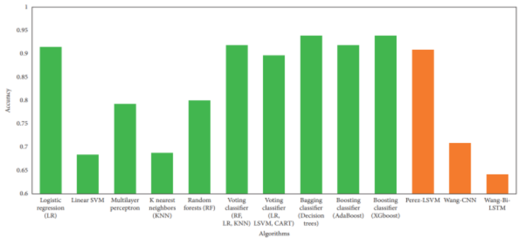


Figure2: "Average Accuracy overall Datasets"

Table 3 "Level of accuracy for the datasets"

	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.98	0.92	0.93	0.88
Linear SVM (LSVM)	0.98	0.31	0.54	0.88
Multilayer perceptron	0.97	0.32	0.93	**0.92**
K-nearest neighbors (KNN)	0.91	0.22	0.85	0.8
Ensemble learners				
Random forest (RF)	0.99	0.3	0.98	0.92
Voting classifier (RF, LR, KNN)	0.96	0.88	0.92	0.86
Voting classifier (LR, LSVM, CART)	0.94	0.86	0.88	0.83
Bagging classifier (decision trees)	0.98	**0.94**	0.93	0.9
Boosting classifier (AdaBoost)	0.98	0.92	0.92	0.86
Boosting classifier (XGBoost)	**0.99**	**0.94**	**0.96**	**0.92**
Benchmark algorithms				
Perez-LSVM	**0.99**	0.79	**0.96**	0.9
Wang-CNN	0.84	0.65	0.48	0.72
Wang-Bi-LSTM	0.92	0.43	0.5	0.65

Table 4 "The Data sets' Recall"

	DS1	DS2	DS3	DS4
Logistic regression (LR)	0.98	0.9	0.92	0.86
Linear SVM (LSVM)	0.98	0.32	**1**	0.86
Multilayer perceptron	**1**	0.36	0.96	0.88
K-nearest neighbors (KNN)	0.87	0.24	0.81	0.74
Ensemble learners				
Random forest (RF)	**1**	0.34	0.93	**0.91**
Voting classifier (RF, LR, KNN)	0.97	0.89	0.96	0.9
Voting classifier (LR, LSVM, CART)	0.97	0.87	0.96	0.89
Bagging classifier (decision trees)	0.97	**0.95**	0.94	**0.91**
Boosting classifier (AdaBoost)	0.98	0.93	0.92	0.86
Boosting classifier (XGBoost)	0.99	0.94	0.94	0.89
Benchmark algorithms				
Perez-LSVM	0.99	0.81	0.97	**0.91**
Wang-CNN	0.9	0.71	0.29	0.75
Wang-Bi-LSTM	0.78	0.59	0.35	0.61

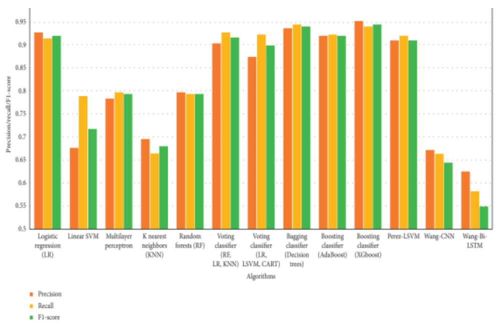


Figure 4: "All datasets for precision recall and F1 Score"

CONCLUSIONS

Manually categorizing news articles calls for expert knowledge of the subject matter and the ability to detect textual irregularities. In order to categorize false news stories, this study use ensemble methods and machine learning models. We gathered all of the data for our analysis from online news articles. In Table 5, you can see the F1-scores for all four datasets. All four of these data sets achieved F1 scores of 0.87, accuracy of 0.98, precision of 0.91, and recall of 0.92 when using logistic regression (LR). Linear Support Vector Machine (LSVM) achieved 0.98 for accuracy, 0.32 for precision, 0.7 for recall, and 0.87 for F1 score. An artificial neural network that uses numerous interconnected layers of neurons is called a multilayer perceptron. The four numbers are 0.98, 0.38, 0.34, 0.95, and 0.98. With K-nearest neighbors (KNN), the results are 0.89, 0.23, 0.83, and 0.77. Learners' grouping algorithms Measurements for the RF model included recall(0.95), accuracy (0.99), precision (0.32), and F1 score (0.91). Voting time rolled around, and the accuracy ratings for Logistic Regression (LR), K-Nearest Neighbours (KNN), and Random Forest (RF) were 0.97, 0.88, and 0.94, respectively. A total of 0.88% was the classifier's accuracy.

With a recall of 0.92, precision of 0.86, accuracy of 0.96, and an F1 score of 0.86, the voting classifier did exceptionally well. Classification and regression trees, LSVM, and R-squared were the tools used for the analysis.

The decision tree-based bagging classifier achieved a precise, accurate, recalling, and F1 score of 0.98. With an F1 score of 0.86, recall of 0.92, precision of 0.92, and accuracy of 0.98., the boosting classifier (AdaBoost) achieved all objectives. Having a recall of 0.99 and a precision of 0.95, the Boost boosting classifier achieved an accuracy of 0.9. Find out how efficient algorithms are. The numbers represent the Perez-LSVM style A, with values of 0.8, 0.96, and 0.9. "Wang-CNN A.I." With a sum of 0.87, 0.67, 0.31, and 0.73, four figures are provided. Wang Bi-Wang Long Short-Term Memory The values are 0.84, 0.44, 0.35, and 0.57, just in case you were wondering. A multi-layer A Perceptron as a support vector machine for logistic regression (LR). Machine learning methods include systems like voting classifiers and K-nearest neighbors (KNN) (RF, LR, KNN). Common machine learning models that combine the output of multiple classifiers into a single voting classifier include CART, LR, and LSVM. Bagging is a classification technique that involves using decision trees.
AdaBoost is a boosting classifier. Enhancing the performance of a classifier using the Boost algorithm. Perez-LSVM The algorithms used are Wang-CNN and Wang-BiLSTM.
Starting with 1 and working our way down, we have the following numbers: 0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, and 0.5. Accuracy, Memory, and F1-score Accuracy Memorization and F1-factor
The precision, F1-score, and recall for all datasets are displayed in Figure 3. Complication. The nine non-political pieces don't restrict themselves to political news but instead

cover a wide range of subjects. The main objective of this research is to find textual patterns that can be utilised to spot counterfeit goods. We utilised a LIWC algorithm to extract various textual fragments from the articles. When training the models, these characteristics served as inputs. The most accurate results were obtained after extensive training and tuning of machine learning models.

REFERENCES

[1] A. Douglas, "News consumption and the new electronic media," 7e International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.

[2] J. Wong, "Almost all the traffic to fake news sites is from fakebook, new data show," 2016. [3] D. M. J. Lazer, M. A. Baum, Y. Benkler et al., ")e science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[4] S. A. Garc´ıa, G. G. Garc´ıa, M. S. Prieto, A. J. M. Guerrero, and C. R. Jimenez, ")e impact of term fake news on the scientific ´ community scientific performance and mapping in web of science," Social Sciences, vol. 9, no. 5, 2020. \

[5] A. D. Holan, 2016 Lie of the Year: Fake News, Politifact, Washington, DC, USA, 2016. [6] S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/ abstract=3237763.

[7] A. Robb, "Anatomy of a fake news scandal," Rolling Stone, vol. 1301, pp. 28–33, 2017. [8] J. Soll, ")e long and brutal history of fake news," Politico Magazine, vol. 18, no. 12, 2016. [9] J. Hua and R. Shaw, "Corona virus (covid-19) "infodemic" and emerging issues through a data lens: the case of China," International Journal of Environmental Research and Public Health, vol. 17, no. 7, p. 2309, 2020.