International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 1, Jan-Feb 2024

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Peak Performance in Fake News Detection: Unveiling the Potential of Deep Learning Optimizers

Dr.M.Deepa[1], S.Keerthika[2], R.Manoj Kumar[3]
N.R.Vikram[4], P.Prakash[5]

[1] Department of Computer Science, Pavai Arts & Science College for Women, Namakkal
[2] Department of Computer Applications, Paavai College of Engineering, Namakkal
[345] Department of Computer Science & Engineerimg, Paavai College of Engineering, Namakkal
[1]legithasai2010@gmail.com, [2]keerthikamcakk@gmail.com, [3]manojkumarbmi@gmail.com,
[4]nrvikram89@gmail.com,[5]prakash.p100@gmail.com

--************************----------------------------------

Abstract:

The World Wide Web's introduction and the quick uptake of social media sites like Facebook and Twitter opened the door to a level of information sharing never seen in human history. Social media platforms are being used by consumers to create and share more information than ever before, some of it false and unrelated to reality. It is difficult to automatically classify a text article as misleading or disinformation. Even a subject-matter expert must consider a variety of factors before determining if an article is true. In this work, we propose to automatically classify news articles using an ensemble machine learning approach. Our research examines many textual characteristics that can be utilized to discern authentic content from counterfeit. We train a variety of machine learning algorithms employing those properties in conjunction with different ensemble approaches, and we assess their performance on four real-world datasets. Our suggested ensemble learner technique outperforms individual learners, as confirmed by experimental evaluation.

Keywords — Fake news, Machine Learning, Text Classification, Social Media, Ensemble Learning.

--************************----------------------------------

I. INTRODUCTION

Fake news is raging through the various social media platforms in large quantities. In this instance, classifying news, posts, stories, and journals as genuine or fraudulent has become essential to distinguishing between the two, and it has also piqued the curiosity of scholars worldwide. As per multiple analysis studies conducted to find out how fake news details affect us once we return, both fake and fictitious news are reported. When someone is exposed to fake news, their basic thought process may be triggered by something that may not be factual.

The pandemic that is sweeping the globe is the best illustration of fake news. The World Wide Web's introduction and the quick uptake of social media sites like Facebook and Twitter opened the door to a level of information sharing never seen in human history. In addition to other applications, news organizations profited from the extensive usage of social media platforms by offering their customers up-to-date news almost instantly. Newspapers, tabloids, and magazines gave way to online news platforms, blogs, social media feeds, and other digital media formats as the news industry changed [1]. Customers can now more easily obtain the most recent news at their fingertips. Seventy percent of traffic to news websites comes from Facebook recommendations [2]. Due to their current capacity to facilitate user discussion and idea sharing on topics like democracy, education, and health, social media platforms are incredibly potent and helpful.

Over the past ten years, there has been a noticeable surge in the dissemination of false information, as evidenced by the US elections of 2016 [4]. Numerous issues have arisen as a result of the widespread dissemination of internet publications that misrepresent the facts, not only in the political sphere but also in the fields of sports, health, and science [3]. The financial markets are one such arena that is impacted by false news [5], where a rumor can have severe effects and even cause the market to stop.

Data is available on the World Wide Web in a variety of formats, including papers, audio files, and movies. It might be challenging to identify and categorize news that has been published online in an unstructured manner (such as news, articles, videos, and audios) because it absolutely requires human skill. Nevertheless, anomalies that distinguish a text article that is dishonest from one that is fact-based can be found using computer methods like natural language processing (NLP). [6]

Other methods examine how fake news spreads in comparison to authentic news [7]. To be more precise, the methodology examines the ways in which a false news piece spreads on a network in contrast to an authentic article. Theoretically, it is possible to distinguish between genuine and fraudulent responses to an article. In order to determine whether or not an article is deceptive, a more hybrid technique can also be utilized to investigate the textual elements and study the article's social reaction.

II. RELATED WORKS

In Ahmed et al.'s study [8], linguistic features like n-grams were extracted from textual articles, and multiple machine learning (ML) models were trained, including K-nearest neighbor (KNN), logistic regression (LR), support vector machine (SVM), linear support vector machine (LSVM), decision tree (DT), and stochastic gradient descent (SGD). SVM and logistic regression achieved the highest accuracy (92%) out of all the models. The study found that overall accuracy dropped as the number of grams computed for a given piece grew. Learning models that are applied to classification tasks have been shown to exhibit this phenomenon.

Shu et al. [9] used textual features with auxiliary data, such as user social engagements on social media, to improve accuracy with various models. The writers also covered the sociological and psychological theories and how to apply them to identify misleading content on the internet. The writers also covered several data mining strategies for building models and common features extraction methods. These models are grounded on social context, including stance and propagation, as well as knowledge, including writing style.

Wang employs a different strategy [10]. The author trained a variety of machine learning models using textual features and metadata. The author primarily addressed the application of convolutional neural networks (CNNs). The bidirectional LSTM layer is employed after a convolutional layer to capture the dependency between the metadata vectors. In order to create the final prediction, the max-pooled text representations were combined with the bidirectional LSTM's metadata representation. This combination was then input into a fully connected layer that used a softmax activation function. The study is carried out using a political dataset that includes quotes from two opposing political parties. In addition, a feature set of metadata is included that includes subject, speaker, employment, state, party, context, and history.

Riedel et al. [11] offer a competing approach in the form of a stance identification algorithm that classifies articles as either "agree," "disagree," "discuss," or "unrelated" based on how closely the article title and article text match. The authors employed a multilayer perceptron (MLP) classifier with one hidden layer and a softmax function on the output of the final layer. They did this by using linguistic aspects of text such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF) as a feature set. Articles with a headline, body, and label were included in the dataset. In test cases, the system works best when it comes to the "agree" label, while its accuracy on the "disagree" label was subpar.

Vosoughi et al. [12] took a novel approach to investigating the characteristics of news dissemination on social media. Specifically, they examined the propagation of news (rumours) on Twitter and examined the ways in which fake news varies from genuine news in terms of its dissemination on the platform. The paper discusses a number of analysis techniques to investigate the online spread of fake news, including the number of unique Twitter users reached at any depth, the number of minutes it takes for true and false rumour cascades to reach depth and number of Twitter users, the size, maximum breadth, structural virility, and mean breadth of true and false rumour cascades at various depths.

III. METHODOLOGY

Both supervised and unsupervised learning techniques have been applied to text classification in the present fake news corpus on several occasions [13]. But the majority of the research focuses on particular datasets or domains, most notably the political sector. In this research, we suggest a machine learning ensemble strategy as a solution to the fake news identification problem. Our research investigates many textual characteristics that may be utilized to differentiate authentic information from counterfeit. We leverage such qualities to train a variety of ensemble methods—some of which are not well-examined in the existing literature—using a combination of different machine learning algorithms. Due to the learning models' propensity to lower error rates through the use of strategies like bagging and boosting, ensemble learners have shown to be beneficial in a wide range of applications [14]. These methods enable the efficient and successful training of various machine learning algorithms. We also carried out in-depth tests on four real-world datasets that are accessible to the public. The outcomes demonstrate that our suggested technique performs better utilizing the four widely-used performance criteria (accuracy, precision, recall, and F-1 score).

We assessed the efficacy of false news detection classifiers using the following learning algorithms in combination with our suggested methodology.

A. The Logistic Regression

A logistic regression (LR) model is employed because it offers a straightforward equation for classifying problems into binary or multiple classes. This is because we are categorizing text based on a vast feature set, with a binary outcome (true/false or true article/fake article) [15]. Before obtaining the maximum accuracies from the LR model, we tried a number of parameters and then did hyper parameter tuning to acquire the best outcome for each particular dataset.

B. Support Vector Machine

Another model for binary classification problems is the support vector machine (SVM), which comes with a number of kernel functions. To classify data points, an SVM model's goal is to estimate a hyper plane, also known as a decision border, based on the feature set [16]. The number of features affects the hyper plane's dimension. Since there are several ways for a hyper plane to exist in an N-dimensional space, the goal is to find the plane that has the greatest margin of separation between the data points of two classes.

C. Multi Layer Perceptron

An artificial neural network comprising an input layer, one or more hidden layers, and an output layer is called a multilayer perceptron (MLP). MLP can be as basic as having all three layers, but in our tests, we have optimized the model by adjusting its parameters and layer count to produce the best possible prediction.

D. Ensemble Learners

Our proposal was to enhance the overall accuracy of classifying an article as true or false by utilizing linguistic qualities as feature input in conjunction with existing ensemble techniques. Because several models are trained using a specific technique to lower the total error rate and increase the model's performance, ensemble learners typically have higher accuracies. The idea underlying ensemble modelling is the same as the one we are accustomed to applying in our daily lives, such as consulting with several experts before making a decision to reduce the likelihood of making a mistake or experiencing an unfavourable result.

E. Random Forest

A more sophisticated version of decision trees (DT), a supervised learning model, is called random forest (RF). RF uses a large number of decision trees operating independently to forecast a class's result; the class with the most votes determines the final prediction. Because there is less association between the trees in a random forest than in other models, the error rate is lower [17]. In order to find the best model that can accurately predict the outcome, a grid search was conducted using varying numbers of estimators to train our random forest model. Depending on whether a decision tree split is needed for regression or classification, there are a number of different algorithms available.

F. Bagging Ensemble Classifier

An early ensemble technique called bootstrap aggregating, also known as bagging classifier, is primarily used to lower the variance (overfitting) over a training set. One of the most popular variations of the bagging classifier is the random forest model. In order to minimize overall variance in a classification issue, the bagging model intuitively chooses the class based on significant votes indicated by the number of trees, and randomly selects each tree's data using replacement sampling from the entire dataset. However, the bagging model averages across numerous estimates for regression issues.

IV. RESULT & DISCUSSION

It is clear that the random forest algorithm and Perez-LSVM together provide a maximum accuracy of 99% on the DS1 (ISOT Fake News Dataset). 98% accuracy was attained with bagging classifiers, boosting classifiers, multilayer perceptrons, and linear SVM. Ensemble learners achieve an average accuracy of 97.67% on DS1, whereas individual learners get an average accuracy of 95.25%. The absolute difference of 2.42% between learners in an ensemble and those learning alone is not statistically significant. The Wang-CNN and Wang-Bi-LSTM benchmark algorithms outperformed the other algorithms in terms of performance. The two best-performing algorithms on DS2, bagging classifier (decision trees) and boosting classifier (XGBoost), both achieve 94% accuracy.

It's interesting to note that Perez-LSVM, random forest, and linear SVM all did poorly on DS2. While the accuracy of ensemble learners is 81.5%, that of individual learners was reported to be 47.75%. For DS3, a similar pattern is seen, with individual learners' accuracy being 80% and ensemble learners' accuracy being 93.5%. On the other hand, Perez-LSVM, which attained an accuracy of 96%, is the top performing algorithm on DS3, in contrast to DS2. The best algorithm on DS4 (DS1, DS2, and DS3 combined) is random forest, with an accuracy rate of 91%. Individual students received an accuracy of 85% on average, while ensemble students received an accuracy of 88.16%. With an accuracy of 62%, Wang-Bi-LSTM is the least effective algorithm.

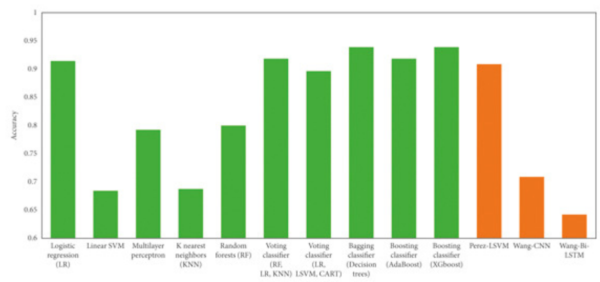

Figure 1:Comparision of various ensemblers

The average accuracy of each algorithm across the four datasets is shown in Figure 1. Overall, Wang-Bi-LSTM (accuracy 64.25%) is the worst performing method, whereas bagging classifier (decision trees) performs best (accuracy 94%). The accuracy of individual learners is 77.6%, whereas that of ensemble learners is 92.25%. Except for DS2, all datasets saw higher accuracy rates when using random forests. Recall, precision, and F1-score are additional metrics we use to assess the performance of learning models as accuracy score alone is not a suitable way to gauge a model's effectiveness.

V. CONCLUSION

In-depth domain knowledge and proficiency in spotting textual irregularities are necessary for the manual classification of news. In this study, machine learning models and ensemble techniques were used to address the problem of classifying bogus news stories. The World Wide Web provides us with a wealth of news stories from other domains, covering the majority of news instead of just categorizing political news, which is why we employed this data for our work. Finding textual patterns that set bogus articles apart from real news is the main goal of the study. Researchers need to focus on the numerous unresolved difficulties surrounding fake news identification. For example, recognizing the critical components involved in the dissemination of news is a crucial first step toward reducing the spread of fake news. Graph theory and machine learning methods can be used to pinpoint the main distributors of false information. Similarly, identifying bogus news in videos in real time may be another potential future path.

REFERENCES

[1] A. Douglas, "News consumption and the new electronic media," *The 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 facebook, new data show," 2016.
[3] D. M. J. Lazer, M. A. Baum, Y. Benkler et al., "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.
[4] A. D. Holan, *2016 Lie of the Year: Fake News*, Politifact, Washington, DC, USA, 2016.
[5] S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/abstract=3237763.
[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
[7] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
[8] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pp. 127–138, Springer, Vancouver, Canada, 2017.
[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
[10] W. Y. Wang, *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.
[11] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," 2017, https://arxiv.org/abs/1707.03264.
[12] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
[13] N. Ruchansky, S. Seo, and Y. Liu, "Csi: a hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, Singapore, 2017.
[14] P. Bühlmann, "Bagging, boosting and ensemble methods," in *Handbook of Computational Statistics*, pp. 985–1022, Springer, Berlin, Germany, 2012.
[15] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, UK, 2000.
[16] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017.