JOURNAL OF DATA SCIENCE | Vol.2024:08
eISSN:2805-5160

Machine Learning for Fake News Detection Analysis

Abhilasha S[1*], Ushasree R[1*], Che Fuzlina Mohd Fuad[2]

[1]Dayananda Sagar Academy of Technology and Management, Karnataka, India
[2]Faculty of Data Science and Information Technology, INTI International University, Malaysia

[*]**Email:** ashuabhilashas65@gmail.com, Ushasree19832000@gmail.com

Abstract

The COVID-19 outbreak has required some health and financial decisions to be made in an unwieldy manner. This has spread uncertainty and lies all over the world. The transmission of false information has been compounded by the problems with fake news. Many of them gave up on newspapers, magazines, and other print media in favor of Internet pleasure. Online entertainment has become the primary news source for a sizable percentage of the population due to its ease of access, low cost, and rapid spread. In some circumstances, bogus information spreads faster than true information to gain popularity over internet entertainment and divert people from the underlying issues. People spread false information using online entertainment for commercial and political benefit. To avoid a harmful influence on society, it is critical to immediately recognize bogus information in all systems. To demonstrate the efficiency of the grouping on the dataset, we produced and tested numerous AI computations independently for this assignment, which looks into research on the recognition of fake news. The Jupyter Notebook stage of this project was used, and the execution was assessed.

Keywords

Fake News, Social Media, Machine Learning, and Classifiers

Introduction

These days, even rural cities may access virtual entertainment, which has long been a part of our lives. While online entertainment has made social contact simpler, there has been a noticeable rise in the amount of persons uploading and disseminating misleading material over the past several years. Ninety percent of people rely on virtual entertainment to get their news because the internet is so accessible and modern technology makes this possible.

Google and Facebook are always working to resolve these issues. One way to recognise fake news is to label it as such and use tools like reality-checking markings, deception websites, and so on (Kumar, K. A., Preethi, & Vasanth, 2020). These tactics have not yet been put to the intended use, despite the fact that it is difficult to determine the authenticity of fake news due to its rapid dissemination and blurry border (Zhou, X., & Zafarani, R., 2020). People ought to

Submission: 12 May 2024; **Acceptance:** 29 May 2024

know what to accept and reject as a result. Fake news detection is becoming more and more important (Sharma, U., Saran, S., & Patil, S. M., 2020).

Methodology

Customers can use the method developed by earlier researchers (Monther Aldwairi et al., 2024) to identify and prioritize locations that include inaccurate and deceptive information. Customers are taken to a website page with content that is significantly biased against their expectations when they click on an association (Aphiwongsophon, S., & Chongstitvatana, P., 2018). We refer to this as a deceptive content source. The customer loses time and is agitated as a result. The strategy includes using a tool to identify and filter out fake complaints from the results that a web crawler or virtual news station offers to a customer. These mechanical assemblies are available for instant download and display in the client's structure.

Dataset

The first and most important step in this endeavor was identifying the dataset that may be used to achieve the goal. News data can be gathered by professional journalists, fact-checking sites, industry markers, and crowdsourced laborers. The new datasets Kaggle provides for our work could change at any time. The exact kind of data for different assessment papers was chosen using these datasets. Three different datasets have been used. They're not numerous enough, they can be monotonous or conflicting, and they nearly surely include a variety of errors in them. Thus, we made a few awkward changes to wrap up the gathering (Kumar, V., et al., 2021).

Once the dataset has been input, data pre-taking is further improved through the depiction computations (Agarwal et al., 2019). Additionally, pre-processing was done on the dataset, including word embedding, TF-IDF vectorization, stopword removal, and the creation of count vectors. All text is converted into a numerical game plan using the TFIDF vectorization, which facilitates the fitting of AI estimations. The input dataset is complete; all properties are there, and it will undergo tokenization. After it has been handled once more, unwanted data will be eliminated from the tokenized dataset. The process of choosing the agreeable fundamental component subsets from the open subsets to generate a final specified subset is frequently referred to as feature assurance.

With this method, the essential components are moved into a new area with fewer perspectives (Samantaray, S. D., & Jodhani, G., 2019). Otherwise, no new components are developed; only a select few features are selected, and the meaningless and boring parts are subsequently eliminated. We collected and managed the data using Genism, Numpy, the Sklearn package, and Pandas (Vogel, I., & Meghana, M., 2020). Wordcloud, Seaborn, and Matplotlib were used to interpret the data. We divided our dataset into preparation and testing datasets in order to finish the assessment cycle.

All of the news content is included in our three datasets. Dataset-1 has 44898 pieces of information with the classification of "valid" or "phoney." 6310 pieces of information in Dataset-2 are marked as "Phoney" or "Genuine," respectively. 4049 records in Dataset-3 have a mark class value of "1" or "0," respectively. Consider '1' for fake and '0' in value without a doubt. The word cloud is used to visualize the news articles in the dataset.

Framework Overview

The fragmented nature of our proposed system is seen in Figure 1. Whether the news is true or false in this case depends on your point of view. Before we can start the interaction, we need to have a clear understanding of the issue, select a model representation, and assess the outcome. The ubiquity of fake news has gained attention since the US presidential election in 2016. People start paying close attention to the topic of fake news after that. There is a discernible rise in the quantity of unique fake news articles discussed and shared on the internet throughout a political campaign. It is even suggested in some blogs that fake news played a role in Trump's election as president.

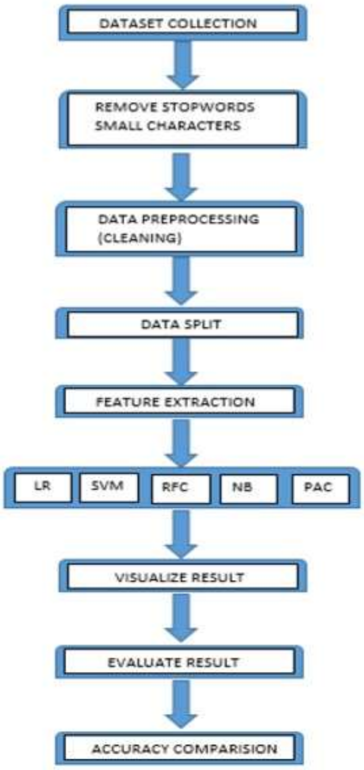

Figure 1. Outline of work

Classifiers

We employed the Jupyter Notebook platform and the Python programming language for the execution of five different types of AI computations in our model. The gullible Bayesian Model, logistic regression, support vector machine, random forest classifier, and inactive forceful classifier were all implemented as our arrangement models using the aforementioned dataset (Aldwairi, M., & Alwahedi, A., 2018). These computations contain features and execution that handle several datasets, and they perform well for different groups. For inspection and organization purposes, we have used four features: id or URL, title or title, text or body, and mark or target. Important details like the date and subject are omitted. A word cloud represents the continuous terms used in the news material (Joju, S., & Kammath, P. S., 2021). Both Figures 2 and 3 explain this matter.

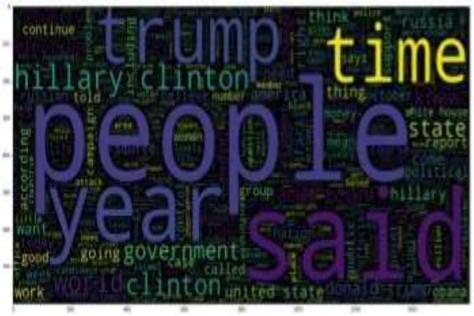

Figure 2. Word cloud for counterfeit news Dataset 2

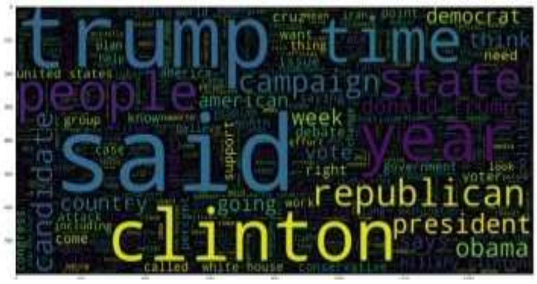

Figure 3. Word cloud for genuine news Dataset 2

Logistic Regression's result is depicted in Figure 4.
"Logistic relapse" is a factual research approach that predicts an information value that validates past perceptions of an information collection. According to Manzoor and Singla (2019), the approach facilitates the calculation of approaching information by an AI application that is backed by verifiable information.

With more meaningful data coming in, the algorithm should get better at predicting groupings inside informative indexes. Strategic relapse can also be used in information readiness exercises by allowing informational indexes to be placed into specially designated containers throughout the concentrate, change, load (ETL) process to organize the material for analysis. A strategic relapse model anticipates the dependent information variable by examining the relationship between the variable and at least one pre-existing free element.

	F1-SCORE	RECALL	PRECISION
DATASET -1	FAKE=98.77 ; REAL=88.66	FAKE=98.97 ; REAL=98.44	FAKE=98.57 ; REAL=98.88
DATASET -2	FAKE=90.12 ; REAL=89.72	FAKE=86.83 ; REAL=93.41	FAKE=93.67 ; REAL=86.32
DATASET -3	FAKE=96.96 ; REAL=97.42	FAKE=95.72 ; REAL=98.51	FAKE=98.24 ; REAL=96.35

Figure 4. Performance evaluation LR

The support Vector Machine's result is depicted in Figure 5.

A support vector machine is a supervised learning method that splits data into two categories. After being trained using a set of knowledge that has previously been split into two groups, it constructs the model. Using an SVM technique, ascertain which category a replacement datum falls under. Consequently, SVM performs the role of a non-binary linear classifier. Support vector machines (SVMs) are machine learning algorithms used to assess data for multivariate analysis and categorization. SVMs are employed in text categorization, image classic classification, and handwriting identification in the sciences.

	F1-SCORE	RECALL	PRECISION
DATASET -1	FAKE=99.48 ; REAL=99.43	FAKE=99.51 ; REAL=99.4	FAKE= 99.46 ; REAL=99.46
DATASET -2	FAKE=92.79 ; REAL=92.76	FAKE=90.84 ; REAL=94.81	FAKE=94.83 ; REAL=90.8
DATASET -3	FAKE=98.69 ; REAL=98.9	FAKE=98.26 ; REAL=99.26	FAKE=99.12 ; REAL=98.54

Figure 5. Performance evaluation SVM

Random Forest's results is depicted in Figure 6.

Random forests are yet another method for supervised learning. Two typical applications for it are regression and classification. It is also the most flexible and straightforward approach. A forest has trees in it. The more trees in a forest, the stronger it is believed to be. Random forests create decision trees based on randomly selected data samples, get forecasts from each tree, and then cast votes to identify the single viable answer. Additionally, it provides a fairly accurate indicator of the feature's importance. Random forests are used in recommendation engines, picture categorization, and selection. It is often used to identify fraud, predict diseases, and classify trustworthy loan applications.

	F1-SCORE	RECALL	PRECISION
DATASET -1	FAKE=98.92 ; REAL=98.81	FAKE=98.89 ; REAL=98.84	FAKE=98.94 ; REAL=98.78
DATASET -2	FAKE=89.46 ; REAL=89.87	FAKE=89.52 ; REAL=89.81	FAKE=89.41 ; REAL=89.93
DATASET -3	FAKE=96.48 ; REAL=96.91	FAKE=93.78 ; REAL=99.42	FAKE=99.34 ; REAL=94.53

Figure 6. Performance evaluation RFC

Passive Aggressive Classifiers are depicted in Figure 7

Beginners and even intermediate Machine Learning enthusiasts may not be familiar with the passive-aggressive algorithms class of machine learning algorithms. Still, they will surely be very useful and successful applications. For the most part, passive-aggressive algorithms are comparable to Perceptron models in that they don't require a learning rate. They do make use of a regularisation parameter, though. When classifications are accurate, it responds passively; when errors occur, it responds forcefully.

	F1-SCORE	RECALL	PRECISION
DATASET-1	FAKE=99.56 ; REAL=99.51	FAKE=99.57 ; REAL=99.5	FAKE=99.54 ; REAL=99.53
DATASET-2	FAKE=92.55 ; REAL=92.74	FAKE=91.96 ; REAL=93.32	FAKE=93.15 ; REAL=92.16
DATASET-3	FAKE=98.9 ; REAL=99.09	FAKE=99.33 ; REAL=98.73	FAKE=98.46 ; REAL=99.45

Figure 7. Performance evaluation PAC

Naïve Bayes Naive result is depicted in Figure 8.

Bayes is one example of a probability-based classification method; it uses the probabilities of all possible traits that have already been observed to predict class membership. This approach is used when a collection of disparate attributes, or evidence, affects the way the target class is chosen. Nave Bayes is capable of considering attributes that, when considered alone, are meaningless but, when considered collectively, significantly influence the probability that an instance belongs to a particular class. Since it is assumed that all features have equal significance, there is no correlation between the value of the first feature and the values of the other features. Put another way, the characteristics stand alone.

	F1-SCORE	RECALL	PRECISION
DATASET-1	FAKE=94.41 ; REAL=93.77	FAKE=93.77 ; REAL=94.5	FAKE=95.06 ; REAL=93.06
DATASET-2	FAKE=79.73 ; REAL=85.39	FAKE=96.17 ; REAL=76.02	FAKE=68.09 ; REAL=97.39
DATASET-3	FAKE=93.93 ; REAL=94.66	FAKE=91.12 ; REAL=97.3	FAKE=96.92 ; REAL=92.15

Figure 8. Performance evaluation NB

Result and Discussion

Fake news is becoming more and more common on social media as more people use them. False information typically leaves individuals confused and distracted. Research is being done to resolve this problem. As our contribution, we commit this work to the detection of fake news.

This study made use of three datasets that were collected from open sources. We also evaluated the performance of five algorithms in addition to the well-known machine. The comparative analysis was carried out using the accuracy value. Accuracy played a major role in selecting the best model. Originally, the dataset was split into training and testing subsets. In Figure 9, the accuracy is displayed.

	LR	SVM	RFC	PAC	NB
DATASET-1	98.72	99.46	98.87	99.54	94.11
DATASET-2	89.92	92.78	89.67	92.65	83.09
DATASET-3	97.21	98.8	96.71	99.0	94.32

Figure 9. Performance evaluation of the accuracy

The accuracy scores of the three datasets' classification methods are displayed in the following graphs. The study discovered that, in terms of accuracy, the PAC approach performed better than the other two datasets (Dataset-1 and Dataset-3). SVM has the maximum accuracy in Dataset 2, one of the three datasets. Among the four classifiers, the Naive Bayes model performs the worst across all datasets. Figures 10, 11, and 12 demonstrate each dataset's correctness in depth.

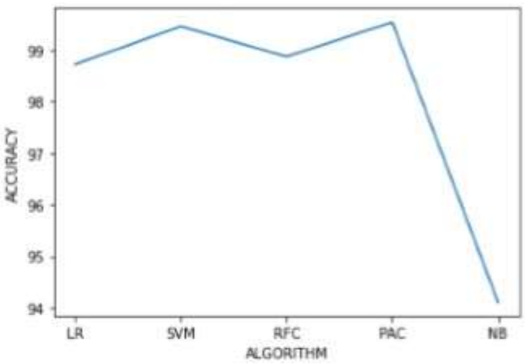

Figure 10. Precision Examination on Dataset 1

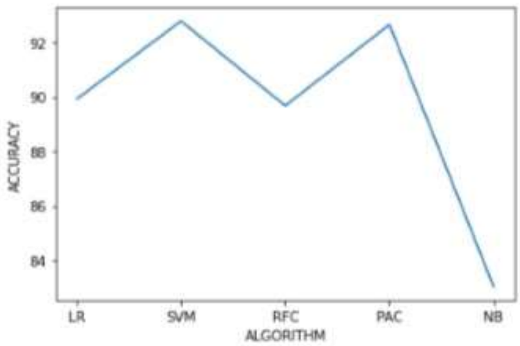

Figure 11. Precision examination on Dataset 2

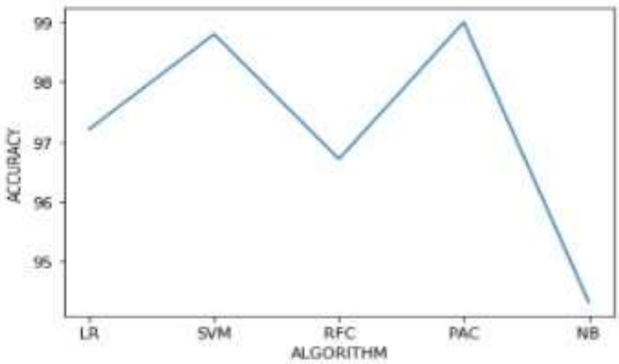

Figure 12. Precision examination on Dataset-3

Conclusion

This work presents an analysis and investigation of five computations devoted to the location of bogus news. Based on our initial study, we have concluded that the Naive Bayes classifier has performed the worst out of the four models. This investigation also showed how well each model identified false information. As was previously mentioned, virtual entertainment is becoming more and more common. Other experts might use this inquiry as a framework to assess if models are successfully achieving their main goals of detecting fake news.

Acknowledgments

The Dayananda Sagar Academy of Technology and Management (DSATM) is truly appreciated by the authors for giving us the tools and space we needed to carry out this research study on "Fake News Detection Analysis Using Machine Learning." The institution's encouragement and assistance have been crucial to this endeavor's successful conclusion.

References

Kumar, K. A., Preethi, G., & Vasanth, K. (2020). A study of fake news detection using machine learning algorithms. *Int. J. Technol. Eng. Syst.(IJTES)*, *11*(1), 1-7.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.

Sharma, U., Saran, S., & Patil, S. M. (2020). Fake news detection using machine learning algorithms. *International Journal of creative research thoughts (IJCRT)*, *8*(6), 509-518.

Khalil, A., Jarrah, M., & Aldwairi, M. (2024). Hybrid Neural Network Models for Detecting Fake News Articles. *Human-Centric Intelligent Systems*, *4*(1), 136-146.

Aphiwongsophon, S., & Chongstitvatana, P. (2018, July). Detecting fake news with machine learning method. In *2018 15th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)* (pp. 528-531). IEEE.

Kumar, V., Kumar, A., Singh, A. K., & Pachauri, A. (2021, November). Fake news detection using machine learning and natural language processing. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)* (pp. 547-552). IEEE.

Agarwal, V., Sultana, H. P., Malhotra, S., & Sarkar, A. (2019). Analysis of classifiers for fake news detection. *Procedia Computer Science*, *165*, 377-383.

Samantaray, S. D., & Jodhani, G. (2019). Fake news detection using text similarity approach. *International Journal of Science and Research (IJSR)*, *8*(1), 1126-1132.

Vogel, I., & Meghana, M. (2020, October). Detecting fake news spreaders on twitter from a multilingual perspective. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 599-606). IEEE.

Aldwairi, M., & Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, *141*, 215-222.

Joju, S., & Kammath, P. S. (2021). Analysis on fake news detection using machine learning. *International Research Journal Of Engineering And Technology (Irjet)*, *8*(7), 181-187.

Manzoor, S. I., & Singla, J. (2019, April). Fake news detection using machine learning approaches: A systematic review. In *2019 3rd international conference on trends in electronics and informatics (ICOEI)* (pp. 230-234). IEEE.