International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**11**, pp.**13-16**, November **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Fake News Detection Using Machine Learning Algorithm Logistic Regression

K. Ramya[1*], M. Yamini[2], K. Prajwala[3], M. Jyothirmai[4]

[1,2,3,4]Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India

[*]*Corresponding Author: kancharlaramya06@gmail.com*

Received: 30/Sept/2023; **Accepted**: 02/Nov/2023; **Published**: 30/Nov/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i11.1316

Abstract: Machine learning is field of Artificial Intelligence that focuses on the development of algorithms and statistical methods. Fake news has caused a lot of issues for our society. Many researchers are trying to determine what fake news is. It is challenging to recognize ambiguous fake news, can only be found after determining meaning and recent pertinent facts. For news, everyone uses a variety of online sources. News quickly disseminated among millions of users in a very short period of time to the increase in the use of social media platforms like Facebook, Twitter, etc. We will enable the user to categorize news as either genuine or real. The logistic regression approach will be used to identify false news. Natural Language processing techniques like Term Frequency Inverse Document Frequency (TF-IDF), text processing etc. In our experiment, we'll demonstrate how our method boosts bogus news' overall performance. We are providing URL search whether the given URL is fake or not.

Keywords: Fake news, Natural Language processing, Logistic Regression, Machine Learning, Term Frequency Inverse Document Frequency, Text Processing

1. Introduction

Fake news has become a major social issue as a result of the growth of social media and digital media. Sensationalism, click-bait headlines, and the fabrication of facts are typical traits of fake news. To stop the spread of false information, fact-checking, media literacy, and critical thinking are crucial strategies. The issues posed by false news in the digital era are being recognised more and more by governments, tech corporations, and individuals.

It can be created intentionally to deceive, misinform, or manipulate readers. The environment for information distribution has changed dramatically as a result of the World Wide Web's phenomenal growth and the quick adoption of social media platforms like Facebook and Twitter. As a result of this change, news organisations can now send timely updates to subscribers in ways that have never been seen before in human history. As news media shifts from traditional formats like newspapers and magazines to a digital space embracing online news platforms, blogs, social media feeds, and various other digital formats, this change is clearly visible.

The COVID-19 pandemic has seen an increase in the dissemination of false information, making it more difficult to control a worldwide health emergency. On numerous online sites, false information about the virus's origins, prevention, and treatment has spread quickly.

Such false information has a number of negative effects, such as encouraging the use of hazardous or ineffective therapies, raising public anxiety levels, and weakening public confidence in reliable health sources. Governments, health organisations, and tech firms have stepped up their efforts to stop the spread of false information, highlighting the need of accurate information and fostering media literacy to distinguish reliable sources from false ones.

Social media platforms in their current form are very effective and helpful for enabling users to debate, share, and discuss topics like democracy, education, and health. However, some organisations also use these platforms negatively, frequently to obtain financial advantage, and occasionally to sway public opinion, influence people's attitudes, or propagate satire or ridiculousness

1.1 Characteristics of Fake News

They frequently make grammatical mistakes. They often have emotive coloring. They typically try to influence readers' attitudes on a wide range of topics. They regularly use catchy terms and news. They're too good to be true. Frequently, their sources are not trustworthy.

1.2 Our Contributions

Our research investigates many textual characteristics that may be utilized to discern authentic content from counterfeit. We utilized the logistic regression machine learning approach to accomplish this.

Future directions for fake news identification are provided, and we address a number of outstanding difficulties.

2. Related Work

One important use of machine learning and natural language processing (NLP) is the detection of fake news. Diverse machine learning methodologies have been employed to tackle this issue. I'll give a summary of the many machine learning methods that are frequently employed to identify false news here. SVMs may divide data into distinct classes based on a hyperplane and are useful for binary classification. To determine the ideal decision border between phony and authentic news pieces, SVMs have been used to the fake news detection domain. Based on Bayes' theorem, naive Bayes classifiers are very helpful for text classification applications. By simulating the conditional likelihood of words given a class (fake or real), they can be used for the detection of fake news. Multiple decision trees are combined into a single forecast using an ensemble learning technique called a random forest. Because they can identify intricate patterns in the data, they are useful in the detection of bogus news. In challenges involving the detection of fake news, gradient boosting algorithms like Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) can achieve high classification accuracy. To fix the mistakes in the earlier models, they progressively construct decision trees. Fake news identification has showed great potential for deep learning techniques, such as transformer-based models (e.g., BERT), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). These models are capable of comprehending semantics and context and are able to capture complex relationships in text. Recurrent neural networks of the LSTM type are capable of modeling text sequences well. They have been used to detect fake news by capturing temporal dependencies in news stories. One popular linear model for binary classification applications is logistic regression. By learning to categorize articles as "real" or "fake" based on attributes taken from the text, metadata, or both, it can be used for the identification of fake news. We know that logistic regression is already existed. But in existing they provided accuracy which is not much as better than us. So we are providing better accuracy, url search in proposed fake news detection by using logistic regression.

3. Theory/Calculation

The chance of a binary event occurring or a binary classification outcome is modeled using the sigmoid function. It converts an input feature set with a linear combination of weights into a probability score between 0 and 1.
Logistic Function (Sigmoid Function):

$$P(Y=1/X)=1/(1+E^{(-Z)}) \quad \textbf{- eq(1)}$$

Term frequency-inverse document frequency, or TF-IDF, is a numerical statistic used in information retrieval and natural language processing to assess a term's (word or phrase's) relevance inside a document in relation to a corpus of documents. In the context of a broader collection of papers, it aids in measuring the relevance of a term to a particular document. Tasks including document retrieval, text categorization, and text mining frequently make use of TF-IDF.
Term Frequency (TF)=(1+log(tf))
Inverse Document Frequency (IDF)=log(N/df(t))
TF-IDF= (Term Frequency) *(Inverse Document Frequency) **- eq(2)**

4. Experimental Method/Procedure/Design

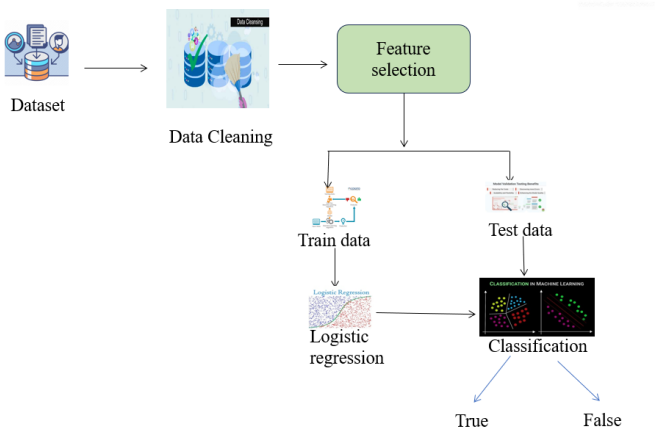

Figure 1: System Architecture

We took a dataset with news articles or text data that has been classified as "real" or "fake" in order to develop a fake news detection system. Preparing the data is crucial to guaranteeing the dataset's quality. Tasks like text normalization fall under this category. Tokenization is the process of dividing a text into discrete words or units. Removal of stopwords: Taking out frequently used terms that don't offer much information. managing outliers or missing data. Classification depends on the significant features that can be extracted from the text data. Common methods include of Term Frequency-Inverse Document Frequency gives words weights according to how important they are in the document and the entire dataset. For the purpose of developing and assessing the model, the dataset is divided into training and testing sets. A typical division is 20% for testing and 80% for training. For binary classification issues such as the identification of fake news, logistic regression is frequently used. The task of classifying news articles as "real" or "fake" is commonly known as binary classification in the context of fake news detection. The proportion of correctly identified occurrences to the total is known as accuracy. For example, https://www.abc.net.au/news/2020-04-04/coronavirus-covid-19-face-masks-paul-kelly-australians/12122042-is the URL of the webpage you wish to categorize. A Python method named ‘check_news_url’, which accepts a URL as input, is defined in the code. The requests are used by the function. Use the get (url) function to transmit the supplied URL as part of an HTTP GET request. It determines whether the response status code is 200, which denotes an

HTTP OK response—a successful response. The web page content is preprocessed if the response status code is 200.It uses preprocessing techniques that are comparable to those used on the training set of data, such as: deleting characters that are not alphabetic, the text is being changed to lowercase, putting words into the text, utilizing the Natural language tool kit Porter stemmer to stem the words, deleting stopwords from English to get rid of common. The classification result is then printed, stating whether the information at the provided URL is categorized as "real news" or "fake news" according to the model's forecast.

5. Results and Discussion

The "fake.csv" dataset was loaded and preprocessed, integrating the "AUTHOR" and "TITLE" columns into a single "content" column and managing missing values.
To get text data ready for modeling, it was tokenized, stemmed, lowercased, and stopword-free.

TF-IDF vectors, a popular method for transforming text data into a numerical representation appropriate for machine learning, were created from the text data.
Using the preprocessed text data, a logistic regression model was trained.
The model's fit to the training set of data is measured by a formula known as training accuracy. It can be a sign of possible overfitting.

To assess how well the model generalized to previously unseen data, testing accuracy was also computed. It gives an indication of the model's potential performance in real life.
A function to categorize news article authenticity according to its URL is included in the code.

Using the trained model, the function retrieves online information, preprocesses it, and determines if it contains bogus or authentic news.
The code walks through the essential procedures for creating a rudimentary false news detecting system.

While the accuracies of testing and training are crucial markers of model performance, we got accuracy of 98% by using logistic regression, additional metrics and reporting could offer a more thorough analysis.

	id	title	author	text	label	content
0	0	House Dem Aide: We Didnâ□□t Even See Comeyâ□□s...	Darrell Lucus	House Dem Aide: We Didnâ□□t Even See Comeyâ□□s...	1	Darrell Lucus House Dem Aide: We Didnâ□□t Even...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	Consortiumnews.com Why the Truth Might Get You...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	Jessica Purkiss 15 Civilians Killed In Single ...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	Howard Portnoy Iranian woman jailed for fictio...
...	...	...	...	...	...	...
20795	20795	Rapper T.I.: Trump a â□□Poster Child For White...	Jerome Hudson	Rapper T. I. unloaded on black celebrities who...	0	Jerome Hudson Rapper T.I.: Trump a â□□Poster C...
20796	20796	N.F.L. Playoffs: Schedule, Matchups and Odds -...	Benjamin Hoffman	When the Green Bay Packers lost to the Washing...	0	Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797	20797	Macyâ□□s Is Said to Receive Takeover Approach ...	Michael J. de la Merced and Rachel Abrams	The Macyâ□□s of today grew from the union of s...	0	Michael J. de la Merced and Rachel Abrams Macy...
20798	20798	NATO, Russia To Hold Parallel Exercises In Bal...	Alex Ansary	NATO, Russia To Hold Parallel Exercises In Bal...	1	Alex Ansary NATO, Russia To Hold Parallel Exer...
20799	20799	What Keeps the F-35 Alive	David Swanson	David Swanson is an author, activist, journa...	1	David Swanson What Keeps the F-35 Alive

Figure 2: Figure showing creation of content

In the above figure author and title are combined and the new column content which has the combination of author and title is created. This process is done for better text analysis

```
(0, 15686)      0.28485063562728646
(0, 13473)      0.2565896679337957
(0, 8909)       0.3635963806326075
(0, 8630)       0.29212514087043684
(0, 7692)       0.24785219520671603
(0, 7005)       0.21874169089359144
(0, 4973)       0.233316966909351
(0, 3792)       0.2705332480845492
(0, 3600)       0.3598939188262559
(0, 2959)       0.2468450128533713
(0, 2483)       0.3676519686797209
(0, 267)        0.27010124977708766
(1, 16799)      0.30071745655510157
(1, 6816)       0.1904660198296849
(1, 5503)       0.7143299355715573
(1, 3568)       0.26373768806048464
(1, 2813)       0.19094574062359204
(1, 2223)       0.3827320386859759
(1, 1894)       0.15521974226349364
(1, 1497)       0.2939891562094648
(2, 15611)      0.41544962664721613
(2, 9620)       0.49351492943649944
(2, 5968)       0.3474613386728292
(2, 5389)       0.3866530551182615
(2, 3103)       0.46097489583229645
```

Figure 3: sparse matrix given by Term Frequency Inverse Document Frequency

A sparse matrix is a matrix in which most of the elements are zero. The key idea behind using sparse matrices is to store and operate only on the non-zero elements efficiently. Sparse matrix representations help conserve memory and speed up certain operations.

The above figure is the sparse matrix. The output of the TF-IDF vectorizer is a sparse matrix where each row corresponds to a document (news article) in the dataset, and each column corresponds to a unique term(word)in the entire corpus. The values in the matrix represent the TF-IDF scores for each term in each document.

6. Conclusion and Future Scope

A Logistic Regression model is trained on a collection of news items using TF-IDF vectorization. The algorithm can predict if a given news report is real or phony. The accuracy scores on the training and testing sets are computed to evaluate the model's performance.

In order to improve the system, experiment with various machine learning models, such as neural networks or Random Forest, and investigate sophisticated text processing methods, such as word embeddings. Use cross-validation for improved performance estimates, investigate ensemble approaches for more resilience, and handle unbalanced data by oversampling or under sampling. The combined effect of these improvements is to increase the model's efficiency, flexibility, and accuracy in news article classification.

Data Availability
We used Dataset from Kaggle this datatset is about US politics Size is 20799x5 means 20799 rows and 5 columns Fake News | Kaggle Attributes used are id, title, author, text, label

Conflict of Interest
We do not have any conflict of interest.

Funding Source
NONE

Authors' Contributions
K. Ramya researched literature and conceived the study. M. Yamini involved in protocol development, gaining ethical approval, patient recruitment, and data analysis. K. Lakshmi Prajwala wrote the first draft of the manuscript. M. Jyothirmai reviewed and edited the manuscript and approved the final version of the manuscript.

Acknowledgements
Our guide SK Mulla Almas played a pivotal role in guiding us through the intricacies of the project, offering invaluable insights and support that were instrumental in the successful completion of our endeavor.

References

[1]. Xinyi Zhou, Reza Zafarani "A Survey of Fake News: Fundamental Theories, Detection Methods and Opportunities", ACM journals, Vol.**53**, Issue.**5**, pp.**1-40, 2020.**
[2]. Aswini thota, Priyanka tilak, simrat, Nibrat, "Fake news detection: A deep learning approach", SMU Data Science Review, Vol.**1**, No.3, Article 10, **2018.**
[3]. Z Khanam,B N Alwasel,H Sirafi,M Rashid, "Fake news detetction using machine learning approaches", IOP science, Vol.**1099, 2020.** DOI:10.1088/1757-889X/1099/1/012040
[4]. Xinyi zhou, Reza zafarani, "Network-based Fake news Detection:A Pattern driven Approach", ACM journals, Vol.**21**, Issue.**2**, pp.**48-60, 2019.**
[5]. Z Khanam, B.N. Alwasel, Sirafi, M Rashid, "Fake news detetction using machine learning approaches", IOP science, Vol.**1099, 2020.** DOI:10.1088/1757-889X/1099/1/012040
[6]. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad," Fake news Detection using machine learning ensemble methods, Hindawi, **2020.**
[7]. M. M. V. Y. a. A. Granik, "Determining Fake Statements Made by Public Figures by Means of Artificial Intelligence" International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Vol.**1**, pp.**424-427, 2018.**
[8]. V.L.Rubin and T.Lukoianova "Truth anddeception at the rhetorical structure level". Journal of the Association for Information Science and Technology, Vol.**66**, Issue.**5**, pp.**905-917, 2015.**
[9]. Conroy,J. Niall,L. Victoria L. Rubin,Y. Chen, "Automatic deception detection: Methods for finding fake news," In the Proceedings of the 2015 Association for Information Science and Technology, Vol.**52**, No.**1**, pp.**1-4, 2015.**

AUTHORS PROFILE

Kancharla Ramya pursuing IV B. tech in the stream of Information Technology at Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dist, Andhra Pradesh.

Mudraboina Yamini pursuing IV B. tech in the stream of Information Technology at Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dist, Andhra Pradesh

Kshatri Lakshmi Prajwala pursuing IV B. tech in the stream of Information Technology at Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dist, Andhra Pradesh

Marreddy Jyothirmai pursuing IV B. tech in the stream of Information Technology at Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dist, Andhra Pradesh

