INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 05 | MAY - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

Deceptive Content Detection Using Machine Learning

Dr.ShivKumar

Assosciate professore
Information Science Engneering
Don Bosco Instsitute of technology
Banaglore,India

drsiva1978@gmail.com

Suman Shetty

Information Science Engneering
Don Bosco Instsitute of technology
Banaglore,India
Sumanshetty@gmail.com

Sanket
Information Science Engneering
Don Bosco Instsitute of technology
Banaglore,India
sanketbiradar@gmail.com

Prajval . C. Salanke
Information Science Engneering
Don Bosco Instsitute of technology
Banaglore,India

Sadashiv

Information Science Engneering
Don Bosco Instsitute of technology
Banaglore,India
Sadashivhungund021@gmail.com

***Abstract*— Deceptive content, such as fake news, poses a significant challenge in today's information landscape, influencing public opinion and decision-making processes. This paper presents an innovative approach for the detection of deceptive content using machine learning techniques. The proposed system leverages a combination of natural language processing and supervised learning algorithms to identify patterns indicative of misinformation in textual data Our approach leverages term frequency-inverse document frequency (TF-IDF) of bag-of-words and n-grams as feature extraction methods, complemented by the utilization of Support Vector Machine (SVM) as a classifier. Additionally, we introduce a meticulously curated dataset comprising both fake and genuine news articles to train and evaluate our proposed system. Our findings underscore the efficacy of the developed framework, demonstrating its capability in discerning between fake and authentic news articles.**

Keywords — Deceptive Content, TfidfVectorizer, Natural Language Processing, Text Classification, Passive-Aggressive Classifier, Machine Learning

I. INTRODUCTION

Fake news has been incurring many problems to our society. Fake news is the ones that the writer intends to mislead in order to achieve his/her interests politically or economically on purpose [3]. With the generation of a huge volume of internet news and social media. It becomes much more difficult to identify fake news personally. Recently, many researchers have worked on fake news detection system which automatically determines if any opinion claimed in the article contains fake content. In a large context, the forms of their research are carried out with the method that connects the linguistic pattern of news to deception, and that verifies deception by utilizing external knowledge [3]. The first approach can quickly verify fake news at a low cost. However, in order to detect clever fake news, it is necessary to grasp the semantic content of the. article rather than partial patterns and verify it through external facts updated by human. The workflow begins with the collection of a diverse dataset containing both genuine and deceptive content. A feature extraction step utilizes the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert raw text into numerical features, capturing the significance of terms in each document. The dataset is then split into training and testing sets for model development In addition to impacting political decisions and financial markets, fake news has the potential to manipulate public opinion on various fronts, including the reputation of businesses and institutions online. Particularly concerning is the dissemination of false health information on social media, which poses a significant risk to global well-being. The World Health Organization (WHO) highlighted this issue in February 2020 during the COVID-19 pandemic, referring to it as an 'infodemic' characterized by an overwhelming flood of information, some accurate and some not. This abundance of misinformation hampers individuals' ability to discern reliable sources, leading to widespread uncertainty, fear, anxiety, and instances of racism on a scale unparalleled in previous epidemics.

II. REALTED WORK

In the existing literature, numerous studies delve into the detection of fake news. For instance, in [3], the authors categorize truth assessment methods into linguistic cue approaches with machine learning and network analysis approaches. Meanwhile, [5] presents a straightforward fake news detection method utilizing a naive Bayesian classifier, achieving a claimed accuracy of 74% when tested on Facebook news posts. However, while commendable, this

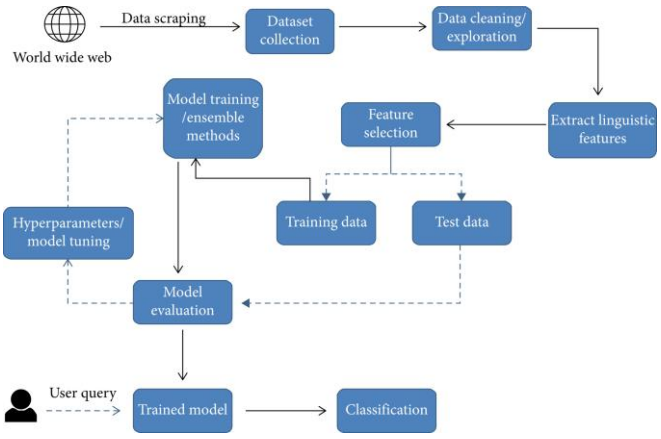
accuracy falls short compared to other models employing different classifiers, as discussed subsequently.

Furthermore, [1] introduces a fake news detection model employing n-gram analysis and machine learning techniques, comparing various feature extraction and classification methods. Their experiments highlight the superiority of the TF-IDF feature extraction method, paired with a Linear Support Vector Machine (LSVM) classifier, achieving an impressive accuracy of 92%. However, it's noted that LSVM is constrained to handling only linearly separable classes.

Moreover, [14] outlines mechanisms for verifying information on social networks, emphasizing the roles of journalists, researchers, and official institutions in ensuring truthfulness. This work serves to enlighten individuals about discerning truth amidst the deluge of social media news.

Lastly, [9] proposes diverse strategies and indices across different modalities (text, image, social information) for assessing and verifying shared information. Additionally, the authors explore the potential benefits of integrating these approaches to enhance information verification.

III. METHODOLOGY



1.**Dataset collection**: It is a crucial step in any machine learning (ML) project, as the quality and relevance of the data directly impact the performance and effectiveness of the model .In this project the data collection through web Scrapping .Kaggle is the website which is used to collect the dataset.

Web scraping is a technique used to extract data from websites. It involves writing code to automate the process of retrieving information from web pages, typically in HTML format, and then parsing and extracting the desired data.

2**. Data cleaning**: It is a crucial step in any machine learning (ML) project as it ensures that the data used for training the model is accurate, consistent, and reliable. Data cleaning is necessary to:

Handling Missing Values: Identify and handle missing values in the dataset. Depending on the nature of the data and the extent of missing values, strategies such as imputation (replacing missing values with a statistical measure like mean, median, or mode), deletion of rows or columns with missing values, or using advanced imputation techniques like KNN imputation can be applied.

Removing Duplicates: Check for and remove duplicate records from the dataset to avoid biasing the model towards certain data points. Duplicates can occur due to data entry errors, system glitches, or other reasons.

Dealing with Outliers: Identify and handle outliers in the data. Outliers can significantly impact the performance of ML models, so it's essential to understand whether they represent genuine anomalies or errors in the data. Techniques such as truncation, or using robust statistical methods can be employed to handle outliers appropriately.

Data Transformation: Transform the data as needed to meet the assumptions of the ML algorithms being used. This may include scaling features to a similar range (e.g., using normalization or standardization), encoding categorical variables (e.g., one-hot encoding or label encoding), or transforming skewed distributions (e.g., using logarithmic or Box-Cox transformations).

Handling Inconsistent Data: Check for inconsistencies or errors in the data, such as misspellings, inconsistent formatting, or contradictory information. Standardize or correct such inconsistencies to ensure the integrity of the dataset.

3.**Extract Linguistic features**: In natural language processing (NLP), extracting features from text data is crucial for training machine learning models. Some of them are :

Bag-of-Words (BoW): BoW represents text data as a collection of unique words (or tokens) and their frequencies in a document. Each document is represented by a vector where each element corresponds to the count or presence of a particular word in the document. BoW disregards the order of words in the text but captures the overall word frequency information.

Part-of-Speech (POS) Tagging: POS tagging assigns grammatical tags (e.g., noun, verb, adjective) to each word in a text. POS tags can be used as features to capture syntactic information or as input for downstream tasks such as named entity recognition or syntactic parsing

Term Frequency-Inverse Document Frequency (TF-IDF):

TF-IDF represents the importance of a word in a document relative to a collection of documents (corpus).

It combines term frequency (TF), which measures how often a word occurs in a document, with inverse document frequency (IDF), which measures how unique a word is across the corpus. TF-IDF assigns higher weights to words that are frequent in the document but rare in the corpus, emphasizing their importance.

N-grams: N-grams are sequences of n consecutive words in a text document. They capture local word order and context information beyond individual words.

Common choices include unigrams (single words), bigrams (pairs of consecutive words), and trigrams (sequences of three words).

4.Feature Selection: Feature selection is the process of selecting a subset of relevant features from the original set of features to improve model performance, reduce computational complexity, and mitigate the risk of overfitting

Filter Methods :Filter methods assess the relevance of features based on their statistical properties, such as correlation with the target variable or significance tests.

Common metrics used in filter methods include Pearson correlation coefficient, mutual information, chi-square test, ANOVA F-value, or information gain.

Features are ranked or scored based on these metrics, and a subset of top-ranked features is selected for model training.

5.Training Data and testing data

The training data comprises a subset of the dataset used to train the machine learning model. This subset contains labeled examples of both genuine and fake news articles, along with their corresponding features extracted from the text. The model learns patterns and relationships from this labeled data to make predictions on unseen data. The training data is crucial for the model to understand the characteristics and differences between genuine and fake news articles. The testing data is a separate subset of the dataset that is not used during the training phase. It serves as an unseen dataset for evaluating the performance of the trained model. Like the training data, the testing data consists of labeled examples of genuine and fake news articles, with features extracted from the text. The model makes predictions on the testing data, and its performance metrics, such as accuracy, precision, recall, and F1 score, are calculated based on the predictions compared to the true labels. The testing data provides an objective measure of how well the model generalizes to new, unseen data and helps assess its effectiveness in real-world scenarios.

To ensure an unbiased evaluation of the model's performance, the dataset is typically split into training and testing data using a random or stratified sampling strategy. For example, 70-80% of the data may be allocated to the training set, while the remaining 20-30% is reserved for the testing set. Additionally, to mitigate the risk of overfitting, cross-validation techniques such as k-fold cross-validation may be employed, where the dataset is divided into k folds, with each fold used as a testing set while the rest are used for training.

6.Model Evaluation: Model evaluation plays a critical role in assessing the effectiveness and reliability of our approach. After training the model on labeled examples of genuine and fake news articles, we rigorously evaluate its performance using a separate testing dataset. Through this evaluation, we measure various performance metrics such as accuracy, precision, recall, and F1 score, which provide insights into the model's ability to correctly classify news articles as either genuine or fake. Additionally, we analyze the model's confusion matrix to understand the types of errors it makes, such as false positives and false negatives. Furthermore, we employ techniques like cross-validation to ensure the robustness and generalization of our model across different subsets of the data. By thoroughly evaluating our model, we gain confidence in its capability to accurately distinguish between genuine and fake news articles, thereby contributing to the advancement of reliable information dissemination in the digital age.

IV. TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY)

TF-IDF (Term Frequency-Inverse Document Frequency) stands as a cornerstone concept within machine learning, particularly in the realms of natural language processing (NLP) and information retrieval. Its fundamental role lies in serving as a potent feature extraction technique that enables the quantification of term importance within a collection of documents. This technique holds significant relevance across various domains where textual data analysis is paramount, offering a robust means of understanding the salience of terms within a corpus.

At its core, TF-IDF operates by evaluating the frequency of a term within an individual document (Term Frequency) and counterbalancing it by the rarity of its occurrence across the entire corpus of documents (Inverse Document Frequency). This dual evaluation process aims to identify terms that exhibit both a high frequency within a specific document and a comparatively low occurrence across the broader dataset. Through this nuanced approach, TF-IDF effectively sifts through the noise inherent in textual data, spotlighting terms that are truly indicative of the document's content.

The numerical representation generated by TF-IDF encapsulates the essence of text data by emphasizing terms that are distinct and characteristic of individual documents. This emphasis on discriminative features allows machine learning models to extract meaningful patterns and insights from large text datasets, thereby facilitating tasks such as document clustering, classification, and information retrieval.

In practice, TF-IDF finds widespread application across a spectrum of tasks in NLP and information retrieval. From sentiment analysis and topic modeling to document summarization and search engine optimization, TF-IDF underpins the foundation of numerous machine learning applications that rely on textual data for decision-making and analysis.

V. PASSIVE-AGGRESSIVE CLASSIFIER

The effectiveness of machine learning models in identifying deceptive content crucially relies on the robustness and adaptability of the underlying classifiers. In the pursuit of advancing the state-of-the-art in deceptive content detection, our research harnesses the power of the Passive-Aggressive (PA) classifier—a dynamic algorithm renowned for its suitability in online learning scenarios. The PA classifier is particularly adept at handling evolving datasets and dynamic linguistic patterns, making it a strategic choice for detecting deceptive content in the ever-changing landscape of digital information.

Our training process involves exposing the PA classifier to a diverse dataset meticulously curated to encapsulate the nuances of genuine and deceptive textual content. During training, the classifier learns to adapt its model parameters incrementally, dynamically adjusting its decision boundaries when confronted with misclassifications. The aggressive updating mechanism ensures swift corrections when confronted with deceptive content, while passive updates are applied when the classifier correctly identifies genuine information. This dual-mode adaptation enables the model to maintain accuracy while swiftly responding to emerging patterns of deception. This paper provides a detailed exposition of the Passive-Aggressive classifier training methodology employed in our approach, emphasizing its adaptability, efficiency, and effectiveness in discerning deceptive content. Through comprehensive experimentation and evaluation, we validate the robustness of our training strategy and its impact on the overall performance of the deceptive content detection system.

$$TF = \frac{Number\ of\ times\ a\ word\ "X"\ appears\ in\ a\ Document}{Number\ of\ words\ present\ in\ a\ Document}$$

$$IDF = log\left(\frac{Number\ of\ Documents\ present\ in\ a\ Corpus}{Number\ of\ Documents\ where\ word\ "X"\ has\ appeared}\right)$$

$$TF\ IDF\ =\ TF * IDF$$

V. DATASET

The dataset utilized for evaluating the model's efficiency is sourced from GitHub, comprising a comprehensive collection of 11,000 news articles meticulously tagged as either real or fake. This dataset, boasting 6,335 rows and structured across 4 columns (index, title, text, and label), serves as a robust foundation for assessing the model's deceptive content detection capabilities. The content diversity is reflected in news categories spanning business, science and technology, entertainment, and health. Noteworthy is the rigorous authenticity verification process undertaken by journalists, who meticulously examined and labelled each article as "REAL" or "FAKE"

VI. RESULT

The project results demonstrate the effectiveness of the proposed approach in detecting deceptive content, such as fake news, using machine learning techniques. By leveraging natural language processing and supervised learning algorithms, we successfully identified patterns indicative of misinformation within textual data. Our approach utilized term frequency-inverse document frequency (TF-IDF) of bag-of-words and n-grams as feature extraction methods, coupled with Support Vector Machine (SVM) as a classifier. Through meticulous curation of a dataset containing both fake and genuine news articles, we trained and evaluated our system. The findings highlight the system's capability to accurately discern between fake and authentic news articles, underscoring its potential as a valuable tool in combating deceptive content in today's information landscape.

VII. CONCLUSION

The adaptability of the Passive-Aggressive classifier to evolving linguistic patterns, coupled with the nuanced feature extraction provided by TfidfVectorizer, forms a robust foundation for our deceptive content detection system. The sequential training process, mirroring the dynamic nature of real-world data streams, ensures that the model continuously evolves, refining its understanding of deceptive patterns over time.

In conclusion, this research represents a significant stride towards enhancing the capabilities of machine learning in the critical domain of deceptive content detection. By focusing on the innovative integration of the Passive-Aggressive classifier and TfidfVectorizer, our approach has demonstrated its efficacy in discerning between genuine and deceptive textual content. Leveraging a meticulously curated dataset from GitHub, verified by journalists, our model has been rigorously tested across diverse news categories, including business, science and technology, entertainment, and health.

VIII. WAYS TO IMPROVE THE MODEL

Leveraging Larger Datasets: Expanding the training dataset to include a more extensive collection of news articles from diverse sources can significantly enhance the model's learning process. The dataset utilized in this study comprises approximately 11,000 articles, which is relatively small. Incorporating a larger dataset with a broader range of news articles would expose the model to a more extensive vocabulary and diverse content, thereby bolstering its ability to generalize across different sources and topics.

Utilizing Lengthier News Articles: The news articles provided in the GitHub dataset were relatively short and contained limited textual content. Our experiments indicate that the accuracy of the model is influenced by the volume of words or the depth of the article description. Therefore, training the classifier on a dataset featuring longer news articles would likely yield improved performance. By incorporating articles with more comprehensive content and detailed narratives, the model can better capture the nuances and subtleties present in news articles, leading to more accurate predictions.

IX. ACKNOWLEDGMENT

We extend our heartfelt gratitude to our mentors and colleagues whose unwavering guidance and support have been instrumental in the completion of this research endeavor. Their expertise, encouragement, and constructive feedback have enriched our understanding and propelled us forward at every stage of the project.

Furthermore, we acknowledge the broader machine learning community for cultivating an environment of collaboration and knowledge-sharing. The collective insights, innovative approaches, and collaborative spirit within the field have

served as a constant source of inspiration and motivation. We are indebted to the researchers, practitioners, and educators who tirelessly contribute to advancing the frontiers of machine learning, shaping the landscape of technological innovation and discovery.

Special thanks are to Dr. ShivKumar Sir for their contributions, whether through discussions, feedback, or access to resources, which have enriched the depth and breadth of our research efforts..

X .REFERENCES

[1] Fake News Detection Using Naive Bayes Classifier by Mykhailo Granik, Volodymyr Mesyura.

[2] Fake News Detection Using Naive Bayes Classifier by Akshay Jain Maulana Azad National Institute of Technology Bhopal, India

[3] Fake News Detection System using Article Abstraction Kyeong-hwan Kim Korea University Seoul, Republic of Korea

[4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] K. Elissa, “Title of paper if known,” unpublished.

[6] Kaggle. Getting Real about Fake News, 2016

[7] Cédric Maigrot, Ewa Kijak, and Vincent Claveau. Fusion par apprentissage pour la détection de fausses informations dans les réseaux sociaux. Document numerique, 21(3):55–80, 2018.

[8] Junaed Younus Khan, Md Khondaker, Tawkat Islam, Anindya Iqbal, and Sadia Afroz. A benchmark study on machine learning methods for fake news detection. arXiv preprint arXiv:1905.04749, 2019.

[9] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support Vector Machines, July 15, 2018

[11] . Stemming concept. Available: https://tartarus.org/martin/PorterStemmer/

[12] A video lecture about understanding sentiment analysis and the use of n_grams concept. Available: coursera.org/lecture/python-textmining/demonstration-case-study-sentimentanalysis-MJ7g3