IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NEWS CLASSIFICATION USING ML ALGORITHMS: AN EXPLORATION OF EFFICIENT INFORMATION CATEGORIZATION

[1]Md Jahirul Islam,[2] Saeed Sarwar Anas

[1]Designation of 1st Author, [2Designation] of 2nd Author

[1] Department of Computer Science and Engineering, [1]City University, Dhaka, Bangladsh

Abstract: In today's rapidly evolving digital era, the vast amount of news articles available online makes it challenging for individuals to navigate through the sea of information. To address this issue, automated news classification systems have gained significant attention. These systems aim to categorize news articles into predefined classes, allowing users to efficiently access the information they seek. This research paper explores the application of machine learning (ML) algorithms for news classification, providing a comprehensive analysis of their effectiveness and performance.

I. INTRODUCTION

The exponential growth of online news sources has created a pressing need for effective methods of information categorization. Manual classification by humans is not only time-consuming but also prone to subjectivity and inconsistency. As a solution, the utilization of machine learning algorithms has emerged as a promising approach for automating the news classification process. The objective of this research paper is to investigate the efficacy of ML algorithms in accurately classifying news articles into predefined categories. By leveraging the power of computational techniques, we aim to enable the development of intelligent systems that can autonomously analyze and categorize news articles, facilitating easy access to relevant information for users. This study explores a wide range of ML algorithms, including but not limited to Naive Bayes, Support Vector Machines (SVM), Random Forest, and Neural Networks, to evaluate their performance in news classification tasks. The algorithms are trained and tested on large-scale datasets comprising diverse news articles, encompassing various topics and domains. Furthermore, the research investigates the impact of different feature extraction techniques and text representation models on the performance of the ML algorithms. It explores traditional approaches like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), as well as more advanced methods like word embeddings and deep learning-based representations. This analysis aims to provide insights into the most effective techniques for feature extraction and text representation in the context of news classification. To evaluate the performance of the ML algorithms, various metrics such as accuracy, precision, recall, and F1-score are employed. Comparative analysis is conducted to identify the strengths and weaknesses of each algorithm, shedding light on their suitability for news classification tasks.The findings of this research paper contribute to the field of automated news classification by providing a comprehensive evaluation of ML algorithms and their performance. The results offer valuable insights for researchers, practitioners, and developers seeking to build robust and efficient news classification systems. Additionally, the study opens avenues for further exploration and refinement of existing ML techniques, with the ultimate goal of enhancing the accessibility and usability of news information in the digital age.

11. DATA ANLYSIS

The BBC News dataset, available on Kaggle, serves as a valuable resource for researchers and data enthusiasts seeking to delve into the realm of text analysis and natural language processing (NLP). With its diverse collection of news articles spanning different categories, this dataset offers a rich opportunity to extract insights and understand patterns in the news landscape. In this article, we embark on an exploration of the BBC News dataset, leveraging its contents to gain valuable insights and showcase the power of textual analysis. The BBC News dataset comprises approximately 2,225 news articles published by BBC News in the early 2000s. Each article is associated with one of five categories: business, entertainment, politics, sport, and tech. By examining these articles, we can gain a deeper understanding of how news content is distributed across various domains and uncover trends that might have existed during that time period. Before delving into analysis, it is crucial to preprocess the text data. This step typically involves removing stopwords, punctuation, and special characters, as well as tokenizing the text into individual words or phrases. Additionally, techniques like stemming or lemmatization may be applied to normalize the words and reduce dimensionality. Once the data is preprocessed, we can represent it in a format suitable for analysis. Traditional methods such as the Bag-of-Words (BoW) or TF-IDF (Term Frequency-Inverse Document Frequency) can be employed to transform the text into numerical vectors. Alternatively, more advanced techniques like word embeddings, such as Word2Vec or GloVe, can capture semantic relationships between words and improve the representation of the text. With the dataset prepared, we can embark on exploratory data analysis (EDA) to uncover patterns and insights. We can start by visualizing the distribution of news articles across the five categories using bar charts or pie charts. This provides an overview of the dataset's composition and helps identify any class imbalances that might impact subsequent analyses. Furthermore, we can examine word frequencies and generate word clouds to visualize the most common words in each category.

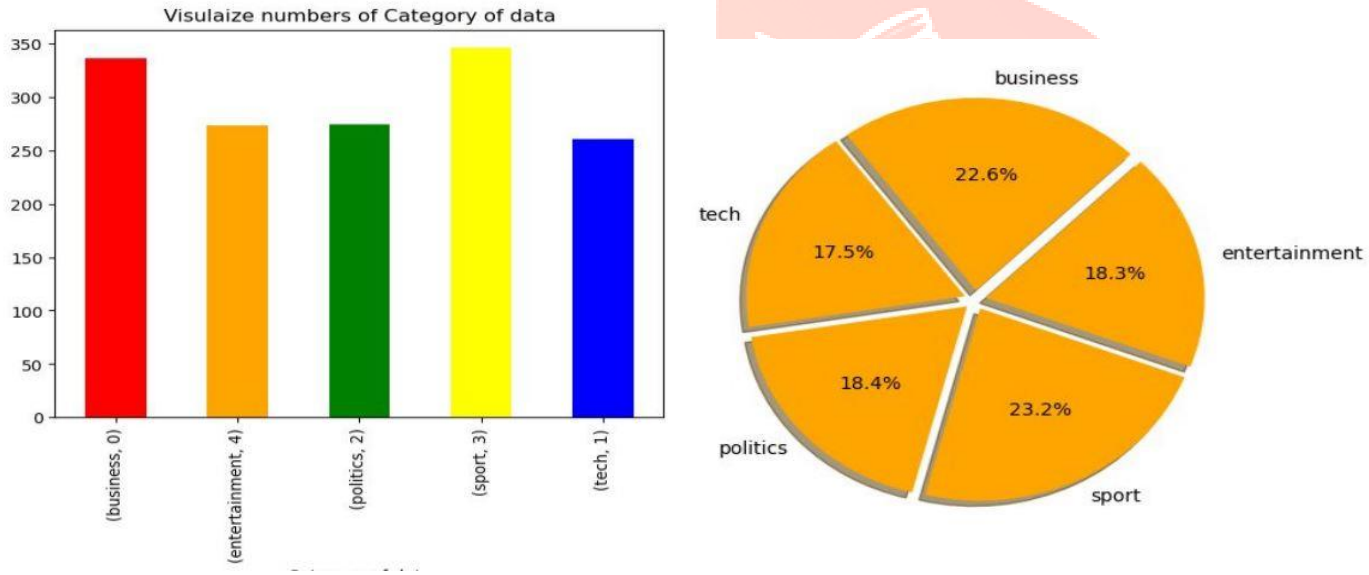

Fig1: Data Visualization

This analysis can reveal domain-specific terminology and highlight the distinguishing characteristics of each category. To delve deeper into the content of the news articles, we can employ topic modeling techniques such as Latent Dirichlet Allocation (LDA) or Non-Negative Matrix Factorization (NMF). These algorithms can identify latent topics within the dataset and assign relevant keywords to each topic. By analyzing the resulting topics and their associated keywords, we gain insights into the dominant themes and subjects present in the BBC News dataset. Sentiment analysis allows us to determine the overall sentiment or emotion expressed in each news article. By utilizing techniques such as lexicon-based analysis or machine learning-based approaches, we can classify articles as positive, negative, or neutral. Analyzing sentiment across different categories can reveal interesting trends, such as variations in sentiment polarity or intensity based on the news domain.

11I.METHDOOLOGY

Support Vector Machines provide a powerful framework for tackling classification and regression problems by finding optimal decision boundaries and leveraging the concept of margins. Their versatility, robustness, and ability to handle non-linear relationships make them a valuable tool in machine learning. We want to optimize the distance between the data points and the hyperplane in the SVM method. The loss function known as hinge loss aids in maximizing the margin.The goal of the SVM method is to increase the distance between the data points and the hyperplane. Hinged loss is the loss function that aids in maximizing the margin.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases} \qquad c(x, y, f(x)) = (1 - y * f(x))_+$$

……...……(i)

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Unlike a single decision tree, which can easily overfit the data, Random Forest uses a combination of several trees to improve generalization and robustness. Each tree in the Random Forest is trained on a random subset of the original data, and the final prediction is determined by aggregating the predictions of all the individual trees.K-Nearest Neighbors (KNN) classification is a popular machine learning algorithm used for both binary and multi-class classification tasks. It is a non-parametric and instance-based learning algorithm, meaning it does not make any assumptions about the underlying data distribution and makes predictions based on the similarity of instances in the feature space. Decision Trees are versatile and interpretable models that can be used for classification and regression tasks. They provide a straightforward and intuitive approach to decision-making based on learned rules from the training data.Gaussian Naive Bayes and Multinomial Naive Bayes are two variants of the Naive Bayes algorithm, which is a popular probabilistic machine learning algorithm used for classification tasks. Naive Bayes models are based on the Bayes' theorem and assume that the features are conditionally independent given the class label.

The statistics and machine learning algorithm logistic regression is commonly used for binary classification applications.The logistic function is used to model the connection between a collection of independent variables (features) and a binary dependent variable (the target or class label).The benefits of logistic regression in numerous fields are its clarity, interpretability, and simplicity.The main features of logistic regression are as follows:Using the logistic function or sigmoid function, logistic regression models the relationship between the features (independent variables) and the likelihood of the binary result (dependent variable)..

logistic function maps a linear combination of features to a probabilistic value from 0 to 1, which represents the positivity of the class. The model is trained by means of maximum likelihood estimation (MLE) or optimization techniques such as gradient descent. The model learns the best weights for each feature based on maximizing the probabilistic value of observed data. The weights represent the effect of each feature for each positive class. The model can then make predictions using the learned weights for each feature. It can apply the learned weights to new instances of the class. It calculates the log (or logit) of the positivity class and turns it into a probabilistic function. The decision boundary (or threshold) between the two classes is determined by the predicted probabilistic value (normally 0.5) of the class.

Logistic regression is linear. However, Logistic regression also has the following drawbacks: It predicts linear relationship between features and log-odd of positive class. It may not capture non-linear patterns. It may be sensitive to non-norms or to influential observations. It is designed primarily for binary classification. Models may need to be modified for multidimensional classification tasks (for example, one-versus-rest, softmaxregression) Logistic regression is widely used in finance and healthcare, marketing and social sciences to solve binary classification problems. It is a basic and interpretable machine learning model. It can be expanded and combined with different techniques for more challenging tasks.

IV. RESULT

Fig2: Accuracy of different Machine learning model.

We use different type of machine learning model such as Support Vector machines Random Forest KNN Classification Decision Trees Gaussian & Multinomial Naive bayes Logistic Regressions is important to note that the statement "random forest is the best model for classifying articles" may not be universally true in all scenarios. The performance of different machine learning models can vary depending on the specific dataset, the nature of the problem, and various other factors. While Random Forest is a powerful and popular algorithm, it is not always guaranteed to be the best choice for every classification task. The choice of the best model for classifying articles depends on several factors, including the size and quality of the dataset, the complexity of the classification problem, the available features, and the specific evaluation metrics used to assess model performance.

REFERENCES

[1] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In ISSP'12.

[2] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In IJCAI'07.

[3] M. H. Hossen and W. Hu, "Hypergraph Regularized SVM and Its Application Emotion Detection," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2021, pp. 133-139, doi: 10.1109/IPEC51340.2021.9421069.

[4] credit card payment system by machine learning classifiers", Proc. SPIE 12339, Second International Conference on Cloud Computing and Mechatronic Engineering (I3CME 2022), 1233927 (28 September 2022);

[5] Sharmin Sultana, Md Helal Hossen, Md Jahidul Islam, Jubayer Ahmed, Afsana Mou (2023). Detection, reduction and filtration of cancer cells through a new DNA polymerization sequence approach. North American Academic Research, 6(1), 230-239. doi: https://doi.org/10.5281/zenodo.7641213

[6] Md Babul Islam, KhandakerSajidul Islam, Md Helal Khan, Abdullah MMA Al Omari, and Swarna Hasibunnahar "Detect deception on banking credit card payment system by machine learning classifiers", Proc. SPIE 12339, Second International Conference on Cloud Computing and Mechatronic Engineering (I3CME 2022), 1233927 (28 September 2022); https://doi.org/10.1117/12.2655113

[7] Hossen M, H., Hasan, M. M., & Hu, W. (2021 Join Public Key and Private Key for Encrypting Data. North American Academic Research, 4(3), 256-267. doi: https://doi.org/10.5281/zenodo. 4661097

[8] Alessandro Bondielli and Francesco Marcelloni, "A survey on fake news and rumour detection techniques", *Information Sciences*, vol. 497, pp. 38-55, 2019

[9] Hossen, Md Helal, et al. "Digital Revolution in the Agriculture Based on Data Science." 2022 2nd Asia Conference on Information Engineering (ACIE). IEEE, 2022.

[10] M. H. Hossen, M. M. Hasan, I. K. Sajidul and W. Hu, "Digital Revolution in the Agriculture Based on Data Science," 2022 2nd Asia Conference on Information Engineering (ACIE), Haikou, China, 2022, pp. 6-12, doi: 10.1109/ACIE55485.2022.00010.

[11] Sharmin Sultana, Md Helal Hossen, Md Jahidul Islam, Jubayer Ahmed, Afsana Mou (2023). Detection, reduction and filtration of cancer cells through a new DNA polymerization sequence approach. North American Academic Research, 6(1), 230-239. doi: https://doi.org/10.5281/zenodo.7641213

[12] Benjamin Horne and Sibel Adali, "This just in: Fake news packs a lot in title uses simpler repetitive content in text body more similar to satire than real news", *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 2017.

[13] Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq et al., "Detecting fake news using machine learning and deep learning algorithms", *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pp. 1-5, 2019.

[14] R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen and R. Hasan, "Heart Disease Detection Using ML," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0983-0987, doi: 10.1109/CCWC57344.2023.10099294.

[15] Manal Iftikhar, Arshad Ali, "Fake News Detection using Machine Learning", 2023 3rd International Conference on Artificial Intelligence (ICAI), pp.103-108, 2023

[16] X. Jose, S. D. M. Kumar and P. Chandran, "Characterization, Classification and Detection of Fake News in Online Social Media Networks," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 759-765, doi: 10.1109/MysuruCon52639.2021.9641517.

[17] A. Farjana et al., "Predicting Chronic Kidney Disease Using Machine Learning Algorithms," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 1267-1271, doi: 10.1109/CCWC57344.2023.10099221.

[18] Ferry Wahyu Wibowo, Akhmad Dahlan, Wihayati, "Detection of Fake News and Hoaxes on Information from Web Scraping using Classifier Methods", *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp.178-183, 2021.