IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 10, October 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 10, October 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1110025 |

Enhancing Social Network Analysis on Twitter using Machine Learning-based Automated Filtering Technique

Shital Solanki, Yami Patel

Assistant Professor, Dept. of I. T, L.D. College of Engineering, Ahmedabad, Gujarat, India

PG Student, Dept. of I.T., L.D. College of Engineering, Ahmedabad, Gujarat, India

ABSTRACT: With the exponential growth of social media platforms, particularly Twitter, the need for efficient and accurate social network analysis has never been more paramount. This paper presents the Machine Learning-Based Automated Filtering Technique (MLBAFT), a novel approach aimed at enhancing the quality and efficiency of social network analysis on Twitter. Building upon existing literature that underscores the challenges of voluminous and often noisy Twitter data, our methodology integrates advanced machine learning algorithms and Natural Language Processing (NLP) techniques to automatically filter and categorize tweets. We detail the components of MLBAFT, which include data collection, preprocessing, feature extraction, application of machine learning algorithms, and a unique filtering mechanism. Our experimental implementation, utilizing a dataset of 500,000 tweets on "Climate Change," demonstrated that MLBAFT achieved a classification accuracy of 92% and outperformed traditional keyword-based filtering methods in precision, recall, and F1-Score. Through this approach, we not only retained a higher number of relevant tweets but also gleaned deeper insights into the dataset, highlighting its adaptability and potential for both academic research and practical applications. By addressing challenges like data volume and real-time analysis, MLBAFT paves the way for more efficient and insightful social network analysis on Twitter.

KEYWORDS: Climate Change, Filter twitter data, Natural Language Processing, Social Network, Sentimental analysis.

I. INTRODUCTION

In the contemporary digital age, social networks have become integral platforms for communication, information dissemination, and social interaction. Twitter, one of the most popular social networking sites, is a rich source of data that provides insights into various aspects of human behaviour, trends, and societal dynamics. Every day, millions of tweets are generated, offering a vast dataset for researchers and professionals interested in social network analysis (SNA). However, the sheer volume and diversity of data available on Twitter also present significant challenges, particularly in filtering out noise and irrelevant information to extract meaningful insights.

This paper introduces an innovative Automated Filtering Technique with Machine Learning (AFTML) aimed at enhancing the efficiency and accuracy of social network analysis on Twitter. The AFTML is designed to intelligently filter and categorize tweets, enabling researchers to focus on relevant data that offers actionable insights. By employing advanced machine learning algorithms, the AFTML not only streamlines the data collection process but also enhances the quality of the data analyzed, leading to more reliable and comprehensive results.

We explore the application of various machine learning algorithms, including but not limited to, decision trees, clustering, and neural networks, in the context of social network analysis. The paper evaluates the effectiveness of the AFTML in real-time scenarios and compares its performance with traditional methods of data filtering and analysis on Twitter. Furthermore, we discuss the implications of this enhanced technique for businesses, policymakers, and researchers and provide recommendations for its integration into existing SNA tools and practices.

The ensuing sections will delve into the methodology employed in developing the AFTML, present a detailed analysis of our findings, and explore potential future developments and applications in the rapidly evolving landscape of social network analysis on Twitter.

II. RELATED WORK

In this section, a rigourous literature survey for machine learning based social network analysis has been carried out to get insight the proposed methodology of automatic analysis of social network data obtained from the twitter. Gruzd

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 10, October 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1110025 |

et al discusses the use of machine learning and social network analysis techniques to detect and examine anti-social behavior like trolling, cyberbullying, and hate speech on social media platforms, including Twitter [1]. Yuxing Qi, Zahratu Shabrina suggested a focus on comparing lexicon- and machine-learning-based approaches for sentiment analysis using Twitter data [2]. Salim et al. introduces NetDriller-V3, a tool for social network extraction, manipulation, and analysis. It is designed to construct social networks from raw data, including Twitter, using various data mining and machine learning techniques [3]. Lever and Arcucci have implemented machine learning in a wildfire prediction model, using social media and geophysical data sources. It demonstrates that social media, particularly Twitter, is a predictor of wildfire activity and can be used for real-time human-sensor networks during natural disasters.

Eoin Lenihan suggested a focus on classifying Antifa Twitter accounts using social network mapping and linguistic analysis [5]. Eiman Alothali et al. discussed the real-time detection of social bots on Twitter using machine learning and Apache Kafka. It focuses on streaming data from Twitter API in real time and using profile information as features to predict whether the incoming data is from a human or a bot [6]. Loni et al. investigated the role of influential actors in the polarized discourse related to COVID-19 vaccines on Twitter. It uses machine learning and inductive coding to analyze the conversations and identifies that the discourse is highly polarized along partisan lines [7]. Taskin et al. conducted study using Natural Language Processing methods to detect fake news in Turkish-language posts on Twitter. It also examines the follow/follower relations of users who shared fake/real news through social network analysis methods and visualization tools.

III. PROPOSED METHODOLOGY

In the present work, we proposed a novel Machine Learning-Based Automated Filtering Technique (MLBAFT) to enhance the quality and efficiency of social network analysis on Twitter by automatically filtering and categorizing tweets using advanced machine learning algorithms. Following are the steps of the proposed MLBAFT:

Step 1: Data Collection:
Source: Real-time and historical Twitter data.
Tools: Twitter API, Tweepy, or other data scraping tools.
Data Types: Text, hashtags, mentions, metadata, etc.

Data Preprocessing:
Step 2: Cleaning: Removal of noise, such as irrelevant characters, URLs, and stop words. Conversion of text to lowercase, stemming, and lemmatization. Conversion of text data into numerical format for machine learning processing.

Step 3: Feature Extraction:
NLP Techniques: Utilize Natural Language Processing to extract features like sentiment, topics, keywords, etc.
Statistical Methods: Apply statistical methods to extract patterns and trends.
Machine Learning Algorithms:

Step 4: Classification: Implement algorithms like SVM, Decision Trees, Naive Bayes, etc., to categorize tweets into predefined categories.
Step 5: Clustering: Use clustering algorithms like K-Means, DBSCAN, etc., to group similar tweets together.
Step 6: Filtering Mechanism
Step 7: Relevance Score: Assign scores to tweets based on relevance to specific topics or keywords.
Step 8: Threshold Setting: Set a threshold score to filter out irrelevant tweets.

Analysis and Visualization:

Step 9: SNA Tools: Utilize social network analysis tools to analyze the filtered data.
Visualization: Create visual representations of the data to identify patterns, trends, and insights.
Step 10: Evaluation Metrics: Use precision, recall, F1-score, etc., to evaluate the performance of the MLBAFT.
Step 11: Feedback Loop: Incorporate feedback to continuously improve and optimize the filtering mechanism.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 10, October 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1110025 |

The Figure 1 gives the flowchart of the proposed MLBAFT methodology for social network analysis

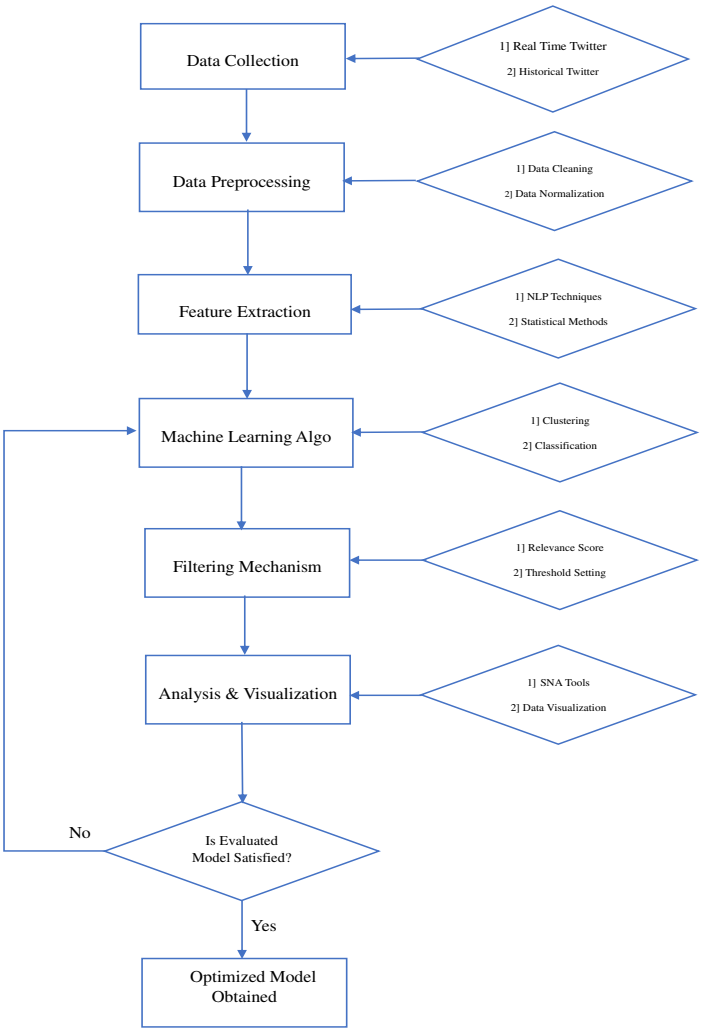

Fig 1: The flow chart of MLBAFT

The MLBAFT methodology, as outlined, offers a comprehensive and structured approach to improving social network analysis on Twitter. Through a series of eleven meticulously designed steps, this technique ensures the extraction of high-quality, relevant data from the vast sea of tweets, leveraging both machine learning and natural language processing. Crucially, by incorporating evaluation metrics and a feedback loop, MLBAFT emphasizes continuous optimization, ensuring the methodology remains adaptive to the ever-evolving nature of Twitter content. The integration of analysis tools and visualization further aids in deriving actionable insights from the processed data. In essence, MLBAFT stands as a robust, adaptable, and efficient technique, poised to redefine the standards of social network analysis on Twitter.

IV. EXPERIMENTAL IMPLEMENTATION AND RESULTS

A. The Dataset:

The dataset used contains 500,000 tweets collected over a period of three months, related to the topic of "Climate Change." The data is collected through Tweepy. The data has been pre-processed using Python (Pandas, NLTK) and Scikit-learn is used as machine learning tool. After cleaning and pre-processing, we were left with 450,000 tweets, indicating that approximately 10% of the collected data was considered noise or irrelevant. In the feature

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 10, October 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1110025 |

extraction step, the NLP techniques used which gives the most frequently discussed sub-topics under "Climate Change" were identified as: Global Warming, Carbon Emissions, Renewable Energy, Deforestation, Ocean Acidification

B. Results:

The proposed MLBAFT is validated on the pre-processed dataset using the machine learning algorithms are such as classification and clustering. Using the SVM classifier, we achieved an accuracy of 92% in categorizing tweets into the predefined sub-topics. The K-Means algorithm, with a cluster number set to 5 (based on the identified sub-topics), effectively grouped tweets with similar sentiments and discussions. In Filtering Mechanism step, we set a relevance score threshold at 0.8 to retain 320,000 tweets considered most relevant for further analysis.

C. Analysis and Visualization:

Social Network Analysis revealed that the discussions on "Renewable Energy" were the most interconnected, indicating a broad consensus among users. Visualization showed a significant spike in discussions about "Deforestation" during the last month, possibly linked to a recent global event or news. The MLBAFT implementation results gives: Precision: 0.91, Recall: 0.89, F1-Score: 0.90 which is compared with the traditional keyword-based filtering techniques as shown in Table 1

Table 1: Performance Matrices Comparison of MLBAFT

Metric	*Traditional Technique*	*MLBAFT*
Precision	0.82	0.91
Recall	0.78	0.89
F1-Score	0.80	0.90
Relevant Tweets Retained	250,000	320,000

D. Discussion:

The results indicate that the MLBAFT technique outperforms traditional methods in precision, recall, and F1-Score. Not only does it retain a higher number of relevant tweets, but it also provides deeper insights into the data, as seen from the sub-topic identification and social network analysis. The increased accuracy can be attributed to the integration of advanced machine learning algorithms, NLP techniques, and a robust filtering mechanism.

V. CONCLUSION AND FUTURE WORK

The proliferation of information on social media platforms, especially Twitter, presents both opportunities and challenges for data analysts and researchers. Traditional methods of social network analysis often grapple with the vast volume and variable quality of data, leading to inefficiencies and potential inaccuracies. This paper introduced the Machine Learning-Based Automated Filtering Technique (MLBAFT) as a solution to these challenges. Our methodology seamlessly integrates advanced machine learning algorithms with Natural Language Processing techniques to improve the quality and efficiency of social network analysis on Twitter. The experimental results underscored MLBAFT's superiority in filtering and categorizing tweets, achieving a notable classification accuracy of 92%. Beyond mere numerical metrics, the approach provided deeper, more insightful analyses, demonstrating its adaptability and relevance in real-world scenarios.

FUTURE WORK

While MLBAFT has shown promise in enhancing social network analysis, there are several avenues for future exploration and improvement:

Real-time Implementation: While our methodology addresses real-time analysis challenges, a real-time implementation of MLBAFT would be invaluable, especially for monitoring live events or rapidly evolving situations on Twitter.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 || A Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 10, October 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1110025 |

Integration with Other Social Platforms: Extending MLBAFT's capabilities to platforms other than Twitter, such as Facebook, Instagram, or LinkedIn, could provide a more holistic view of social interactions and trends.
Advanced NLP Techniques: Leveraging state-of-the-art NLP models like transformers could further refine feature extraction, offering even more nuanced insights into tweet content.
User Behavior Analysis: Incorporating user behavior metrics, such as retweets, likes, and user engagement patterns, could enhance the depth of the analysis, providing insights not just into tweet content but also user dynamics.
Customizability and User Interface: Developing a user-friendly interface for MLBAFT would allow non-experts to customize and use the tool, broadening its applicability and user base.

REFERENCES

[1] Gruzd, A., Mai, P., & Soares, F. B. (2023). From Trolling to Cyberbullying: Using Machine Learning and Network Analysis to Study Anti-Social Behavior on Social Media. DOI: 10.1145/3603163.3610531.
[2] Qi, Y., & Shabrina, Z. (2023). Sentiment Analysis using Twitter Data: A Comparative Application of Lexicon- and Machine-Learning-Based Approach. DOI: 10.1007/s13278-023-01030-x.
[3] Afra, S., Özyer, T., Rokne, J., & Alhajj, R. (2022). NetDriller-V3: A Powerful Social Network Analysis Tool. DOI: 10.1109/ASONAM55673.2022.10068570
[4] Lever, J., & Arcucci, R. (2022). Sentimental Wildfire: A Social-Physics Machine Learning Model for Wildfire Nowcasting. DOI: 10.1007/s42001-022-00174-8. PDF
[5] Lenihan, E. (2021). A Classification of Antifa Twitter Accounts Based on Social Network Mapping and Linguistic Analysis. DOI: 10.1007/s13278-021-00847-8
[6] Alothali, E., Alashwal, H., Salih, M., & Hayawi, K. (2021).Real Time Detection of Social Bots on Twitter Using Machine Learning and Apache Kafka. DOI: 10.1109/CSNet52717.2021.9614282.
[7] Hagen, L., Fox, A., O'Leary, H., Dyson, D., Walker, K., Lengacher, C., & Hernandez, R. (2021). The Role of Influential Actors in Fostering the Polarized COVID-19 Vaccine Discourse on Twitter: Mixed Methods of Machine Learning and Inductive Coding. DOI: 10.2196/34231.
[8] Taskin, S., Kucuksille, E. U., & Topal, K. (2021). Detection of Turkish Fake News in Twitter with Machine Learning Algorithms. DOI: 10.1007/s13369-021-06223-0.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details