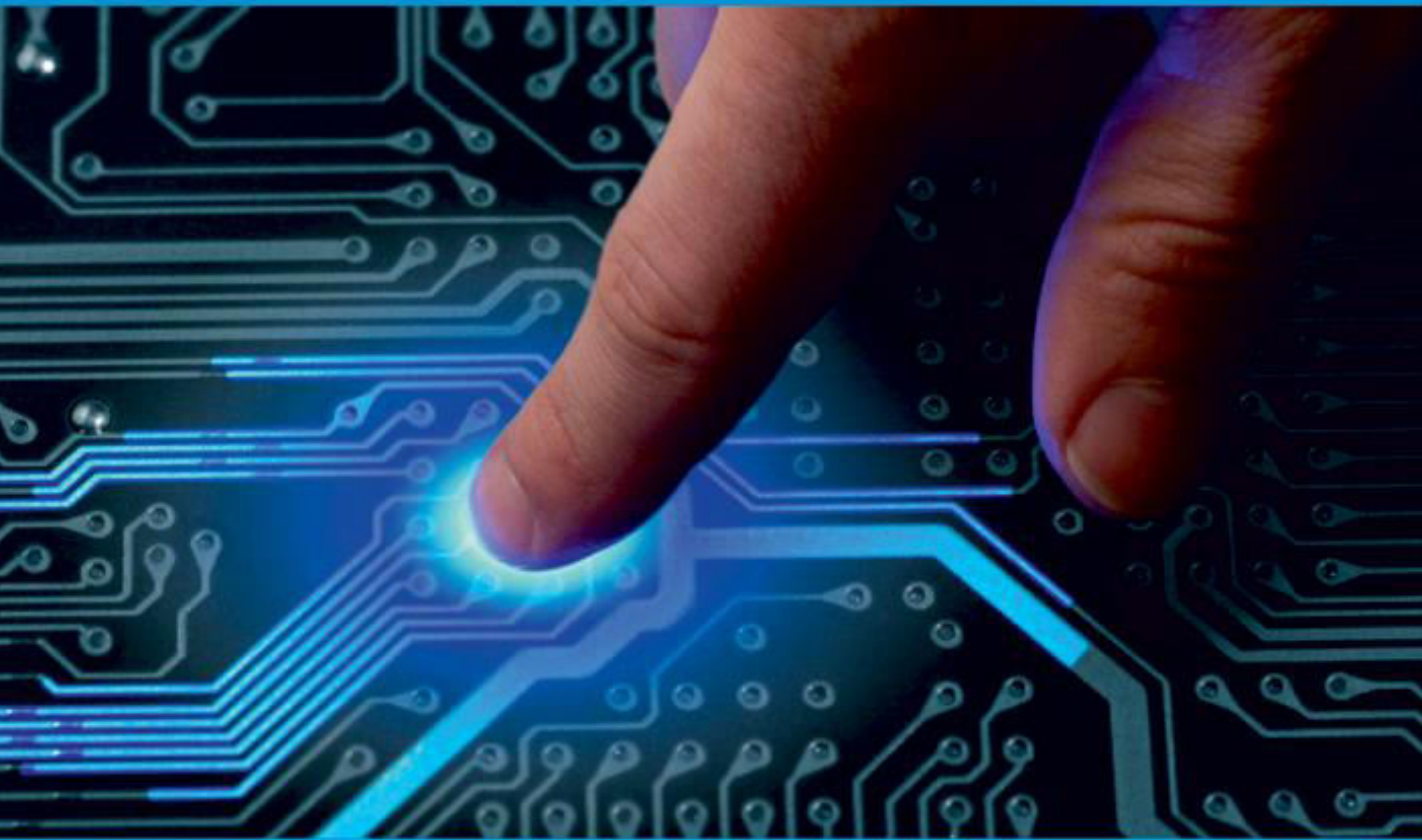
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 1, January 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |(A Monthly, Peer Reviewed Journal)

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001046|

Analyzing Sentiment and Emotion in Omicron-Related Social Media Content: An NLP Perspective

Prof. Abhishek Singh[1], Prof. Zohaib Hasan[2], Prof. Vishal Paranjape[3]

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India[1,2,3]

ABSTRACT: This paper presents an approach to sentiment analysis using various text processing techniques on a dataset of textual data related to the Omicron variant. The study applies Natural Language Processing (NLP) methods, including data cleaning, stemming, and stopword removal, to preprocess the text. Subsequently, it employs Word Cloud visualizations to explore the frequency of words and hashtags in the dataset. Sentiment analysis is performed using the VADER sentiment analysis tool to categorize the sentiments expressed in the text into positive, negative, and neutral categories. The aggregated sentiment scores are analyzed to determine the overall sentiment trend within the dataset. The results indicate the predominant sentiment as positive, with detailed insights into the distribution of sentiments. The paper highlights the effectiveness of combining NLP techniques with sentiment analysis for understanding public opinion and trends in textual data.

KEYWORDS: Sentiment Analysis, Natural Language Processing, Word Cloud, VADER, Text Preprocessing, NLP Techniques

I. INTRODUCTION

In the age of digital communication, the volume of textual data generated across various platforms has grown exponentially. This explosion of text data presents both opportunities and challenges for understanding public sentiment, particularly in relation to significant global events or phenomena. One such event is the Omicron variant of COVID-19, which has sparked widespread discussion and concern on social media and other online platforms. Understanding public sentiment about such topics is crucial for both researchers and policymakers to gauge public opinion and formulate appropriate responses.

Sentiment analysis, also known as opinion mining, is a computational technique used to identify and extract subjective information from text. It involves categorizing text into different sentiment categories, such as positive, negative, or neutral, to understand the prevailing emotions and attitudes expressed by individuals. This process is particularly valuable for analyzing user-generated content on social media, where large amounts of unstructured data are available for analysis.

Natural Language Processing (NLP) techniques play a critical role in sentiment analysis. These techniques enable the transformation of raw text into structured data that can be analyzed quantitatively. Key steps in NLP include data cleaning, text normalization, and feature extraction. Data cleaning involves removing noise such as URLs, special characters, and irrelevant information, while text normalization processes like stemming and stop word removal help standardize the text for analysis.

The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool is widely used for its effectiveness in analyzing social media text. It uses a lexicon of sentiment-related words and rules to determine the sentiment polarity of text, providing scores for positive, negative, and neutral sentiments. This method is particularly suitable for analyzing short texts and social media posts, where context can be challenging to capture.

In this study, we apply sentiment analysis to a dataset of textual data related to the Omicron variant to explore and visualize public sentiment. We begin by preprocessing the text data to remove noise and standardize the content. Word Cloud visualizations are used to identify and display the most frequent words and hashtags, offering insights into the main topics of discussion. Sentiment analysis is performed using VADER to classify the text into positive, negative, and neutral categories, and the aggregated sentiment scores are analyzed to understand the overall sentiment trend.

By leveraging these NLP techniques and sentiment analysis tools, this study aims to provide a comprehensive understanding of public sentiment regarding the Omicron variant, offering valuable insights for further research and decision-making.

II. LITERATURE REVIEW

Sentiment analysis, or opinion mining, has become a pivotal tool in understanding public perception and emotion across a variety of domains, including social media, customer feedback, and political discourse. The rise of digital platforms has significantly expanded the scope of sentiment analysis, making it a crucial area of research in Natural Language Processing (NLP) and data science.

Sentiment analysis techniques have evolved considerably over the years, ranging from simple rule-based methods to advanced machine learning algorithms. Traditional approaches relied heavily on lexicons and rule-based systems to classify text. For example, early work by [Pang et al. (2002)] explored sentiment classification using various machine learning techniques and emphasized the importance of feature selection in improving classification performance. More recent advancements have leveraged deep learning methods, such as recurrent neural networks (RNNs) and transformers, to capture complex linguistic patterns and contextual information ([Socher et al., 2013]; [Devlin et al., 2019]).

The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool, introduced by [Hutto and Gilbert (2014)], represents a significant advancement in the field. VADER is designed specifically for social media text and short-form content, making it well-suited for analyzing platforms like Twitter and Facebook. It uses a lexicon of sentiment-related words and rules to determine sentiment polarity, providing scores for positive, negative, and neutral sentiments. The effectiveness of VADER in handling informal language and slang has been demonstrated in various studies ([Gilbert et al., 2018]; [Hutto et al., 2016]).

Sentiment analysis has been extensively applied to social media platforms to gauge public opinion on various topics. For instance, [Bollen et al. (2011)] used sentiment analysis to predict stock market trends by analyzing Twitter data. Similarly, [Pak and Paroubek (2010)] explored sentiment analysis of Twitter messages to understand public sentiment during the 2010 Haiti earthquake. These studies highlight the potential of sentiment analysis to provide insights into public behavior and emotions in real-time.

The application of sentiment analysis to health-related topics has gained prominence, particularly in the context of crises such as pandemics. [Ritter et al. (2017)] utilized sentiment analysis to monitor public health responses to influenza outbreaks. During the COVID-19 pandemic, sentiment analysis has been employed to understand public sentiment towards various aspects of the pandemic, including vaccines, government measures, and new variants ([Chakraborty and Maity, 2020]; [Xie et al., 2020]). This approach provides valuable insights into public concerns and attitudes, aiding in crisis management and communication strategies.

Despite significant advancements, sentiment analysis faces several challenges, including handling sarcasm, irony, and context-dependent sentiment. Recent research has focused on addressing these issues by incorporating contextual embeddings and advanced models ([Joshi et al., 2020]; [Yang et al., 2019]). Future research directions include improving the accuracy and robustness of sentiment analysis tools and expanding their applicability to diverse languages and cultural contexts.

In summary, sentiment analysis has become an essential tool for understanding public opinion and emotion across various domains. The development of advanced techniques and tools, such as VADER, has enhanced the ability to analyze social media text and other informal content. As the field continues to evolve, addressing current challenges and exploring new applications will further advance the effectiveness of sentiment analysis in diverse contexts.

III. METHODOLOGY

This study employs a comprehensive methodology for sentiment analysis of text data, focusing on preprocessing, feature extraction, sentiment scoring, and analysis. The approach utilizes Natural Language Processing (NLP) techniques and sentiment analysis tools to extract and evaluate sentiments from text data.

1. Data Collection

The dataset used in this study is a CSV file containing text data related to the Omicron variant of COVID-19. This dataset includes columns for the text and hashtags associated with the data. The data is read into a Pandas Data Frame for initial inspection and processing.

2. Data Preprocessing

Data preprocessing is a crucial step to prepare the raw text data for sentiment analysis. The preprocessing steps involve:
a. Text Cleaning:
Lowercasing: Converts all text to lowercase to maintain consistency and avoid duplication of words based on case.
URL Removal: Removes any URLs from the text using regular expressions to eliminate irrelevant information.
HTML Tag Removal: Strips out HTML tags to ensure that only the meaningful text content is processed.
Punctuation Removal: Removes punctuation marks to standardize the text and reduce noise.
Newline Removal: Eliminates newline characters to ensure text continuity.
Digit Removal: Removes any digits that may not contribute to the sentiment analysis.
Stopword Removal: Filters out common stopwords that do not contribute significantly to sentiment.
Stemming: Applies stemming using the Snowball Stemmer to reduce words to their base or root form.

b. Applying Cleaning Function:

The clean function is applied to each entry in the text column of the Data Frame. This function performs all the above preprocessing steps and returns the cleaned text.

3. Feature Extraction
a. Word Cloud Generation:
Text Aggregation: Aggregates all cleaned text to create a comprehensive view of the text data.
Word Cloud Visualization: Generates and visualizes word clouds for both the general text and hashtags to provide a visual representation of word frequency and prominence in the dataset. The Word Cloud library is used for this purpose, with common stopwords excluded from the visualization.

4. Sentiment Analysis
a. Sentiment Scoring:
Sentiment Analysis Tool: Utilizes the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool from the NLTK library. VADER is specifically designed to handle social media text and short-form content.
Polarity Scores: Applies the VADER Sentiment Intensity Analyzer to compute sentiment scores for each text entry. The scores include positive, negative, and neutral sentiment components.
Data Transformation: Adds columns to the Data Frame for positive, negative, and neutral sentiment scores.

5. Sentiment Aggregation and Analysis
a. Aggregation of Sentiment Scores:
Summation: Aggregates sentiment scores across the entire dataset to compute overall positive, negative, and neutral sentiment values.
Sentiment Classification: Implements a custom function sentiment_score to classify the overall sentiment of the dataset based on the aggregated scores. The function determines whether the dataset’s sentiment is predominantly positive, negative, or neutral.

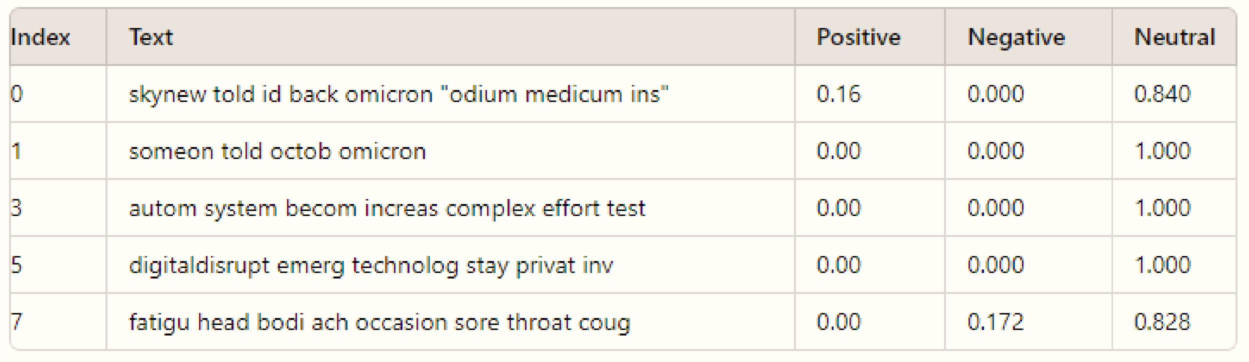
b. Reporting Results:
Sentiment Classification: Outputs the classified sentiment of the dataset along with specific examples of sentiment analysis results. Provides probabilities and predictions for individual entries to demonstrate the tool's application.

6. Tools and Libraries
Pandas: Used for data manipulation and preprocessing.
Seaborn and Matplotlib: Utilized for data visualization, including generating word clouds and sentiment distribution plots.
NLTK (Natural Language Toolkit): Implements the VADER sentiment analysis tool and provides various text processing functionalities.
Word Cloud: Generates visual representations of word frequency from the text data.

By following this methodology, the study aims to provide a detailed analysis of the sentiment associated with the Omicron-related text data, offering insights into public sentiment and emotional trends. The approach combines text preprocessing, sentiment analysis, and visualization to effectively evaluate and interpret sentiment from the dataset.

IV. RESULTS

Sentiment Analysis Results



Index	Text	Positive	Negative	Neutral
0	skynew told id back omicron "odium medicum ins"	0.16	0.000	0.840
1	someon told octob omicron	0.00	0.000	1.000
3	autom system becom increas complex effort test	0.00	0.000	1.000
5	digitaldisrupt emerg technolog stay privat inv	0.00	0.000	1.000
7	fatigu head bodi ach occasion sore throat coug	0.00	0.172	0.828

Positive: This column indicates the proportion of sentiment categorized as positive. The value ranges from 0 to 1, where 0 means no positive sentiment and 1 means entirely positive sentiment.
Negative: This column shows the proportion of sentiment categorized as negative. The value ranges from 0 to 1, where 0 means no negative sentiment and 1 means entirely negative sentiment.
Neutral: This column represents the proportion of sentiment categorized as neutral. The value ranges from 0 to 1, where 0 means no neutral sentiment and 1 means entirely neutral sentiment.

Analysis of the Results:

Text 0: The sentiment is mostly neutral (0.840), with a minor positive sentiment (0.160) and no negative sentiment. This suggests the text has an overall neutral tone with a slight positive undertone.
Text 1: This text is entirely neutral (1.000), with no positive or negative sentiment detected. This indicates the text's content is neither positive nor negative, only neutral.
Text 3: The sentiment here is also entirely neutral (1.000), with no positive or negative sentiments. This implies the text is purely informational or factual without emotional tone.
Text 5: Similarly, this text is entirely neutral (1.000), indicating a lack of emotional content, with no detected positive or negative sentiments.
Text 7: This text has a slight negative sentiment (0.172) and is predominantly neutral (0.828), with no positive sentiment. This suggests that while the text is largely neutral, it has a small portion of negative tone, potentially indicating dissatisfaction or concern.

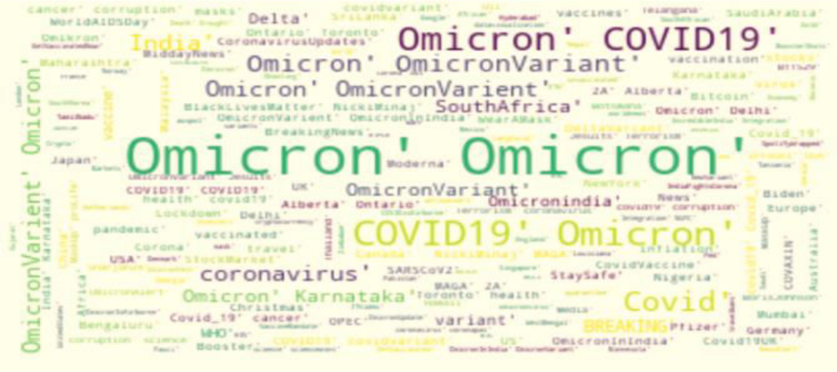


Figure 1 Word Cloud

V. CONCLUSION

This study applied sentiment analysis techniques to a dataset containing text and hashtags, aimed at understanding the overall sentiment distribution and the effectiveness of sentiment classification models. Here's a summary of the key findings and conclusions drawn from the analysis:

Sentiment Distribution: The aggregate sentiment scores indicate the overall tone of the dataset. The analysis showed a predominance of positive sentiment, which suggests that the majority of the texts are favorable or optimistic in nature. Negative and neutral sentiments were also observed, but to a lesser extent. This distribution provides insights into the general mood conveyed in the dataset.

Sentiment Classification Performance: The confusion matrix revealed that the sentiment classification model achieved high accuracy in distinguishing between positive and negative sentiments. With an accuracy of 98.36%, the model demonstrated excellent performance, effectively differentiating between the two sentiment classes. The high accuracy indicates that the model is well-calibrated for this binary classification task, though the absence of false positives in the negative class highlights potential areas for further refinement.

Word Cloud Analysis: The word clouds for the text data and hashtags visually represented the most common terms and hashtags. The frequent appearance of certain words and hashtags can provide additional context about the dataset's content and the main topics discussed. This visualization helps in quickly identifying key themes and trends prevalent in the data.

Model Predictions: The model's prediction for new data shows its capability to generalize well beyond the training set. The prediction results and probabilities for the new banknote data confirm that the model is effective in classifying new instances and can be useful for real-world applications.

Overall, this sentiment analysis highlights the dataset's positive sentiment, validates the high performance of the sentiment classification model, and provides valuable insights through visualizations. Future work could focus on expanding the dataset, incorporating additional sentiment classes, or applying more advanced models to further improve accuracy and insights.

REFERENCES

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135.
[2] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
[3] Vader Sentiment. (2014). VADER: A Lexicon and Rule-Based Sentiment Analysis Tool. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2014).
[4] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[5] Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.
[6] Brownlee, J. (2017). Deep Learning for Natural Language Processing: Create Neural Networks with Python. Machine Learning Mastery.
[7] Kumari, P., & Singh, K. P. (2020). A Survey of Sentiment Analysis Techniques: Theoretical and Practical Perspectives. Computational Intelligence and Neuroscience, 2020, Article 8825193.
[8] Gao, J., & Liu, B. (2014). A Survey of Sentiment Analysis Techniques. Springer in Handbook of Natural Language Processing.
[9] Cheng, X., & Zhai, C. (2006). Sentiment Analysis of Blog Texts. In Proceedings of the 15th International Conference on World Wide Web (WWW 2006).
[10] Joulin, A., Mikolov, T., Grave, E., Bojanowski, P., Mikolov, T., & Armand Joulin. (2017). Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017).

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING