International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Comprehensive Sentiment Profiling and URL Mining of WhatsApp Conversations using VADER and NLP Techniques

D.Madhavi Latha[1], P.S.Pooja Sree[2], A.Gayatri[3], N.Raju Nayak[4]

[1]*Assistant Professor, Dept of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology,Telangana,India*

[2,3,4]*Students, Dept of Electronics and Communication Engineering, Geethanjali College of Engineering and Technology, Telangana, India*

Abstract - *With over 487.5 million active addicts, WhatsApp is one of the most popular platforms for informal discussion. The point is estimated to take a new communication every 0.5 seconds, denoting it produces a large knob of user- generated content every single day. This paper focuses on the performing sentiment analysis and URL extraction from WhatsApp conversation using Natural Language Processing (NLP) algorithms like VADER (Valence Aware Dictionary and sEntiment Reasoner). The methodology consists of pre-processing conversation data which includes drawing up converse data and organizing the users sentiments into positive, neutral, or negative tags, along with orderly recovery of participated links. URL extraction processes and pattern matching enabled with regular expressions alongside VADER's preset formulas for scoring sentiments yield precious perceptivity into user exertion, colorful analyses indicated positive connections between content participated externally and the dominating sentiments which redounded in links being participated in the more positive exchanges. Using sentiment pie maps alongside URL frequence graphs allows for farther understanding of the data. These approaches add onto hypotheses concerning stoner actions analytics and can be applied to advancement in covering systems for real- time usages in digital marketing and social media monitoring.*

Key Words: Sentiment Analysis, WhatsApp Chat Analysis, Natural Language Processing (NLP), VADER Tool, Python Programming, URL Extraction, Data Visualization, Text Mining, Regular Expression.

1.INTRODUCTION

Sentiment analysis is one of the branches of Natural Language Processing (NLP) that is actively evolving because it works towards obtaining an opinion or sentiment from a given text. As new communication mediums like WhatsApp emerge, the need to analyze user sentiments in conversations becomes important for understanding social relationships, feedbacks, and even personal sentiments. Python, along with VADER (Valence Aware Dictionary and sEntiment Reasoner), has incorporated sentiment analysis features which improves processing speed with large datasets. In this research, the analysis of sentiment in WhatsApp chat data using NLP libraries in Python and VADER while also building a module for extracting links to find the most shared URLs.

In this paper the development of an automated system to analyze the emotion underlying WhatsApp conversations and derive the useful conclusions as pertaining to the URLs contained within the messages. Due to the amount of data shared over messaging applications, deriving the sentiment captured in these messages could be greatly useful for certain organizations, individuals, and researchers. With sentiment analysis, it is now possible to classify interactions as positive, negative, or neutral which helps in assessing emotional expressions especially in written communications.

In addition, extracting URLs from these conversations enables analysis of the resources commonly shared among them, which can give further context to the users' sentiment. This pairing of sentiment analysis and URL extraction gives a complete methodology for analyzing WhatsApp conversations.

2. EXISTING SYSTEMS

Sentiment analysis has also received a lot of attention over the last few years with applications in social media monitoring, customer feedback analysis, and product review systems. A variety of methods for sentiment analysis have been explored in previous research, especially techniques that apply to informal text and chat-based communication.

The earlier sentiment analysis approaches were dependent a lot on classic machine learning models like Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression. These approaches demanded large amounts of labeled data and used feature extraction methods like bag-of-words and n-grams to transform textual data [1]. Yet, these models tended to be weak at processing informal and unstructured text like chat data since they were poor at handling context and subtle language.

As technology improved with the help of artificial intelligence, deep learning-based approaches were developed using models like Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) [2]. These showed better performance but were computationally intensive and demanded huge datasets, which are not always available when working with chat data.

As a supplement, lexicon-based methods have been increasingly popular, especially with informal short-text. VADER (Valence Aware Dictionary and sEntiment Reasoner) has been particularly effective at dealing with social media parlance, emoticons, acronyms, and colloquials. Created by Hutto and Gilbert [3], VADER integrates lexical heuristics and a sentiment

dictionary to provide polarity scores for text and therefore is well-tuned to analyze WhatsApp chat communications.

In addition to sentiment analysis, URL extraction has been a critical focus of research with the ubiquity of link sharing in online discourse. URL extraction methods often employ regular expressions or machine learning to find and identify URLs in text [4]. By examining frequency and content of shared URLs, trending topics and the kinds of resources impacting discourse dynamics may be identified. Although there has been progress in sentiment analysis and URL extraction, current systems are likely to handle these two as distinct tasks. The integration of sentiment scoring and URL analysis for chat environments such as WhatsApp has not been explored much.

3. PROPOSED SYSTEM

To fill this void, the system proposed in this paper combines sentiment analysis with the VADER tool and URL extraction methodologies to achieve a holistic analysis of WhatsApp chat data. This method allows for parallel examination of the emotional content and the exchanged content in messages.

VADER is chosen based on its successful track record in analyzing informal text with high accuracy and low resource utilization. It performs well in the nature of WhatsApp chats, such as using slang, abbreviations, emojis, and different styles of punctuation. This qualifies it as the best option for sentiment analysis within the context of a messaging application. For URL extraction, regular expressions are used to find and extract links that are embedded in messages. After extraction, the URLs are analyzed for frequency, allowing for insights into the most shared domains or resources in the conversation. By integrating sentiment analysis with URL frequency analysis, this system offers a two-layered understanding of WhatsApp chats:

- Sentiment insights provide the emotional context of the conversation.
- URL analysis identifies the external resources exchanged and possibly shaping these sentiments.

This combined approach offers a new way of analyzing conversational data, providing deeper insights than individual sentiment analysis or URL extraction. The system lays the groundwork for future improvements in the comprehension of digital communication and is a valuable addition to the study of natural language processing and social media analysis.

This integrated method presents a novel approach to conversational data analysis, offering more meaningful insights than sentiment analysis or URL extraction alone. The system sets the foundation for future enhancements in understanding digital communication, making it a valuable contribution to the field of natural language processing and social media analytics.

4. METHODOLOGY

The methodology employed in this exploration is designed to efficiently break down WhatsApp conversation data for sentiment analysis and URL birth. The process can be divided into several stages data collection, data pre-processing, sentiment analysis, URL birth, and affect analysis. Each stage is described in detail below.

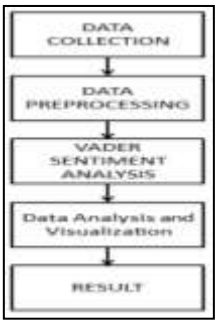


Fig: Functional Flow Diagram of the Proposed System

4.1 **Data Collection :** The dataset for this exploration consists of WhatsApp conversation logs. These conversation logs are exported from the WhatsApp usage in the form of manual lines(. txt). Each manual line includes people communications, timestamps, and other metadata. The dataset can either be collected from real- world people relations or faked data, depending on the range and ethical considerations. For the purpose of this study, the data used was anonymized to assure isolation and confidentiality.

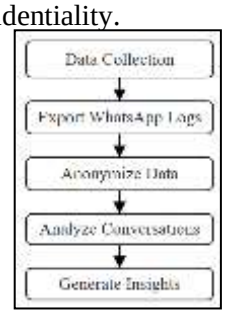


Fig 4.1 : Data Collection

4.2 **Data Preprocessing :** Before WhatsApp conversation text data is fully versatile for meaningful analysis, it needs to be converted from potential raw text into usable corpora by processing through the steps listed below.

- ***Text Cleaning:*** All unwanted characters, special symbols, and unwanted metadata (timestamps and usernames) are removed from the text while we keep emojis and emoticons because they can be very informative to sentiment.
- ***Tokenization:*** The cleaned text is tokenized into smaller bits of text like words or phrases from which to conduct text analysis which must happen before we can analyze anything quantitatively or

qualitatively. Tokenization is valuable because it allows us to analyze every token from the text.

- ***Lower-Casing:*** The text is lower-cased into lower-case text; lower-casing has to occur in order to maintain standardization, and we don't want to confuse terms that differ only by case (i.e. "happy" and "Happy").
- ***Stopword Removal:*** Stopwords are common words which are less meaningful, e.g., "the", "is", and "and" or "so". We want to remove stop-words so that we are removing noise in the data.
- ***Lemmatization:*** The words are changed to base or dictionary terms, e.g., change "running" to "run"; we just want to be more general for ease in further analysis.

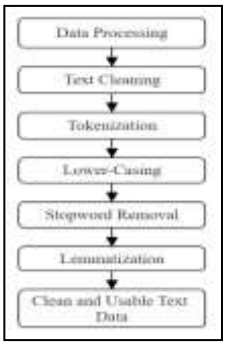


Fig 4.2 : Data Preprocessing

All of these processing procedures will allow for the text data to be clean, coherent, and in exploitable format that can be analyzed for sentiment analysis and URL identification/extraction.

4.3 **Sentiment Analysis Using VADER:** In this process of sentiment analysis, We used VADER (Valence Aware Dictionary and sEntiment Reasoner) to process and analyze social media and informal text. VADER is a rule-based model that uses a sentiment lexicon to assign polarity scores to textual data, and is therefore perfect for processing WhatsApp chat messages. The process for conducting a sentiment analysis is as follows:

- ***Text Input:*** The text for the sentiment analyzer is the pre-processed WhatsApp chat data.
- ***Sentiment Scores:*** VADER assigns a compound sentiment score to each chat message between -1 (most negative) to +1 (most positive). Messages that have a score close to 0 are considered neutral.
- ***Classification:*** Based on the computed scores, each message can be classified as positive, negative, or neutral. Each of these classifications have thresholds applied so the sentiments scores are correctly mapped to each classification.
- ***Visualization:*** The final results of the sentiment analysis were visualized in both bar charts and pie charts. These graphics serve as an easy to interpret visual representation of the sentiment analysis dataset. The process allows for a systematic and interpretable way to analyze the emotional tone in WhatsApp conversations, thereby better understanding user sentiment.

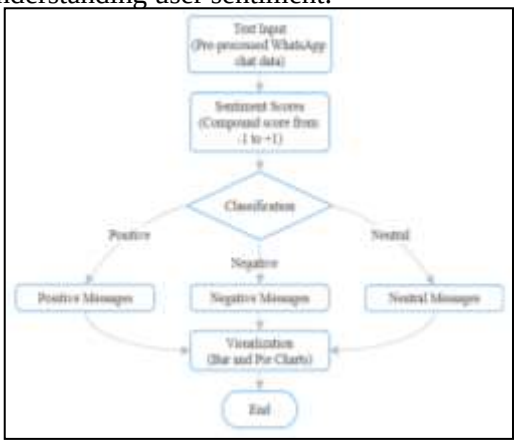


Fig 4.3: Sentiment Analysis Using VADER

4.3 **URL Extraction** ***:*** A URL extraction module is incorporated in the system to identify and evaluate URLs exchanged throughout WhatsApp chat conversations. The extraction module has a process to extract, filter, and interpret URLs that have been exchanged for further analysis:

- ***Pattern Recognition:*** URLs are extracted from the chat messages using regular expressions. Regular expressions are prepared to extract the standard web address patterns, i.e. those that start with "http://" or "https://".
- ***URL Filtering:*** Duplicates will be filtered out along with non-useable and non-existent URLs, resulting in the accurate URLs.
- ***Frequency Analysis:*** The list of URLs will be scanned for the frequency of every unique URL, when available. This will highlight the most commonly exchanged resources, along with identifying recurring topics or references made throughout the conversation.
- ***Categorization:*** URLs could optionally be categorized by purpose (e.g. media outlet, social media, e-commerce, etc.) to help with identification of the general character and purpose of the exchanged URL post.

Fig 4.4: *URL Extraction*

This phase of analysis complements any sentiment analysis completed as it presents some context to the information propelling the content of conversations, as well as allowing you to see a two dimensional view of emotional tone and content exchanged.

4.4 **Data Analysis and Visualization :** After sentiment analysis and URL extraction, the outcomes are analyzed in depth and visualized to extract useful insights from the WhatsApp chat dataset. The following methods are used:

- ***Sentiment Distribution:*** Histograms and pie charts are created to show the distribution of sentiment scores among the messages.
- ***URL Frequency:*** The most shared URLs across the chat data are represented using bar charts. The bar charts assist in determining trending topics or frequently used resources among the users.
- ***Correlation Analysis:*** Where possible, a correlation analysis is performed to determine whether there are any relationships between message sentiment and the nature or category of forwarded URLs.

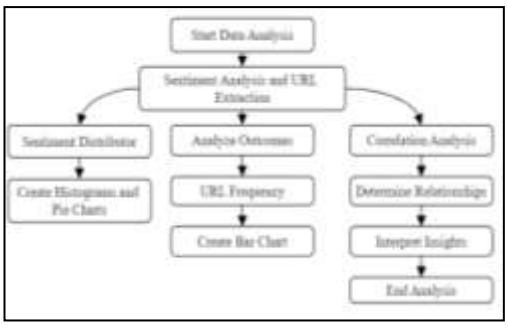


Fig 4.5: Data Analysis and Visualization

These visualizations and analysis methods help in interpreting raw data into usable insights, enriching the understanding of both emotional expression and content sharing behavior within WhatsApp conversations.

4.5 Tools and Technologies : The suggested approach is based on a range of tools and technologies that support different activities such as text pre-processing, sentiment analysis, URL extraction, and data visualization. The key technologies are:

- ***Python:*** It is one of the most uncomplicated programming languages and it has numerous libraries for natural language processing and data analysis.
- ***VADER Sentiment Analysis:*** VADER is a software tool that is a Python package used for sentiment analysis of informal texts (e.g. chat messages in WhatsApp).
- ***Natural Language Toolkit (NLTK):*** NLTK is utilized to conduct pre-processing tasks such as tokenization, stopword removal, and lemmatization.
- ***Regular Expressions:*** the built-in re module in Python is utilized to define patterns to extract URLs from the chat messages.
- ***Matplotlib and Seaborn:*** These visualization libraries are used to create visual representations of the analyses results, for example, sentiment distributions, and URL frequency charts.

In summary, these tools and technologies may help the fast building and operation of the system as they allow nondisruptive accurate processing, analysis, and visualization of the chat data.

5. RESULTS

By applying the suggested system on the WhatsApp chat dataset, we gained new insights about sentiment classification, URL frequency analysis, and participant behavior monitoring. To help with understanding the user behavior, the commonly shared content, emotional tone, and any changes in their chats over time, a variety of visualizations were created based on the processed data. The figures that we are shown below represent the main things we learned from the processed data which were both statistical and behavioral learning's of the conversations.

Fig 5.1: *Obtaining a data frame mapping Messages and their Sentiments.*

Fig 5.2: Obtaining the most messaged person and most used Emoji's from the Chat

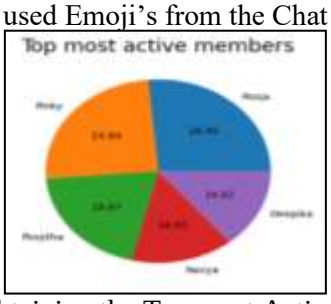


Fig 5.3: Obtaining the Top most Active Members

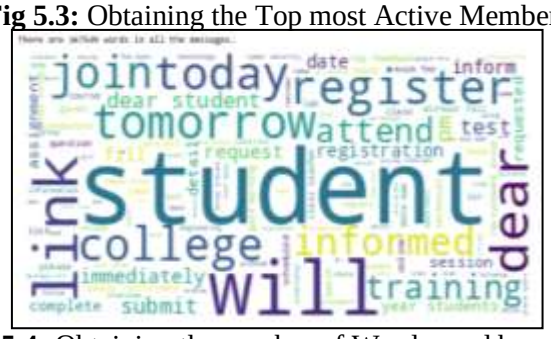
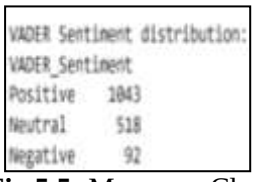


Fig 5.4: Obtaining the number of Words used based on their frequency of usage



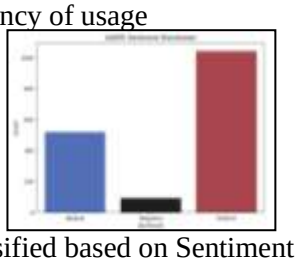


Fig 5.5: Messages Classified based on Sentiment

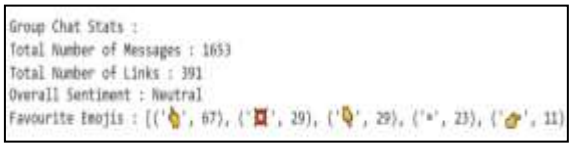


Fig 5.6: *Obtaining the group chat statistics of Data set.*

Fig5.7 : Analysis of most messages on the Day and Year

Fig 5.8: Analysis of Time when chat was highly Active

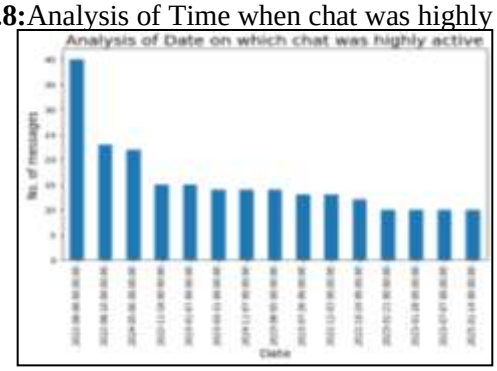


Fig 5.9: Analysis of Date when chat was highly Active

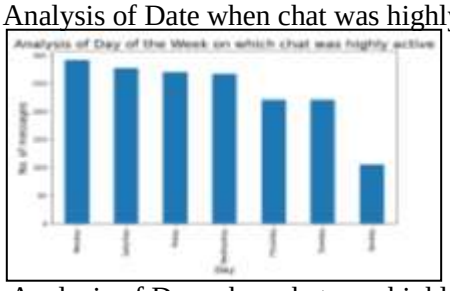


Fig5.10 : Analysis of Day when chat was highly Active

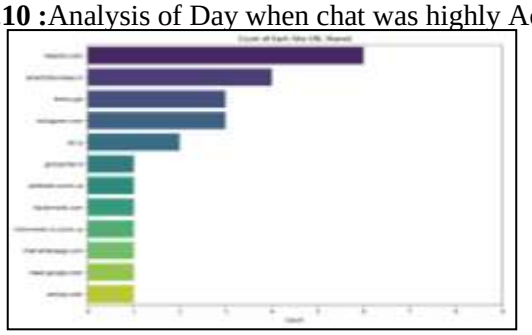


Fig :5.11 : Obtaining the number of URLs used based on their frequency of usage

6. APPLICATIONS

- ***Sentiment Analysis in Social Media:*** Sentiment analysis on social media is good for market research, monitoring public opinion, or brand management.
- ***WhatsApp Chat Sentiment:*** Chat sentiment can provide personalized customer support, analyze feedback, and optimize user engagement, among a large number of other uses.
- ***URL Extraction:*** Extracting URLs has many potential applications in data mining, identifying trends, and analyzing links being shared across social media.
- ***NLP in Domains:*** NLP can be applied to customer reviews, opinion mining, social media analytics, and automated content analysis, just to name a few examples.

7. BENEFITS

- ***Better Knowledge about Customers:*** Allows companies to respond rapidly to customer sentiment and feedback as it arises.
- ***Better Decision Making:*** Companies can use sentiment-related trends to make better choices in their marketing, product development, and support.
- ***Automation:*** Eliminates the time required for manual sentiment analysis, and gives a scalable approach to handling large amounts of text data.
- ***Enhanced User Engagement:*** Enhanced understanding of customer engagement can lead to more reliable engagement strategies from the organization based on real-time analysis and assessment of feedback.

8. CONCLUSION

A complete methodology was established in order to study WhatsApp chat data that integrated sentiment analysis utilizing the VADER software with URL extraction operations. From the utilization of Natural Language Processing (NLP) processes (text cleaning, tokenization, lemmatization, and the removal of stopwords), the initial chat data was formatted into a form ready to be processed and analyzed.

The sentiment analysis concluded that the messages were positive, negative, or neutral as appropriate for determining the emotional tone of the conversations. Concurrently, the URL extraction discovered the URLs and gave analysis of the most frequently shared URLs for an identification of the kind of content viewed by the users during the conversations. Collectively, both methods provided a two-layer emotional stance on the basis of sentiment of responses and a behavior measure based on data exchanged sent and received as relative to what was initially shared.

The graphics, such as the bar chart and pie chart, made the information more readable that illustrated user's activity, user's number of messages, and any sentiment trend that seems to be present over time. Such analysis can be especially useful in determining user intent and activity in such scenarios as customer service, online community navigation, and marketing analysis. Upcoming studies can look at the rise of improved sentiment detection accuracy by using context models like transformer models.

REFERENCES

1. Yadollahi A, Shahraki AG, Zaiane OR "Current state of text sentiment analysis from opinion to emotion mining", vol 5, no(7),pp.850-858,2017.
2. Lui B "Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol" vol 5 no(1),pp.1–167, 2012.".
3. Casillo M, Clarizia F, D'Aniello G, De Santo M, Lombardi M, Santaniello D "Chat-bot: a cultural heritage-aware teller-bot for supporting touristic experiences. Pattern Recognition" vol 1 no(131), pp. 234-243, 2020.
4. D'Aniello G, Gaeta M, Orciuoli F, Sansonetti G, Sorgente F "Knowledge-based smart city service system" vol 9, no(6), pp.1–22, 2020.
5. Rosenthal S, Farra N, Nakov P SemEval-2017 task 4:" Sentiment analysis on Twitter. In: Proceedings of the 11th international workshop on semantic evaluation" pp. 502-518, 2018.
6. Wang H, Castanon JA "Sentiment expression via emoticons on social media. In: 2015 IEEE international conference on big data (Big Data)", pp 2404–2408, 2015.
7. Tubishat M, Idris N, Abushariah M "Explicit aspects extraction in sentiment analysis using optimal rules combination", no(114), pp. 448–480, 2021.
8. Farhadloo M, Rolland E "Fundamentals of sentiment analysis and its applications." Springer, Cham, vol 6, pp 1–24, 2016.
9. Novielli N, Calefato F, Lanubile F "Love, joy, anger, sad, fear, and surprise: Se needs special kinds of ai: a case study on text mining and se." vol 3 no(37) pp. 86–91, 2020.
10. Munezero M, Montero C, Sutinen E, Pajunen J "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text" vol 5 no(02) pp.101–111, 2014.