# Sentiment Classification of WhatsApp Data Context and Topic Analysis Using R Tool

**Pavithra Poovalingam[1], J.Melba Rosalind [2]**

*[1]Department of Computer Science, Lady Doak College, Madurai Kamaraj University*
*Madurai, TamilNadu, India*

*[2]Department of Computer Science, Lady Doak College, Madurai Kamaraj University*
*Madurai, TamilNadu, India*

***Abstract-*** The medium of communication varies periodically due to the evolution of various social media applications. WhatsApp is one such application that is used for transferring variety of information in the form of text, audio, video and other compatible files. Nowadays, Schools and Colleges also make use of WhatsApp as the primary means of Communication. This project entitled, "Sentiment Classification Of WhatsApp Data Context And Topic Analysis Using R Tool" aims to analyze the WhatsApp group chat conversations of students and used to measure the effectiveness of WhatsApp for academic and non-academic purposes by categorizing the contexts according to the major topics. It is also used for measuring the polarity of each WhatsApp chat on their topic of interest.

Student's WhatsApp group chat conversations of Department of Computer Science, Lady Doak College is used for the analysis. This project will be really helpful to the department faculty for better understanding the Student's topic of interest and the polarity of their sentiments. For student community, it reveals the context distribution among them on WhatsApp group chat conversations.

***Keywords*** -WhatsApp analysis, Topic modeling, Sentiment classification, LDA Algorithm, Unsupervised learning.

## I. INTRODUCTION

### *A. Sentiment Analysis*

Sentiment Analysis is the process of computationally identifying and categorizing the different opinions expressed in a piece of text, especially in order to determine person's interest and attitude towards a given topic. The attitude may be a judgment or evaluation, affective state of the person or the intended emotional communication.

It refers to the use of text analysis, computational linguistics, natural language processing and biometrics to systematically identify, extract, quantify and study affective states and subjective information. It is used for classifying and categorizing the text based on the polarity, sentiments and by context.

Usage of social media such as blogs and social networks are rising and it has fueled interest in sentiment analysis. By analyzing one's social media data, we can easily make desirable decisions about their region of interest.

### *B. Challenges in Sentiment Analysis*

There are many challenges in sentiment analysis techniques that may affect accuracy of the developed system.

- Usage of Bi-lingual languages increases complexity to the data abstracted from the source and it becomes difficult for us analyze the sentiments in such contexts.
- Many of the textual data contains misspelled words, spelling errors, irregular typography and ungrammatical sentences.
- Possibility of sarcastic expressions in the textual data.[1]
- Multiple sentiments could be expressed in one sentence/document.
- Users prefer to use short form of words in the sentences which again lead to transliteration problems.

## II. SYSTEM ARCHITECHTURE

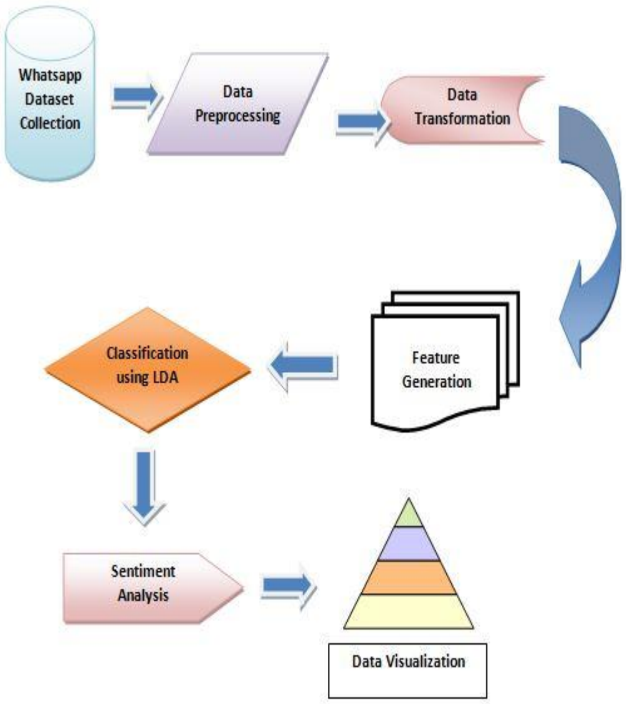

**Fig.1 System architecture for the proposed system.**

## III. DATA COLLECTION AND PREPROCESSING

Fig.1 shows the process flow involved in the proposed system starting from Dataset Collection, followed by generation of features and finally ends through Visualization. Data Collection [2] is the primary task of any data mining process. Scope of the project is limited to the computer science department students of Lady Doak College, particularly the Post graduate students. Their group chat data is collected from their respective class representatives.

Dataset is collected by exporting the WhatsApp group chat conversations from the students. Since, the project mainly focuses on the textual data, audio and video media files can be ignored while exporting the text file.

Data preprocessing [3] is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

In this study, data preprocessing steps are carried out to remove unnecessary punctuation marks, stop words, whitespaces and tab spaces. It makes the project scope to be more specific to the particular area of research. Without Data Preprocessing, the system will not be robust and is easily prone to complex problematic results.

Data Transformation is a necessary step in data mining that converts data or information from one format to other. Based on the need of the project, data transformation can include a range of activities such as converting data types, cleansing data by removing nulls or duplicate data, enriching the data, or performing aggregations and so on.

## IV. FEATURE GENERATION

Feature generation is also known as feature construction, feature extraction or feature engineering. It focuses on the construction of features from raw data, creating a mapping to convert original features to new features, creating new features from one or multiple features

Two goals of feature generation can be dimensionality reduction and accuracy improvement. When the goal of a feature generation method is dimensionality reduction, then the result will be a feature space which contains less features than the original feature space. However, when the goal is accuracy improvement, the resulting feature space will most likely contain more features than the original feature space [4].
There is always a need to remove dimensions/features that are not of the project's interest. One such example is trying to remove 'Thanglish[5]' words and other short form of words from the dataset. In the proposed system, these non English words don't help to reach the goal and needs. More complex transliteration techniques may be needed for converting them to English Words. These non English words had been removed with the help of textclean package and in comparison with WordNet dictionary.

The approach implemented in the system consists of two levels.

- Generating a new text representation based on merging terms with their associated concept.
- Selecting the characteristic features for creating the categories profiles.

The Fig.2 shows the sample data for feature selection procedure and mapping them to relevant concepts.

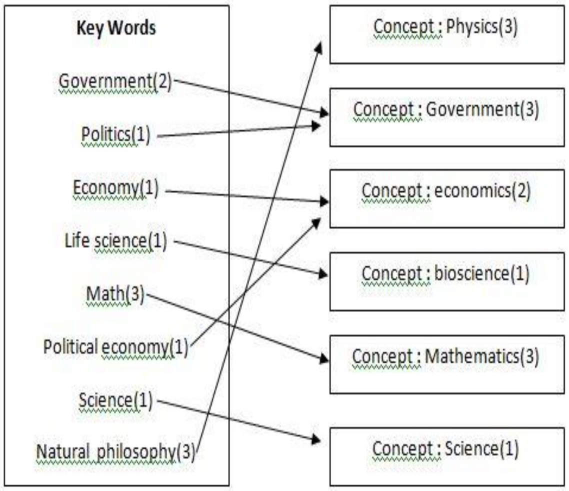

**Fig2. Sample data for mapping features to concepts.**

## V. DATA CLASSIFICATION – UNSUPERVISED LEARNING

### *A. Latent Dirichlet allocation (LDA)Algorithm:*

Latent Dirichlet allocation (LDA)[6]is a method for fitting a topic model. It handles each document as a mixture of topics, and each topic as a mixture of words. So it allows documents to "overlap" each other in terms of content, rather than separated into discrete groups. There are two specific principles for conceptual understanding.

**Every document is a mixture of several topics.** For e.g. each document may contain words from several topics in particular proportions i.e. in a 2 topic model, we could say "Document 1 contains 40% of topic A and 60% of topic B, while Document 2 contains 20% of topic A and 80% of topic B."

**Every topic is a mixture of many words.** Let's consider a sample data, with a two-topic model of American news, with one topic for "politics" and one for "entertainment." Most common words in the politics topic might be "Congress", "President", and "government" and in the entertainment topic, it may be made up of words such as "movies", "television", and "actor". Also, the words can be shared between topics; a word like "budget" might appear in both equally.

This algorithm is a mathematical method for estimating both of these at the same time: finding the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document. Many numbers of existing implementations of this algorithm are available. We had used LDA algorithm for categorizing the generated features into certain topics of discussion to identify the major topics of interest among students.

### *B. Analyzing Sentiments*

With the explosion of digital and social media, there are various emoticons and emojis that can be embedded in text messages, emails, or other various social media communications, for use in expressing personal feelings or emotions.

Emotions may also be expressed in textual forms using words. R packages are used for analyzing emotions in the text and also the expression of anger, fear, anticipation, trust, surprise, sadness, joy, and disgust has been filtered. Polarity of sentiment is analyzed and classified as positive, negative and neutral based on the WhatsApp group chat conversations.

## VI. DATA VISUALIZATION

Data Visualization and Presentation of obtained result to the users are the most important essence in a data mining task. The Predicted output can be displayed in form of either word cloud, bar chart etc. It allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner. Grouping can be represented by color, symbol, size, and transparency.

### *A. Word Cloud*

Word cloud is a text mining technique that allows us to highlight the most frequently used keywords in paragraphs of text.

The Word clouds in the Fig.3 shows that the most frequently used word in I PG students WhatsApp chat conversation is 'percent' and for II PG students, the most frequently used word is 'media'. This obtained result can be further grouped under topics for better understanding of student's topic of interest.

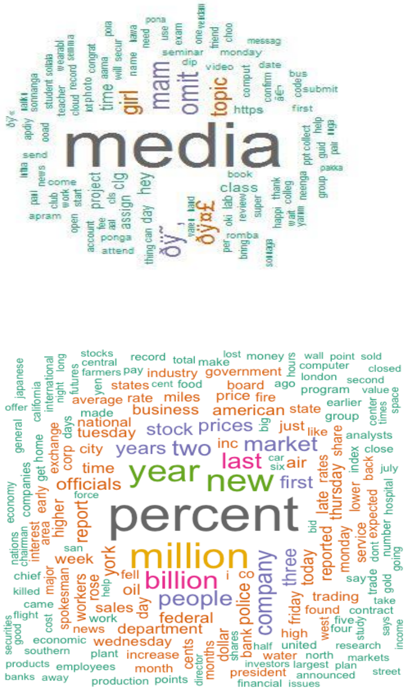

**Fig3. Word Clouds of WhatsApp group chat conversations**

### *B. Sentiment Visualization*

Sentiment Visualization is a research challenge in information visualization and visual analytics to analyze sentiment discovered in text data. The following Histogram in the Fig.4 reveals the sentiment score of student's group chat context. It classifies the sentiments of PG students of Department of Computer Science. Their usage of words in the WhatsApp Group chat implies that the students mostly share positive sentiments with nearly 300 instances in their group conversations.Anticipation and trust are the second major sentiments shared in the WhatsApp conversations with approximately 200 instances.

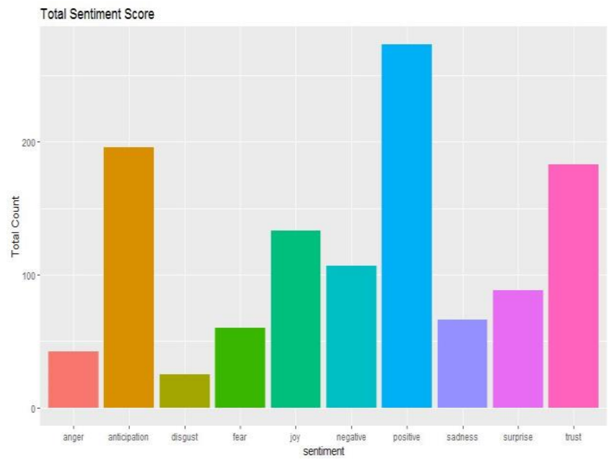

**Fig4. Classification of WhatsApp group chat conversations**

### *C. Topic Modeling*

Topic modeling is a statistical approach for discovering "abstracts/topics" from a collection of text documents based on statistics of each word. In simple terms, the process of looking into a large collection of documents, identifying clusters of words and grouping them together based on similarity and identifying patterns in the clusters appearing in multitude. Latent Dirichlet Allocation Algorithm is implemented for performing topic modeling on student's group chat. This allows to easily predict the topic of discussion/interest of group of students.

In the topic visualization shown in Fig. 5, student's topic of interest is relatively high on category 3.General topics namely business, government are relatively high when compared to academic related context.

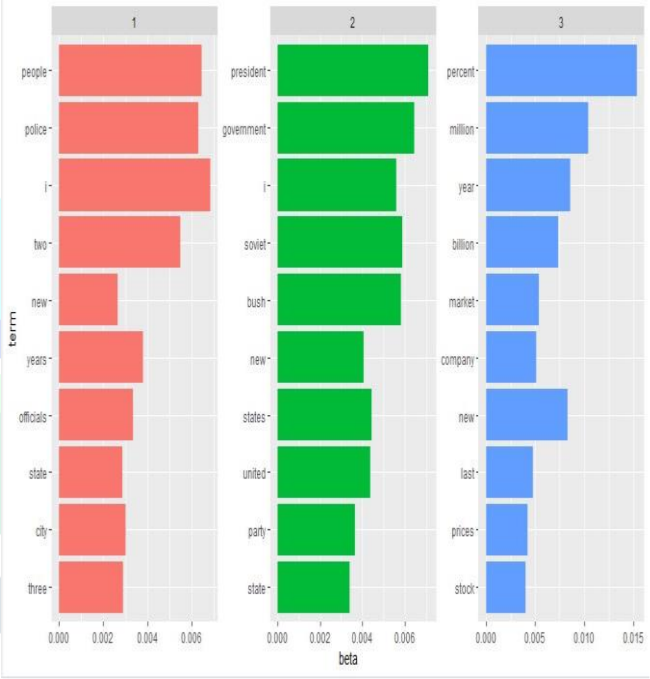

**Fig5. Percentage based on topic relevance**

## VII. CONCLUSION

From the performed analysis and visualization it is revealed that the II PG and I PG students from the Department of Computer Science shared Positive sentiments in their WhatsApp group irrespective of the topic discussed. It is also found that the percentage of student's topic of interest in 'General Knowledge' is relatively more when compared to academic contextual chats.

## VIII. FUTURE ENHANCEMENT

This project can be further enhanced in such a way that some more options can also be included for getting input dataset in other formats like audio (Speech to Text Conversation) and as images (Optical Code Recognition) from which data can be further analyzed.

Transliteration techniques and implementation of Natural Language Processing can be carried out for better text analysis.

## REFERENCES

[1] S.M. Mohammad, Challenges in sentiment analysis, in *A Practical Guide to Sentiment Analysis* (pp. 61-83). Springer, Cham,2017.Available:https://saifmohammad.com/WebDocs/sentiment-challenges.pdf

[2] Sanchita Patil, "*WhatsApp Group Data Analysis with R*", International Journal of Computer Applications (0975 – 8887) Volume 154 – No.4, November 2016.

[3] (2019) The Journocode Website. [Online]. Available: https://journocode.com/2016/01/31/project-visualizing-WhatsApp-chat-logs-part-1-cleaning-data/

[4] Suzanne van den Bosch, "*Automatic feature generation and selection in predictive analytics solutions*", Master thesis Computing Science, Faculty of Science, Radboud University.

[5] P.Sudhandradevi, V.Bhuvaneswari, "*Pattern Mining or WhatsApp Chats Identifying Thanglish Words in WhatsApp Chats*", ISSN(Online): 2347 – 2820, Volume – 6, Issue-1_2, 2018.

[6] Julia Silge, David Robinson, *"Text Mining with R: A tidy Approach"*, in Chapter 6: Topic Modeling, Published by Sebastopol, CA: O'Reilly Media, 2017. Available: https://www.tidytextmining.com/topicmodeling.html

[7] D.Radha, R. Jayaparvathy, D. Yamini, "*Analysis on Social Media Addiction using Data Mining Technique*", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, pp.23-26, April 2016.

[8] Bing Liu, "*Sentiment Analysis and Opinion Mining*", Published by Morgan &ClayPool Publisher, May 2012.

[9] V.Bhuvaneswari, "*Data Analytics with R - A Practitioner's Approach*", First Edition-2016, Published by Department of Computer Applications, Bharathiar University, ISBN 978-81-929131-2-.4