IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

Whatsapp Group Chat Analysis

D V Swetha Ramana, Anusha K M, Bhagya K, Keerthana M, Lakshmi G
Dept. of CSE,
Rao Bahadur Y Mahabaleswarappa Engineering College, Ballari

Abstract: In the past few years, the large-scale dissemination of misinformation through social media has become a critical issue, harming the trustworthiness of legit information, social stability, democracy and public health. Thus, developing automated misinformation detection methods has become a field of high interests both in academia and in industry. In many developing countries such as Brazil, India, and Mexico, one of the primary sources of misinformation is the messaging application WhatsApp. Despite this scenario, due to the private messaging nature of WhatsApp, there still few methods of misinformation detection developed specifically for this platform. In this paper, the WhatsApp data was collected from dataset repository. Then, we have to implement the pre- processing techniques. Then, the system is developed the NLP techniques. Then, we have to implement the machine learning algorithm such as **DT** and **SVM**. The experimental results shows that the accuracy.

Keywords: Machine learning, Authentication System, Chat Analysis, Data preprocessing

I. INTRODUCTION

WhatsApp is an instant messaging application that allows users to send text messages, chat and share media files like images, audio and video files. Users can also share documents and applications. With WhatsApp, users have the opportunity to communicate with several other users at the same time in a group. In addition, a user can send a broadcast message to up to two hundred and fifty six (256) users at a single message stance. This feature makes the message to appear as though it was sent to each individual alone.

This Project is based on data analysis and processing. Data pre- processing plays a major role when it comes to machine learning. it is to understand the implementation and usage of various python inbuilt modules. The Python libraries are used such as NumPy, pandas, matplotlib, seaborn etc.

The application software can be used on different Internet Operating Systems (iOS) such as Android, Apple and Windows iOS. According to WhatsApp is an application that facilitates the exchange of instant messages, pictures, videos and voice calls through an Internet connection. It enables easy communication via text or voice messages between two or more persons. Basically it helps people to stay connected.

The fact that WhatsApp can be termed as cost free, clearly explains the success of WhatsApp. Also, its function across different smartphone types like Apple, Android, etc, and its international functionality are also important contributors to this popularity. According to, WhatsApp service had up to 450 million monthly average users as at 2014. Also, a data analysis by, showed that the use of WhatsApp accounted for 19.83% of all smartphone in 2015 as compared to Facebook which takes only 9.38%. Also, female folks were reported to be using WhatsApp for significantly longer period of time than the male. Furthermore, they stated that younger people use WhatsApp for a longer duration of time. WhatsApp can be described as free but since it makes use of data, one can be said to be paying for it since you cannot make use of it without data. It works on iPhone, Android, BlackBerry, Symbian, and Windows devices. The information needed by a user is ones username name and phone number. Messages on WhatsApp are said to be encrypted.

II. LITERATURE REVIEW

Many researches have been carried out for analyzing the WhatsApp and their features. Some of the them are as follows In paper [1] *Examining the Acceptance of WhatsApp Stickers Through Machine Learning Algorithms, 2020.*
Authors: Rana A. Al-Maroof ,Ibrahim Arpaci, Mostafa Al- Emran, Said A. Salloum & Khaled Shaalan
Methodology:
WhatsApp stickers are gaining popularity among university students due to their pervasiveness, specifically in educational WhatsApp groups. However, the acceptance of stickers by university students is still in short supply. Thus,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

this research aims to empirically examine the determinants affecting the acceptance of WhatsApp stickers through a proposed theoretical model by integrating the technology acceptance model (TAM) with the uses and gratifications theory (U&G). A questionnaire survey was circulated to collect data from 372 university students who have been engaged in a "Group Talk" in WhatsApp. A novel approach was employed to analyze the hypothesized relationships among the constructs in the research model through the use of machine learning algorithms. The results pointed out that IBk and RandomForest classifiers have performed better than the other classifiers in predicting the actual use of stickers with an accuracy of 78.57%. The research findings are believed to provide future directions for stickers developers to better promote stickers in educational activities.

In paper [2] *Text Classification based Behavioural Analysis of WhatsApp Chats,2020*
Author: Sonika Dahiya; Astha Mohta; Atishay Jain
Methodology:
WhatsApp is used by millions of users to express emotions and share feelings. The model is presented in this paper aims to perform sentimental and emotional analysis using textual messages and emojis used in WhatsApp chats. Code switching, which is quite prevalent over online conversations, is handled by the model by unifying and converting all the texts to a standard form. For each subject, multiple chats are taken; translated and using a neural network, each sentence and emoji is scored in a dimensional form. The composition of the emotions expressed by the subject (out of Happy, Sad, Bored, Fear, Anger and Excitement) are defined. The scores are added up for each subject. Throughout the analysis, the behavioral traits are extracted. It is determined that, if the subject likes to use emojis and if they use it as a replacement for words or as an add-on to express their emotions better. It is also observed that if the subject behaves differently on text according to the person in front of them with regard to these emotions and finally, if the subject is an introvert or extrovert.

In paper [3] Unsupervised WhatsApp Fake News Detection using Semantic Search, 2020.
***Author*:** Jaynil Gaglani; Yash Gandhi; Shubham Gogate; Aparna Halbe
Methodology:
Social media has become the backbone of today's lifestyle. It has a widespread effect on nearly every walk of life. One of the well-known social media applications WhatsApp Messenger is a free and cross-platform text messaging software that also provides services for sending and receiving multimedia messages. But at the same time in recent years, its easy accessibility has served a way for propagating fake and biased news articles, blogs and messages. Fake news and messages have paved their way for Political polarization, ethnic tensions, unwanted panic and mass hysteria. A solution is proposed that uses Natural Language Processing for analyzing the messages and leverage Transfer Learning Models to detect the authenticity of the information. Claims are filtered from the bulk of forwarded messages disseminated on WhatsApp. The solution comprises of a semantic search mechanism between each claim and associated news sources. The similarity comparison done by the model predicts the truthfulness of the claim

III. EXISTING SYSTEM

In existing system, User-generated content in the form of reviews, ratings, and comments can be analysed for greater insights for enterprise use. The analysis of such consumer behaviour is helpful to understand the consumer's requirements and predict their future intentions towards the service. Through this cognitive study, E-commerce Organizations can track the usage and sentiments attached to their products and take appropriate marketing approaches to provide a personalized shopping experience for their consumers, thereby increasing their organizational profit. This paper aims to employ data- driven marketing tools, such as data visualization, natural language processing, and machine learning models that help in understanding the demographics of an organization. We also build recommender systems through collaborative filtering, neural networks, and sentiment analysis.

Disadvantages of Existing System

- The results is low when compared with proposed
- It doesn't efficient for large volume of data's

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

- Time consumption is high
- It doesn't implement the API key
- Theoretical limits.

IV. PROPOSED SYSTEM

In this system, the whastapp dataset was taken as input. The input data was taken from datset reporitory. Then, we have to implement the data pre-processing step. In this step, we have to handle the missing values for avoid wrong prediction, and to encode the label for input data. Then, we have to analyses the sentiment by using the natural language processing. In this step, we have to remove punctuations, stop words and stemming. Then, we have to split the dataset into test and train. The data is splitting is based on ratio. In train, most of the data's will be there. In test, smaller portion of the data's will be there. Training portion is used to evaluate the model and testing portion is used to predicting the model. Then we have to implement the vectorization. It means, to encode the text as integers or numeric value to create the feature vectors. Then, we have to implement the classification algorithm (i.e.) machine learning. The machine learning algorithms such as DT and SVM. Finally, the experimental results shows that the performance metrics such as accuracy, precision and recall.

V. OBJECTIVE

The main objective of our project is,

- To classify the whatsApp into good and bad.
- To implement the machine learning algorithms and NLP techniques.
- To enhance the overall performance for classification algorithms.
- To classify or predict the whatsApp message effectively.

VI. ARCHITECTURE

The system architecture for WhatsApp group chat analysis can be designed using various technologies and tools based on the specific requirements of the analysis. Generally, the architecture can include the following components:

- Data collection: The first component of the system architecture is data collection. It involves collecting the data from the WhatsApp group chat, including text messages, media files, and other relevant information.
- Data pre-processing: The collected data needs to be pre- processed to clean the data, remove unwanted information, and structure the data for analysis. This process can include tasks such as data cleaning, data normalization, and data transformation.
- Analysis engines: The next component of the system architecture is the analysis engines. This includes the various algorithms and techniques used to analyze the data collected from the WhatsApp group chat. For example, machine learning algorithms can be used for sentiment analysis, topic modeling, and network analysis.
- Visualization and reporting: The final component of the system architecture is the visualization and reporting layer. This layer involves visualizing the analyzed data in various formats such as graphs, charts, and dashboards. It also includes generating reports based on the insights obtained from the analysis.

The system architecture for WhatsApp group chat analysis can be deployed on-premise or on the cloud, depending on the size and complexity of the analysis. Various tools and technologies such as Python, R, Hadoop, Spark, and Tableau can be used to implement the different components of the system architecture.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

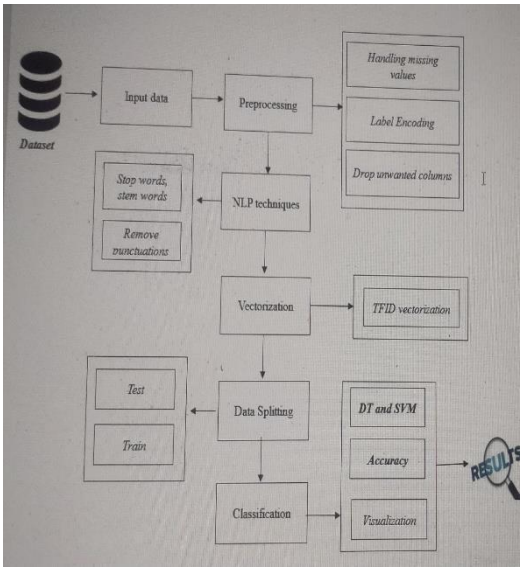


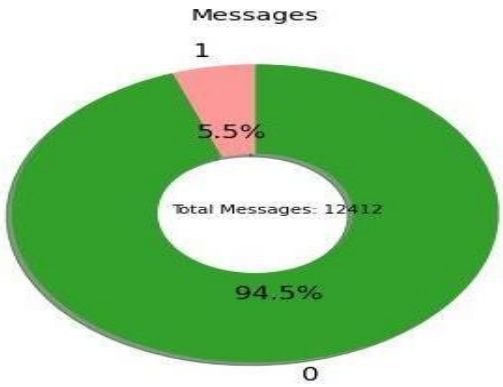
Fig 1. System Architecture

VII. ALGORITHM SPECIFICATION

- Import chat data
- Clean the chat data
  - Remove irrelevant messages (e.g. system messages)
  - Correct typos and spelling mistakes
  - Remove stop words (e.g. 'the', 'and', 'in')
- Perform text mining
  - Tokenize messages into individual words
  - Perform keyword extraction to identify commonly used words and phrases
  - Perform topic modeling to identify common themes within the chat data
- Perform sentiment analysis
  - Determine the sentiment of each message (e.g. positive, negative, neutral)
  - Calculate the overall sentiment of the chat data
- Perform network analysis
  - Identify the most active members
  - Analyze communication patterns (e.g. who talks to who most often)
  - Determine the overall structure of the group (e.g. is it hierarchical or flat?)
- Generate visualizations
  - Create word clouds to visualize commonly used words and phrases
  - Create graphs and charts to visualize communication patterns and sentiment trends
- Present findings and insights
  - Summarize the analysis results and key insights
  - Provide recommendations for improving group dynamics and communication
- Save and export analysis results
  - Save the analysis results in a structured format (e.g. spreadsheet, database)
  - Export the analysis results for further analysis or for use in other systems

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

VIII. RESULTS



IX. CONCLUSION

We conclude that, the WhatsApp data was collected as input. We are implemented the NLP techniques and classification algorithms (i.e.) machine learning algorithm. Then, machine learning algorithms such as DT and SVM. Finally, the result shows that the accuracy for above mentioned algorithm and visualize the output in the form of graph. Then, analyse the WhatsApp chat is positive or negative.

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like, Precision, Accuracy and Recall.

X. FUTURE ENHANCEMENT

In the future, we should like to hybrid the two different machine learning. In future, it is possible to provide extensions or modifications to the proposed classification algorithms to achieve further increased performance. Apart from the experimented combination of data mining techniques machine algorithms can be used to improve the detection accuracy. Finally, the sentiment analysis detection system can be extended as a prevention system to enhance the performance of the system.

REFERENCES

[1] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2019.

[2] N. Ahmad and J. Siddique, “Personality assessment using Twitter tweets,” Procedia Comput. Sci., vol. 112, pp. 1964– 1973, Sep. 2019.

[3] T. Ahmad, A. Ramsay, and H. Ahmed, “Detecting emotions in English and Arabic tweets,” Information, vol. 10, no. 3, p. 98, Mar. 2020.

[4] A. Bandi and A. Fellah, “Socio-analyzer: A sentiment analysis using social media data,” in Proc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64,

[5] F. Harris, S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp. 61–67.

[6] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in Twitter,” in Proc. ICCC, 2014, pp. 155–162.

[7] R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, “COVID 2019 outbreak: The disappointment in Indian teachers,” Asian J. Psychiatry, vol. 50, Apr. 2020, Art. no. 102047.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2021.

[9] P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha, and G. Röst, “Risk assessment of novel coronavirus COVID-19 outbreaks outside China,” J. Clin. Med., vol. 9, no. 2, p. 571, Feb. 2020.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 5, May 2023

[10] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," Information, vol. 9, no. 5, p. 127, May 2018.

[11] X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering," 2001, arXiv:cs/0109015. J. P. Carvalho, H. Rosa,

[12] G. Brogueira, and F. Batista, "MISNIS: An intelligent platform for Twitter topic mining," Expert Syst. Appl., vol. 89, pp. 374–388, Dec. 2017.

[13] B. K. Chae, "Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research," Int. J. Prod. Econ., vol. 165, pp. 247–259, Jul. 2015.

[14] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2013, pp. 3267–3276.

[15] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pandemic of social media panic travels faster than the COVID-19 outbreak," J. Travel Med., vol. 27, no. 3, Apr. 2020, Art. no. taaa031.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[17] M. E. El Zowalaty and J. D. Järhult, "From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans—Call for a one health approach," One Health, vol. 9, Jun. 2020, Art. no. 100124.