JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

WHATSAPP CHAT ANALYZER

Samiksha Sanjay Malage & Aayush Rajendra Pawar

Guide: Asst. Prof. Gauri Mhatre

Keraleeya Samajam’s Model College, Khambalpada Road, Thakurli, Dombivli (East), Maharashtra

ABSTRACT

The Growth of WhatsApp as a primary communication platform has led to a surge in spam messages, posing significant risks to user privacy and security. The complicated and dynamic nature of spam on WhatsApp, where messages are frequently informal and diversified and include slang, abbreviations, and multimedia features, makes traditional rule-based spam detection algorithms inadequate. This study investigates the creation of a sophisticated machine learning-based system intended to examine WhatsApp conversations and efficiently identify spam.

Utilizing Natural Language Processing (NLP) techniques, the purpose of this research is to develop a strong spam detection algorithm that can handle WhatsApp messaging's complexities. Data collection from WhatsApp conversations, preprocessing to clean and normalize the text data, feature extraction to identify crucial components indicating of spam, and training of a variety of machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and sophisticated deep learning models like Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), are some of the crucial steps in the methodology.

INDEX TERM

Whatsapp chat data, Analysis, Sentiment Analysis, NLP, Spam detection, Features and models.

1 INTRODUCTION

WhatsApp has emerged as a dominant communication tool globally, boasting over 2 billion active users as of 2023. This platform is a flexible tool for individual and group interactions because it supports a wide range of interaction formats, including voice notes, photos, videos, and simple text messages. But since WhatsApp is so widely used, bad actors have also been drawn to the network, using it to transmit spam messages. These spam messages can contain everything from harmless adverts to links to malware and other potentially harmful content that jeopardizes user security and privacy.

As chat interactions on WhatsApp are informal, it might be difficult to identify spam. Unlike emails or SMS, which are more formal forms of communication, WhatsApp messages frequently include slang, acronyms, emoticons, and other multimedia components. It is challenging for conventional rule-based spam detection techniques to be effective because of this variation in communication styles. Moreover, spammers constantly modify their strategies in order to avoid detection, calling for increasingly complex and adaptable solutions.

With its ability to learn from and adapt to massive datasets, machine learning presents a potent method for addressing WhatsApp's spam detection issue. A branch of machine learning called natural language processing (NLP) is especially important since it studies how computers and human language interact. We can efficiently analyze and categorize the enormous and diverse textual data found in WhatsApp messages by utilizing NLP techniques like tokenization, stemming, lemmatization, and advanced algorithms like Long Short-Term

Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT).

Data collection is undertaken with strict adherence to ethical guidelines, ensuring user consent and anonymization to protect privacy. To prepare the text for analysis, preprocessing techniques include tokenization, stemming, lemmatization, and the elimination of stop words and special characters. Word embeddings and Term Frequency-Inverse Document Frequency (TF-IDF) are two feature extraction techniques used to extract the semantic meaning and significance of words in the texts.

The efficacy of these models in recognizing spam communications is evaluated by the research using criteria including accuracy, precision, recall, and F1-score. The results show how effective deep learning models are, namely BERT, which showed excellent accuracy and dependability in spam detection. Significant improvements are seen when compared to conventional rule-based methods, indicating the potential of machine learning techniques in this area.

The objective of this research is to create a sophisticated machine learning-based spam detection system by analyzing WhatsApp talks. The goal of this system is to overcome the limitations of conventional approaches by effectively identifying and filtering out spam messages in real-time. The research will gather data from WhatsApp chats, preprocess the text to make it readable for analysis, extract features from the data to find the most relevant information, then train and test multiple machine learning models.

This task is important because it has the potential to greatly enhance WhatsApp users' experiences by decreasing the amount of spam they get, improving their safety and privacy. In addition, the approaches created in this study can be modified for use with other messaging services, supporting wider initiatives in digital communication and cybersecurity. The goal of this research is to offer a solid answer to the continuous problem of spam identification in modern communication platforms by utilizing the power of machine learning.

1.1 PROBLEM STATEMENT

Spam in WhatsApp interaction are generally understood to be unsolicited, frequently incorrect or irrelevant messages sent through a platform. These messages have the ability to delay communication and give rise to more severe problems such as virus attacks, fraud, or phishing scams. Typical forms of spam consist include:

1. Advertisements: unsolicited advertisements for goods or services that the receiver has not indicated a desire to purchase.
2. Phishing Attempts: messages that aim to trick the receiver into disclosing credit card details, passwords, or other private information.
3. Scams: fraudulent communications that try to fool recipients into sending money or disclosing personal information.
4. Malware Links: messages with links that direct recipients to fraudulent websites with the intention of installing malware on their device.
5. Chain Messages: messages with fictitious prizes or warnings of repercussions that entice recipients to forward them to others.

1.2 SCOPE OF THE STUDY

The scope of this study on WhatsApp chat analysis is to improve our knowledge and provide useful tools for analyzing WhatsApp chats. It includes a number of important topics. Sentiment analysis and spam identification, which make use of complex machine learning and natural language processing (NLP) techniques, will be the main applications of focus. The ethical and privacy issues regarding the analysis of private chat data will also be addressed by the study.

1.3 LITERATURE SURVEY

The analysis of WhatsApp chats has emerged as a significant research area due to the platform's widespread adoption and the rich, unstructured data it generates. Sentiment analysis is a primary focus, aiming to classify the emotional tone of messages. Machine learning techniques have essentially replaced the lexicon-based methods used in earlier approaches, which used predetermined dictionaries of positive and negative words. The effectiveness of supervised learning algorithms and deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), is demonstrated in studies by Socher et al. (2013) and Kim (2014). WhatsApp conversation is informal and diverse, with emojis, slang, and acronyms contributing to its specific issues. According to Sarker and Gonzalez (2017) and Nuruzzaman et al. (2019), accuracy can be improved by using hybrid models that combine machine learning with lexicon-based techniques. Another important factor is multilingual sentiment analysis because WhatsApp has users all over the world. The models created by Mozafari et al. (2020) and Zhou et al. (2016) use cross-lingual embeddings and transfer learning techniques to handle multiple languages effectively.

In WhatsApp, spam detection is crucial to preserving user experience and security. The efficacy of classifiers like Random Forest and Gradient Boosting has been demonstrated in research by Almeida et al. (2013) and Biggio et al. (2014), illustrating how traditional keyword-based filtering techniques have developed into complex machine learning models. Deep learning models, such as CNNs and Long Short-Term Memory (LSTM) networks, can capture complex patterns in spam communications, greatly increasing detection rates, according to recent research by Gomez Hidalgo et al. (2020) and Zong et al. (2019). Studies by Zhang et al. (2019) and Liu et al. (2020) concentrate on optimizing models for real-time performance utilizing lightweight neural networks and effective data preprocessing strategies. Real-time detection is critical.

2 METHODOLOGY

2.1 DATA COLLECTION

Collecting WhatsApp chat data involves obtaining consent from users and extracting messages from the WhatsApp application or its backups. Here's a description of the typical process for collecting WhatsApp chat data:

1. User consent: It is necessary to have the users whose data will be gathered to give their express approval before collecting WhatsApp chat data. This ensures adherence to ethical principles and privacy laws. Users must be made aware of the reason for data collection, the intended use of their information, and any potential dangers or ramifications.
2. Message Extraction: Chat data on WhatsApp can be retrieved directly from the app, from backup files kept on the user's device, or from cloud storage. Chat data can be extracted using a variety of tools and techniques, such as scripts, third-party software, and manual techniques. Timestamps, metadata, multimedia files, and text messages are frequently included in the collected data.
3. Data formatting: In order to prepare the WhatsApp chat data for analysis, it might be necessary to format and process it after it has been extracted. This could entail cleaning the data to get rid of duplicates, pointless messages, and sensitive information, as well as turning it into a common format like CSV or JSON.
4. Storage and Security: To avoid unwanted access or data breaches, the gathered WhatsApp chat data needs to be kept in a secure location. The confidentiality and integrity of the data should be safeguarded by the implementation of encryption, access controls, and secure storage procedures.

2.2 DATA PREPROCESSING

Data preprocessing is a critical step in preparing WhatsApp chat data for analysis, involving several steps to clean and preprocess the raw text data. Here are the key steps for cleaning the data and techniques for handling imbalanced data:

Steps for Cleaning the Data:

1. Removing Stop Words: Common words like "and," "the," and "is" that have little relevance in text analysis are considered stop words. By concentrating on more informative terms, stop words are eliminated, which lowers noise and increases the effectiveness of text analysis algorithms.
2. Handling Misspellings: Spelling errors are frequent in casual chat communications and can have an impact on how accurately text analysis activities perform. Misspellings can be handled and the text data can be standardized by using techniques like spell checking, normalization, and correction.
3. Removing Special Characters and Punctuation: In order to simplify the text data, special characters, punctuation, and symbols can be eliminated if they don't add much to text analysis activities.
4. Tokenization: Tokenization is the process of breaking the text data into smaller units, such as letters, words, or subwords. By

dividing the material into digestible chunks, tokenization makes additional analysis easier.

5. Stemming and Lemmatization: Lemmatization and stemming are methods for breaking words down to their most basic or root form. Lemmatization connects words to their canonical forms, whereas stemming eliminates prefixes and suffixes from words. These methods aid in decreasing the dimensionality of the feature space and normalizing word variants.
6. Handling Emojis and Emoticons: Emojis and emoticons, which represent feelings and emotions, are frequently included in WhatsApp chat data. The particular analysis task at hand will determine whether these symbols are kept or swapped out for their textual equivalents.

Imbalanced data refers to datasets where the number of instances in one class is significantly higher or lower than the number of instances in other classes. Handling imbalanced data is crucial to prevent bias and ensure accurate model performance. Here are some techniques for handling imbalanced data:

1. Resampling: Resampling techniques involve either oversampling the minority class or undersampling the majority class to balance the dataset. Oversampling techniques include duplication, synthetic generation (e.g., SMOTE), while under sampling techniques involve randomly removing instances from the majority class.
2. Weighted Loss Functions: During model training, instances from the minority class are given higher weights to indicate their relative importance. This reduces bias towards the majority class and improves the model's ability to learn from unbalanced data.
3. Ensemble Methods: To increase overall performance, ensemble methods integrate predictions from several classifiers. When dealing with unbalanced data, strategies like bagging, boosting, and stacking can be applied to produce reliable and accurate models.
4. Selecting the Right Algorithm: Various machine learning algorithms react differently to unbalanced input. When dealing with unbalanced datasets, algorithms like gradient boosting, random forests, and decision trees often function well.

3 FEATURE EXTRACTION AND MODELS

Feature extraction is a crucial step in text analysis, where raw text data is transformed into a numerical representation suitable for machine learning algorithms. Here are some methods for extracting features from text and a discussion on the importance of feature selection:

Methods for Extracting Features from Text:

1. Term Frequency-Inverse Document Frequency (TF-IDF) : A statistical metric called TF-IDF assesses a word's significance in a document in relation to a corpus of documents. It calculates the product of inverse document frequency (IDF), which penalizes terms that appear often across documents in the corpus, and term frequency (TF), which counts how often a word appears in a document. Words that are common in a document but uncommon in the corpus are given greater weights by TF-IDF, which increases their discriminative power for classification tasks.
2. Word Embeddings: Using techniques like Word2Vec, GloVe, or BERT, dense vector representations of words extracted from massive text corpora are known as word embeddings. These embeddings help models better capture similarities and contrasts between words by encoding contextual information and capturing semantic links between words. In text analysis tasks including named entity recognition, document classification, and sentiment analysis, word embeddings are frequently utilized.

FEATURE SELECTION:

1. Dimensionality Reduction: The abundance of features (words, for example) in text data can cause the curse of dimensionality and raise computing complexity. By selecting only the most pertinent and instructive features, feature selection approaches assist decrease the dimensionality of the feature space while enhancing the performance and efficiency of the model.

2. Enhanced Model Generalization: Feature selection reduces overfitting and enhances the capacity of machine learning models to generalize by picking pertinent characteristics and eliminating unnecessary or redundant ones. Less noise in the data is likely to be remembered by models trained on a lesser number of features, and they are also more likely to generalize well to new examples.
3. Interpretability: By determining which features are most crucial to the model's predictions, feature selection can improve the interpretability of text analysis models. This understanding can be helpful for domain specialists as well as academics and practitioners in determining which features of the text data are influencing the model's judgments.
4. Effective Use of Resources: By focusing on a subset of pertinent features, feature selection lowers the amount of computing power needed for model training and inference. This can result in reduced memory utilization, faster model training durations, and better scalability—especially for large-scale text analysis a position.

MODEL SELECTION

Criteria for choosing machine learning models include:

1. Performance: High recall, accuracy, precision, and F1-score should be attained by the models.
2. Scalability: Effective models should be able to handle big datasets and lots of computing power.
3. Interpretability: Models ought to shed light on the processes involved in making decisions.
4. Robustness: Models ought to manage noisy inputs and generalize well to previously unknown data.

Overview of selected models:

1. Support Vector Machines (SVM): SVMs are useful for binary classification jobs because they identify the hyperplane that best divides the data.
2. Recurrent Neural Networks (RNN): Text temporal dependencies are captured by RNNs, which are suited for sequential data.
3. BERT: Modern transformer-based model for tasks involving the understanding of natural language.

MODEL TRAINING

The training process for machine learning models involves several steps to ensure the model learns from the data effectively:

1. Data Preparation: Organize the dataset into sets for testing and training. The testing set assesses the model's performance, and the training set is used to teach it.
2. Feature Extraction: Create numerical characteristics from raw text data by applying techniques such as word embeddings and TF-IDF.
3. Model Initialization: Select an appropriate machine learning model (such as SVM, RNN, or BERT) and set up its parameters.
4. Training Loop: To get predictions, feed the model with the training set of data. Compute the error (loss) that exists between the model's predicted labels and the actual labels. Backward Pass: To minimize the loss, compute gradients via backpropagation and update the model's parameters.
5. Iteration: For several epochs, repeat the forward and backward passes until the model converges to a performance level that is acceptable.
6. Validation: To keep an eye on the model's performance and avoid overfitting, periodically assess it on a validation set.

MODEL EVALUATION

Metrics for Evaluating Model Performance:

1. Accuracy: The proportion of cases that were accurately predicted to all instances.
2. Precision: The proportion of correctly anticipated positive forecasts to all positive predictions.
3. Recall: The proportion of all actual positives to genuine positive predictions.
4. F1-Score: The harmonic mean of recall and precision yields a single measure that strikes a balance between the two.

4 Examples of Chat Data and Model Predictions And Error

1. Spam Message: "Congratulations! You have won a free vacation. Click here to claim your prize."

Model Prediction: Spam

Explanation: Keywords like "Congratulations" and "free vacation" along with a call to action (click here) are typical indicators of spam.

2. Non-Spam Message: "Hey, are we still on for dinner tonight?"
Model Prediction: Non-Spam
Explanation: The message's casual conversational tone and lack of promotional language classify it as non-spam.

Error Analysis

1. False Positives: Messages that are not spam but are mistakenly categorized as spam because they contain specific terms or phrases that are commonly seen in spam but are used in other contexts.
2. False Negatives: Spam communications that the model is unable to recognize, frequently as a result of subtlety or previously untrained spam techniques.

5 Ethical Considerations and Anonymization of Data

Ethical considerations are paramount when collecting and using WhatsApp chat data for research or analysis. Here are some key ethical considerations and practices for handling WhatsApp chat data:

1. Informed Consent: As previously said, gathering users' informed consent is a prerequisite to gathering their WhatsApp chat history. Users need to be properly informed about the reason behind data collection, the intended use of their data, and any possible dangers or repercussions.
2. Anonymization: Prior to analysis, WhatsApp chat data must be anonymized in order to safeguard user privacy. In order to do this, personally identifiable information like names, phone numbers, and other identifying characteristics must be deleted or made anonymous. Anonymization lowers the possibility of privacy violations and helps avoid the identification of specific people based on their conversation data.
3. Data Security: To avoid misuse, illegal access, and data breaches, it is crucial to ensure the security of WhatsApp conversation data. Access restrictions should be put in place to limit access to only authorized workers, and data should be encrypted both while it's in transit and when it's at rest.
4. Respect for Privacy: Researchers should be mindful of WhatsApp users' privacy and not reveal or share private information gleaned from chat data without authorization. Any analysis or conclusions should be communicated in a way that respects the participants' confidentiality and privacy.

6 Challenges in Detecting Spam Due to the Informal Nature of Chat Language

Detecting spam in WhatsApp messages presents unique challenges primarily because of the informal and diverse nature of chat language. These challenges include:

1. Diversity in Language Use: Slang and Abbreviations: Slang, abbreviations, and shorthand are frequently used in WhatsApp communications and can differ significantly throughout user groups and geographical areas. Common abbreviations that might not be included in traditional dictionaries include "LOL" for "laugh out loud" and "BRB" for "be right back".
2. Emojis and GIFs: Emojis and GIFs add a layer of complexity when they are used to express emotions, sentiments, and even complete texts. Emojis and text can be used together in a spam message to get past basic keyword-based screening.Informality: Conversational Tone: Standard text analysis tools have a tougher time applying to WhatsApp messages because they are typically more conversational in nature than official emails or documents. Sentence structure, punctuation, and informal grammar are frequently used, which makes detection more difficult. Typos and misspellings: People commonly make typos or misspellings in their messages, and spammers may do so on purpose to avoid being detected by algorithms.
3. Adaptive Strategies: Changing Tactics Spammers are always changing their tactics

to avoid being discovered. To avoid setting off spam filters, they could utilize clever strategies to resemble real discussions or change the format of their messages. Content Variability: The format and style of spam can differ greatly. The variety of formats—from embedded links and plain text to graphics and videos—requires flexible detection methods.

4. Contextual Knowledge: Why Context Is Important The context of the discourse can have a significant impact on how a message is understood. For example, a term that seems appropriate in one situation may be considered spam in another. It takes sophisticated natural language processing (NLP) skills to understand context.
5. User privacy: End-to-end encryption on WhatsApp makes sure that only the users who are chatting can access the messages, making it difficult to identify spam. Any detection system has to work within the limitations of protecting user security and privacy.

Limitations

The research encountered several limitations:
Data Variety: While multimedia information, including photographs, videos, and voice messages, is also common in WhatsApp discussions, the study only examined text-based spam.
Dataset Size: The training dataset's small size and lack of diversity may have hindered the model's capacity to generalize across various spam subtypes and colloquial language variants.
Real-Time Processing: The difficulties of integrating models into live chat systems and real-time spam detection were not sufficiently covered in the research.

7 Comparison with Existing Techniques and future work

Traditional Spam Detection Techniques:Conventional techniques like keyword matching and rule-based filtering are straightforward and simple to use, but they are not sophisticated enough to manage the variety and nuance of spam messages. Because these methods can't grasp semantics and context, they frequently produce greater percentages of false positives and false negatives.

Modern Techniques: The methods examined in this work, along with other contemporary approaches that make use of machine learning and deep learning, provide notable advancements. For instance, through learning from labeled data, machine learning algorithms like SVM and Naive Bayes outperform rule-based systems in terms of performance. Spam detection is further improved by deep learning models such as LSTM and BERT, which recognize the sequential and contextual aspects of text, respectively.

Future Work:

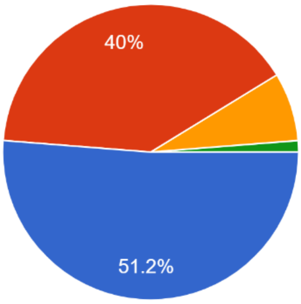
There are several areas for further research:
Multimedia Spam Detection: By adding methods for picture and video analysis, the system can be expanded to identify spam in multimedia content.
Greater Generalization and Robustness Against Different Spam Tactics: Optimize your model by utilizing datasets that are both larger and more varied.
Real-Time Implementation: Examine the difficulties and fixes associated with putting these models into practice in real-time, making sure that they integrate with messaging services such as WhatsApp in an effective and seamless manner.
Adaptive Learning Systems: Provide systems that are capable of continuously updating the model with fresh information and changing spam tactics while sustaining a high level of detection accuracy.
User Feedback Mechanisms: By integrating user-reported false positives and negatives into the training process, user feedback mechanisms can be used to enhance model performance and dependability.

8 PUBLIC SURVEY

We first conducted a poll of people through Google form creator and data collection service to acquire information regarding people's awareness.

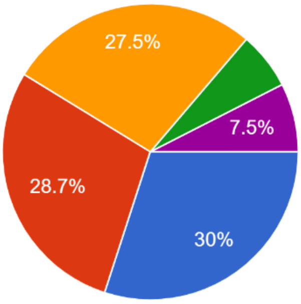
AGE

80 responses



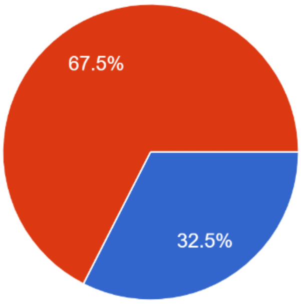
For what purposes do you primarily use WhatsApp?

80 responses



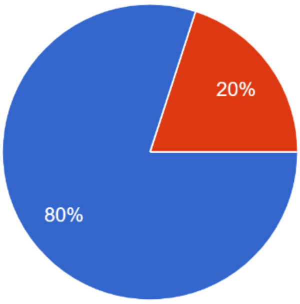
GENDER

80 responses



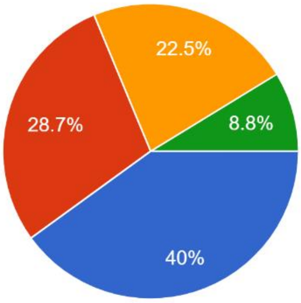
Have you ever received spam messages on WhatsApp?

80 responses



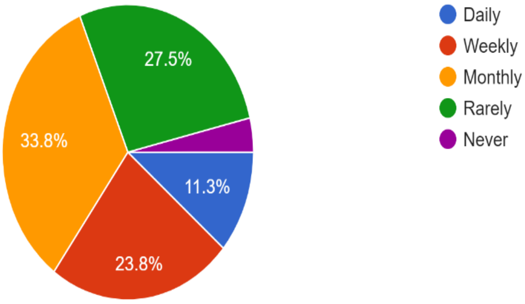
How often do you use WhatsApp?

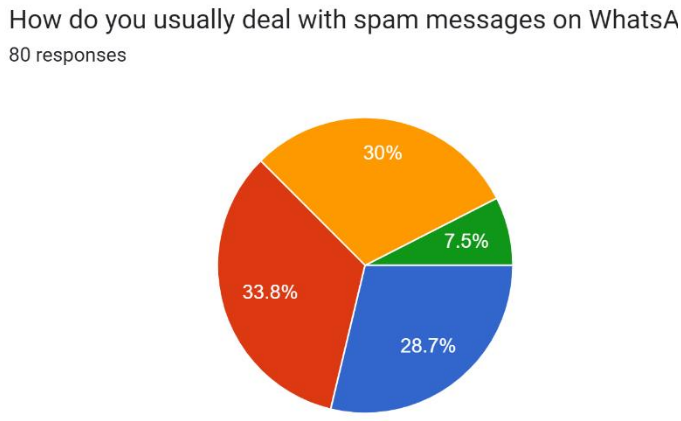
80 responses



How frequently do you receive spam messages on WhatsApp?

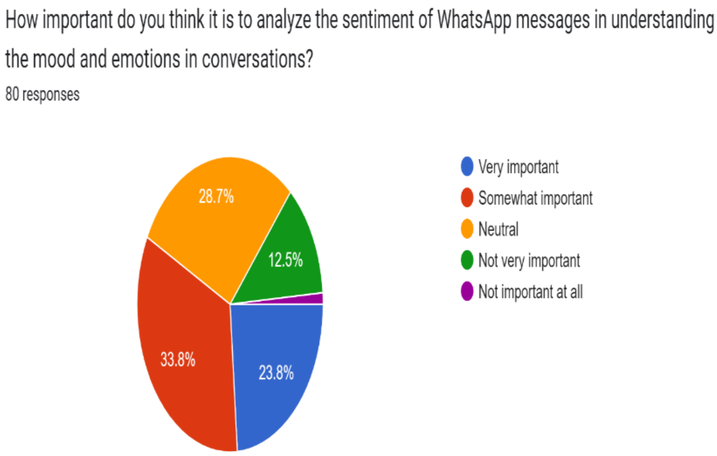
80 responses





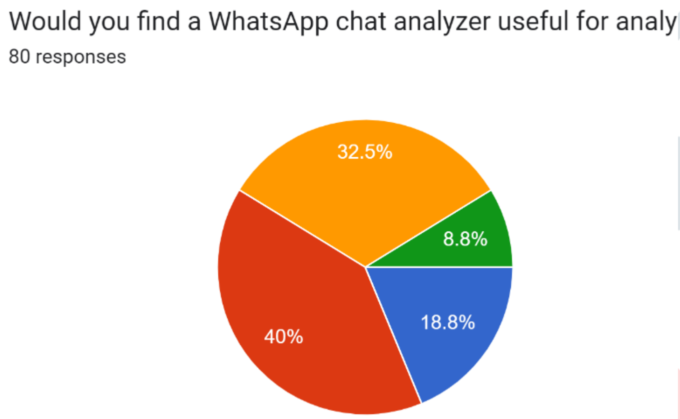
How do you usually deal with spam messages on WhatsA

80 responses



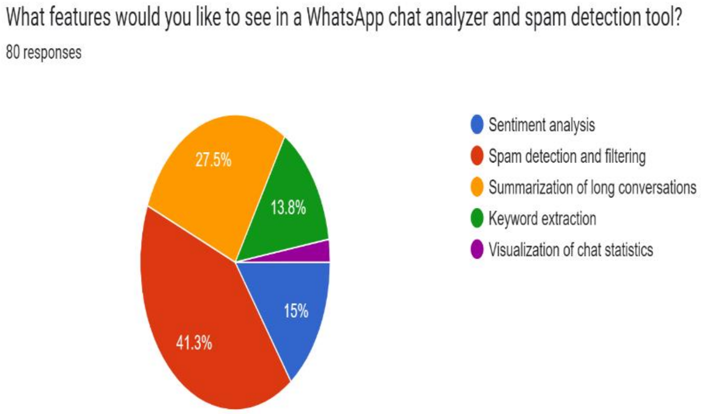
How important do you think it is to analyze the sentiment of WhatsApp messages in understanding the mood and emotions in conversations?

80 responses



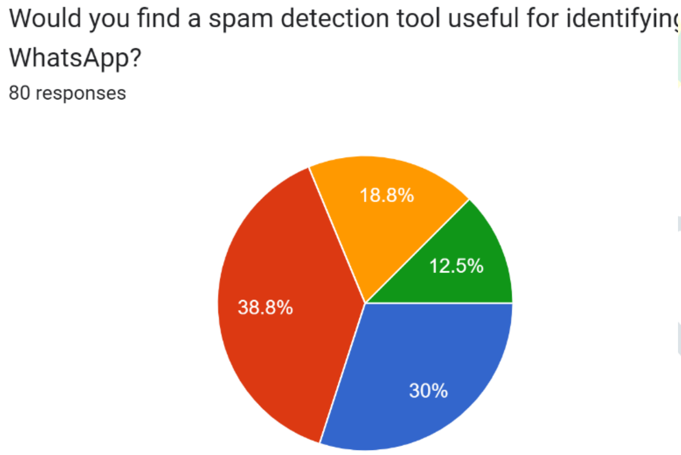
Would you find a WhatsApp chat analyzer useful for analy

80 responses



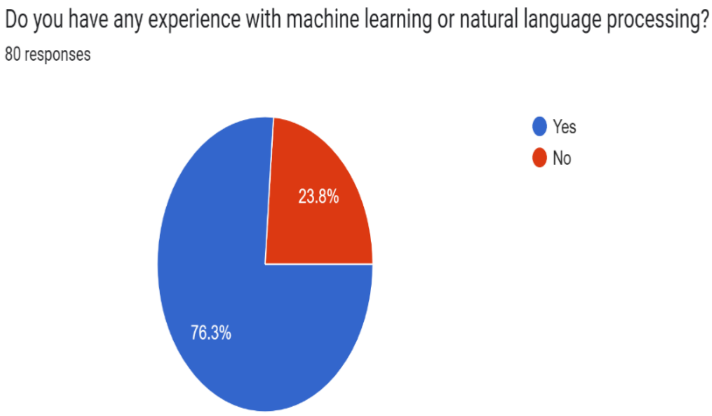
What features would you like to see in a WhatsApp chat analyzer and spam detection tool?

80 responses



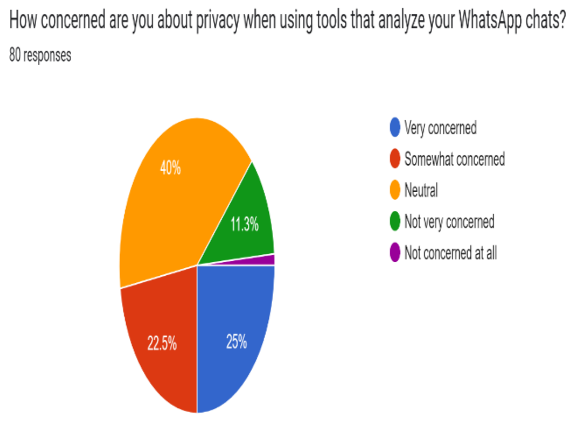
Would you find a spam detection tool useful for identifying WhatsApp?

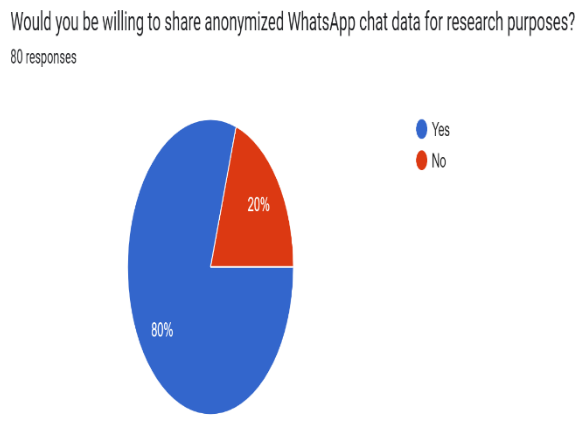
80 responses



Do you have any experience with machine learning or natural language processing?

80 responses





Descriptive Statistics

Descriptive statistics means of describing features of a data set by generating summaries about data samples. Here are some results which will helps us in finding the actual response of people.

Have you ever received spam messages on WhatsApp?	
Mean	1.2
Standard Error	0.045003516
Median	1
Mode	1
Standard Deviation	0.402523684
Sample Variance	0.162025316
Kurtosis	0.34527972
Skewness	1.528815916
Range	1
Minimum	1
Maximum	2
Sum	96
Count	80
Largest(1)	2
Smallest(1)	1
Confidence Level(95.0%)	0.089577258

How often do you use WhatsApp?	
Mean	2
Standard Error	0.111093529
Median	2
Mode	1
Standard Deviation	0.993650729
Sample Variance	0.987341772
Kurtosis	-0.847817863
Skewness	0.555797151
Range	3
Minimum	1
Maximum	4
Sum	160
Count	80
Largest(1)	4
Smallest(1)	1
Confidence Level(95.0%)	0.221126138

How frequently do you receive spam messages on WhatsApp?	
Mean	2.8875
Standard Error	0.117991834
Median	3
Mode	3
Standard Deviation	1.055351042
Sample Variance	1.113765823
Kurtosis	-0.701087563
Skewness	-0.167781538
Range	4
Minimum	1
Maximum	5
Sum	231
Count	80
Largest(1)	5
Smallest(1)	1
Confidence Level(95.0%)	0.23485687

Do you have any experience with machine learning or natural language processing?	
Mean	1.2375
Standard Error	0.047878241
Median	1
Mode	1
Standard Deviation	0.428236005
Sample Variance	0.183386076
Kurtosis	-0.430354805
Skewness	1.257394362
Range	1
Minimum	1
Maximum	2
Sum	99
Count	80
Largest(1)	2
Smallest(1)	1
Confidence Level(95.0%)	0.095299255

Would you be willing to share a2nymized WhatsApp chat data for research purposes?	
Mean	1.2
Standard Error	0.045003516
Median	1
Mode	1
Standard Deviation	0.402523684
Sample Variance	0.162025316
Kurtosis	0.34527972
Skewness	1.528815916
Range	1
Minimum	1
Maximum	2
Sum	96
Count	80
Largest(1)	2
Smallest(1)	1
Confidence Level(95.0%)	0.089577258

How do you usually deal with spam messages on WhatsApp?	
Mean	2.1625
Standard Error	0.104402686
Median	2
Mode	2
Standard Deviation	0.933806014
Sample Variance	0.871993671
Kurtosis	-0.921951482
Skewness	0.24034141
Range	3
Minimum	1
Maximum	4
Sum	173
Count	80
Largest(1)	4
Smallest(1)	1
Confidence Level(95.0%)	0.207808349

9 CONCLUSION

The analysis of WhatsApp conversations presents a significant chance to capitalize on the enormous volumes of unstructured data produced every day. Raw chat data can be effectively transformed into meaningful insights through machine learning and natural language processing (NLP) approaches, especially when it comes to sentiment analysis and spam detection. Hybrid models in sentiment analysis have demonstrated increased accuracy in identifying emotions in casual and multilingual messages by fusing sophisticated machine learning techniques with lexicon-based methods. But improving these models to properly represent complex emotions is still a work in progress.

In order to ensure a safe communication environment, advanced classifiers and deep learning models have greatly improved the identification and filtering of spam messages. Lightweight neural networks and real-time detection significantly enhance performance for real-world applications. WhatsApp chat analysis has several real-world uses, including as improving customer support, keeping an eye on mental health, and protecting company communications. These uses highlight the importance and usefulness of efficient conversation analysis technologies.

BIBLOGRAPHY

I. [1] Available from: http://www.statista.com/statistics/260819/number of-monthly-active-WhatsApp-users.

Number of monthly active WhatsApp users worldwide from April 2013 to February 2016(in millions).

II. [2] Ahmed, I., Fiaz, T., “Mobile phone to youngsters: Necessity or addiction”, African Journal of Business Management Vol.5 (32), pp. 12512-12519, Aijaz, K. (2011).

III. [3] Aharony, N., T., G., The Importance of the WhatsApp Family Group: An Exploratory Analysis. “Aslib Journal of Information Management, Vol. 68, Issue 2, pp.1-37” (2016). [4] Access Data Corporation. FTK Imager, 2013. Available at http://www.accessdata.com/support/product-downloads.

IV. [5] 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.

V. [6] Jessica Ho, Ping Ji, Weifang Chen, Raymond Hsieh, “Identifying google talk”, IEEE International Conference on Intelligence and Security Informatics, ISI ‘09, pp. 285-290, 2009.

VI. [7] Mike Dickson, “An examination into AOL instant messenger 5.5 contact identification.”, Digital Investigation, ScienceDirect, vol. 3, issue 4, pp. 227-237, 2006.

VII. [8] Mike Dickson, “An examination into yahoo messenger 7.0 contact identification”, Digital Investigation, ScienceDirect, vol. 3, issue 3, pp. 159-165, 2006.