*Original Research Article*

WhatsApp Romanized Sinhala (Singlish) Group Chat Summarization Using NLP Techniques

abstract

|  |
| --- |
| With the growing popularity of WhatsApp group chats, especially in Sri Lanka, users increasingly face challenges of information overload, leading to missed or unread important messages. While solutions exist for summarizing English-typed messages, there has been no significant attempt to summarize Singlish, a unique typing style where Sinhala words are written using the English alphabet. This research aims to address this gap by developing a Natural Language Processing (NLP)-based system to automatically summarize Singlish-typed WhatsApp group chats over 24 hours. Using exported chat data without media attachments, a customized data pre-processing pipeline was developed to clean, tokenize, and extract keywords from the chats. Two popular Summarization models, facebook/bart-base and sshleifer/distilbart-cnn 12-6, were employed to generate concise summaries, which were then distributed to users via email. The system was evaluated through information retrieval metrics and human assessments to ensure relevance and quality. The study highlights the challenges of processing Singlish due to its informal variations and lack of language resources and sets a foundation for future improvements in chat summarization for low-resource languages. The developed solution not only enhances user productivity but also contributes to the broader field of localized NLP research. |

*Keywords: Extractive summarization, Abstractive summarization, Chat Summarization, Chat Data Preprocessing*

1. INTRODUCTION

In today’s hyper-connected world, communication has shifted rapidly from traditional voice calls and SMS to real-time messaging platforms such as WhatsApp, Telegram, and Facebook Messenger. Among these, WhatsApp stands out as one of the most popular applications globally (Yuzbasioglu et al., 2020), including in Sri Lanka, due to its ease of use and accessibility (Statista, 2024). The rise of WhatsApp group chats, particularly in educational, professional, and social contexts, has enabled seamless interaction among large communities. However, this growing convenience has introduced a new challenge—information overload.

In large group chats, hundreds or even thousands of messages may be exchanged daily. Users who check messages infrequently often miss important updates, leading to confusion, reduced productivity, and communication gaps. This is especially problematic in Sri Lankan WhatsApp groups, where most conversations are written in “Singlish”—Sinhala language typed using the English alphabet. Singlish is an informal, non-standardized, code-mixed language that lacks consistent spelling rules or linguistic resources (Kurera et al., 2022), making it difficult for conventional Natural Language Processing (NLP) models to interpret and summarize accurately.

While there has been significant progress in the development of NLP tools for structured languages like English (Fayed & Ghantous, 2025), tools that can handle the irregularities of Singlish remain underdeveloped. Current summarization systems trained in high-resource languages struggle with Singlish text due to the informal structure, phonetic spelling variations, and limited annotated datasets. Additionally, Singlish chat content often includes short utterances, emojis, and slang, further complicating automated analysis.

This research addresses the gap by proposing a customized NLP pipeline capable of summarizing WhatsApp group chats typed in Singlish. The system focuses on extracting daily summaries from chat exports, using preprocessing techniques tailored for Singlish, and applying transformer-based models such as DistilBART. The aim is to deliver concise, relevant summaries that reduce the time and cognitive load for users.

By focusing on a low-resource, culturally specific language variant, this study contributes to the growing field of localized NLP research and provides a foundation for future work in summarizing informal, mixed-language digital communication.

**1.1 Related Works**

Automatic text summarization has been a significant research area in Natural Language Processing (NLP), aiming to reduce large bodies of text into concise and meaningful summaries. Numerous studies have introduced summarization techniques, primarily focused on English and other high-resource languages, but there is a noticeable gap in work concerning code-mixed or low-resource languages like Singlish—Sinhala expressed using the English alphabet.

Several efforts have been made to address this gap from different perspectives. One such study explored the development of NLP tools for processing Sinhala text typed in English letters, focusing on translation and summarization by identifying slang, informal structures, and phonetic variations (Kurera et al., 2022). Another relevant work investigated automatic summarization of Sinhala government gazettes using a hybrid approach that combined extractive and abstractive techniques, demonstrating the potential of such models in low-resource linguistic environments (Asal et al., 2017; Jayawardane, 2022).

Unsupervised keyword extraction has also played a crucial role in aiding summarization. Techniques like YAKE! (Yet Another Keyword Extractor) (Campos et al., 2020) and TextRank (Mihalcea & Tarau, 2004) have shown effectiveness in identifying meaningful content from raw text without requiring large, annotated datasets. These models are domain independent and have proven useful in various summarization and information retrieval tasks.

Recent research has extended into the realm of chat and conversational summarization. Studies have examined group sentiment, misinformation patterns, and chat summarization using deep learning models and transformer-based architectures (Resende et al., 2019; Mihalcea & Tarau, 2004). Specifically, RankAE, an unsupervised autoencoder-based model, was introduced to summarize chat logs by ranking utterances based on centrality and diversity (Zou et al., 2021). Similarly, large language models (LLMs) such as BART and DistilBART have been leveraged for abstractive summarization, primarily in the news domain, with growing interest in adapting them for conversational data (Zhang et al., 2023).

Despite these advancements, there remains a significant lack of attention to ward Singlish—a highly informal, non-standardized (Kurera et al., 2022), and code-mixed writing style. Its phonetic inconsistencies, lack of grammar rules, and mixed vocabulary make traditional summarization methods less effective. Moreover, the absence of labeled datasets and pre-trained language models tailored to Singlish limits the applicability of existing NLP solutions.

This research seeks to bridge that gap by adapting established summarization techniques to handle Singlish WhatsApp group chats. By focusing on a linguistically underrepresented yet widely used communication style, the study contributes to the growing body of work on localized NLP and presents an early step toward developing effective tools for informal, low-resource language environments.

2. methodology

This study presents a structured pipeline for summarizing WhatsApp group chats typed in Singlish, an informal variant of Sinhala using the English alphabet. The methodology comprises several key stages, from data collection and preprocessing to summarization, evaluation, and distribution.

**2.1 Data Collection**

To study the automatic summarization of Singlish group chats, several Singlish WhatsApp group chats were used to collect the relevant data. We collected over 15 group chats with more than 2500 words each, and 8 groups were selected (see Table 1) to export the chats for research purposes. Groups with excessive off-topic messages, prolonged inactivity, or inconsistent linguistic patterns were excluded to maintain focus on coherent conversational flows that support meaningful summarization. Rather than exporting chats daily, which could have resulted in hundreds of small and fragmented files, the export strategy was designed to focus on periods of high activity within the chat groups. Each exported .txt file typically represented a continuous 1 to 3-day span of meaningful conversation. This method ensured that the files preserved contextually rich and dense segments of chat data suitable for summarization. Over the course of one month, each of the 8 selected groups contributed 2–3 high-quality exports, resulting in a balanced dataset for analysis and model training.

The collected group chat data was stored in .txt file format on the device used to develop the model. WhatsApp group chats were exported using the platform’s built-in export feature, selecting the “without media” option to retain only textual content (see Figure 1). The exported ‘.txt‘ files included each message’s timestamp, sender name, and textual content. Media elements such as images, videos, audio messages, and emojis were intentionally excluded. Ethical considerations were upheld throughout: participants provided informed consent, and any user who chose to withdraw had their data completely omitted.

**Table 1: Used dataset for research**

|  |  |
| --- | --- |
| Number of WhatsApp groups | 8 |
| Number of days selected for summarization | 100 |
| Total number of exported words | 35642 |
| Average number of chats per group | 4826 |
| Average number of words per sentence | 3 |
| Total number of distinct words | 3424 |



**Figure 1: Structure of a WhatsApp chat. A screenshot taken on 02 February 2024**

**2.2 Data Preprocessing**

Preprocessing involved multiple steps to clean and prepare the text [6]:

* **Regex Filtering**: Regular expressions were used to remove timestamps, emojis, metadata, and other non-textual noise.
* **Normalization**: Singlish words with varied spellings (e.g., kanawada, knwada, kanawd) were unified to a standard form.
* **Tokenization**: The NLTK library was used to segment text into sentences and words, helping identify language patterns.
* **Stopword Removal**: A combined list of English and custom Singlish stop words was applied to remove low-value terms.

**2.3 Tools and Technologies**

The summarization system was built using Python due to its flexibility and vast NLP ecosystem. Libraries included Pandas and NumPy for data handling, Regex for pattern matching, and NLTK for tokenization and stopword filtering (See Figure 2). YAKE! (Campos et al., 2020) and TextRank were used for keyword extraction to help focus summarization on high-salience content (Mihalcea & Tarau, 2004).



**Figure 2: Overview of the Singlish WhatsApp Chat Summarization System**

**2.4 Tools and Technologies**

Transformer-based models were explored for the summarization task. DistilBART CNN-12-6 was selected for its speed and efficiency. As a distilled version of BART trained on the CNN/DailyMail dataset, it provides near state-of-the-art performance with reduced inference time. Despite being fine-tuned on news data, its summarization capabilities were promising for conversational data when aided by keyword extraction.

**2.5 Evaluation Metrics**

To assess the effectiveness of the proposed Singlish WhatsApp chat summarization system, two complementary evaluation strategies were employed: (1) ROUGE-based quantitative evaluation and (2) manual evaluation using Precision, Recall, and F1-score. These methods allowed for a comprehensive understanding of both content overlap and sentence-level accuracy when compared to human-written summaries.

**2.6 ROUGE Evaluation**

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric was used to quantitatively compare the model-generated summaries against manually written summaries [13]. Specifically, ROUGE-N (for unigram and bigram overlap) and ROUGE-S (skip-bigram) were considered. Due to the informal and code-mixed nature of Singlish chats, ROUGE scores alone were not sufficient, so a simplified calculation was introduced to measure content reduction:

$$Words\left(O-HS\right)=\frac{Number of words in Human Summary}{Total words in original chat} (1)$$

$$Words\left(O-MS\right)=\frac{Number of words in Machine Summary}{Total words in original chat} (2)$$

$$ROUGH=\frac{Words(O-MS)}{Words(O-HS)} \left(3\right)$$

An acceptable summary was considered to have a word count between 33% and 66% of the original content. This heuristic helped identify whether the generated summaries were proportionally concise.

**2.7 Manual Evaluation: Precision, Recall, and F1-Score**

In addition to ROUGE, a manual evaluation was conducted to assess the relevance and coverage of the summaries (Yuzbasioglu et al., 2020). A sentence-level comparison between machine-generated summaries and human-written ground truth summaries was performed. The following metrics were calculated:

$$Precision=\frac{Number of relevant sentences in machine summary}{Total sentences in machine summary} (4)$$

$$Recall=\frac{Number of relevant sentences in machine summary}{Total sentences in human summary} \left(5\right)$$

$$F1 score=2·\frac{Precision · Recall}{Precision + Recall} (6)$$

**2.8 Distribution of Summaries**

After generating summaries, they were embedded into a simple email-based distribution system. Group members received daily summarized chat updates, enabling users to stay informed without reading full conversation threads. This step validates the practical use of the summarization pipeline.

3. results and discussion

This section presents the performance of the proposed Singlish WhatsApp Chat Summarization system based on two evaluation approaches: manual F1-score comparison and automated ROUGE analysis. The experiments were conducted across multiple WhatsApp group chat datasets representing various topics, participant behaviors, and message volumes

* 1. **F1 Score Evaluation**

A total of 10 chat exports were manually evaluated to compare machine-generated summaries with human-written references. These 10 were selected from the full dataset to represent a range of linguistic challenges, including varying levels of Singlish complexity, message volume, and conversation focus. The selection also considered manageability for manual evaluation, as each summary required sentence-level analysis and annotation by human reviewers. The F1-score was computed for each file based on sentence-level precision and recall. As shown in Table 2, the F1 scores range from 0.45 to 0.71. The highest F1-score of 0.71 was recorded for Bodima chat13.txt, where the conversation was topic-focused and linguistically clean. Lower scores (e.g., 0.45 in App chat3.txt) were observed in chats with inconsistent spelling, frequent code-mixing, and noisy data. The average F1-score across all evaluated files was approximately 0.59, indicating that the summarizer captured a meaningful portion of relevant content despite the complexity of Singlish.

**Table 2: F1 Score Evaluation of Machine Summaries Compared to Human Summaries**

|  |  |
| --- | --- |
| **Sample chat File** | **F Score** |
| Trip\_chat2.txt | 0.64 |
| Trip\_chat3.txt | 0.53 |
| Trip\_chat22.txt | 054 |
| Bodima\_chat2.txt | 0.58 |
| Bodima\_chat12.txt | 0.63 |
| Bodima\_chat13.txt | 0.71 |
| App\_chat2.txt | 0.47 |
| App\_chat3.txt | 0.45 |

The results indicate that while some summaries closely matched human written content, others struggled with informal or noisy messages. Overall, the hybrid evaluation approach confirmed that the proposed method is capable of generating concise and contextually relevant summaries for Singlish WhatsApp chats.

* 1. **ROUGE-Based Evaluation**

To complement the manual evaluation, ROUGE-N and ROUGE-S metrics were used to assess lexical overlaps between machine and human summaries. The ROUGE ratio was calculated using a simplified heuristic involving word count compression, where acceptable summaries were defined as being 33–66% of the original message length.

* The majority of summaries fell within the acceptable compression threshold.
* Chat files with higher F1-scores also exhibited balanced ROUGE ratios, suggesting consistency between sentence-level relevance and lexical similarity.
* No cases of over-summarization (e.g., summaries shorter than 20%) or under summarization (e.g., summaries longer than 75%) were found.
	1. **Observations**

The results demonstrate that the proposed system is capable of generating concise and contextually relevant summaries for Singlish chats. Performance was strongest when:

* Chats remained focused on a specific topic
* Fewer emojis or media placeholders were present
* Participants used relatively consistent spelling

Conversely, performance degraded in the presence of highly informal, off topic, or extremely abbreviated language, which further emphasizes the challenges of working with low-resource and unstandardized text like Singlish.

4. Conclusion

This research introduced a novel NLP-based system designed to summarize WhatsApp group chats typed in Singlish, a non-standard, code-mixed language commonly used in Sri Lanka. By leveraging transformer-based models such as DistilBART and implementing tailored preprocessing steps, including normalization of Singlish variants and keyword extraction, the system effectively reduced message volume while preserving conversational context. Evaluation using both ROUGE metrics and F1-score comparisons against human-written summaries demonstrated that the proposed approach achieved reasonably high performance despite linguistic irregularities and the absence of domain-specific training data.

While the results are encouraging, several areas remain for future improvement. A primary direction is the development of a dedicated Singlish corpus to fine-tune summarization models for this unique linguistic style. Incorporating real-time summarization capabilities into messaging platforms and enabling users to control the level of compression could further enhance usability. Moreover, future evaluations may benefit from integrating deeper semantic metrics such as BERT Score and involving human annotators for qualitative feedback. These extensions would bring the system closer to practical deployment and contribute to advancing NLP for low-resource, informal languages.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper

Disclaimer (Artificial intelligence)

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1. ChatGPT (GPT-3.5)

2.

3.

References

Statista Research Department. (Apr. 2024). Number of monthly active WhatsApp users worldwide from April 2013 to April 2024. <https://www.statista.com/statistics/260819/number-of-monthly-active-whatsapp-users/>

Kurera, W. S. A., Rajapaksha, R. K., Rupasinghe, H. A. P., Liyanage, K. L. D. U. B., Kasthuriarachchi, S., & Rajapakshe, S. (2022). Building NLP tools to process Sinhala text data written using English letters. In Proceedings of the 22nd International Conference on Advances in ICT for Emerging Regions (ICTer) (pp. 63–68). IEEE. <https://doi.org/10.1109/ICTer58063.2022.10024080>

Jayawardane, H. M. R. Y. (2022). Automatic Sinhala text summarization for government gazettes using abstractive and extractive methods (B.Sc. thesis). University of Colombo School of Computing. <https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4609>

Fayed, K., & Ghantous, M. (2025). An NLP-based WhatsApp group chat summarizer. In A. Abdelgawad, A. Jamil, & A. A. Hameed (Eds.), Intelligent Systems, Blockchain, and Communication Technologies (pp. 316–329). Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-82377-0_27>

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 509, 257–289. <https://doi.org/10.1016/j.ins.2019.09.013>

You, X. (2019). Automatic summarization and keyword extraction from web page or text file. In Proceedings of the 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET) (pp. 154–158). IEEE. <https://doi.org/10.1109/CCET48361.2019.8989315>

Joshi, S. (2019). Sentiment analysis on WhatsApp group chat using R. In R. K. Shukla, J. Agrawal, S. Sharma, & G. S. Tomer (Eds.), Data, Engineering and Applications (pp. 47–55). Springer. <https://doi.org/10.1007/978-981-13-6347-4_5>

Ramos, J. G., Navarro-Alatorre, I., Becerra, G. F., & Flores-Sánchez, O. (2019). A formal technique for text summarization from web pages by using latent semantic analysis. Presented at Tecnológico Nacional de México, Instituto Tecnológico de La Piedad, Michoacán, Mexico.

Mishra, N., Sahu, G., Calixto, I., Abu-Hanna, A., & Laradji, I. H. (2023). LLM aided semi-supervision for extractive dialog summarization. arXiv preprint, arXiv:2311.11462. <https://arxiv.org/abs/2311.11462>

Yuzbasioglu, H. B., Helvaci, S. C., Cevik, E. E., & Kurnaz, M. A. (2020). Examination of the conversations in a WhatsApp group created for communication in a bachelor’s level astronomy course. International Journal of Education in Mathematics, Science and Technology, 8(2), 168. <https://doi.org/10.46328/ijemst.v8i2.674>

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. arXiv preprint, arXiv:2301.13848. <https://arxiv.org/abs/2301.13848>

Zou, Y., et al. (2021). Unsupervised summarization for chat logs with topic-oriented ranking and context-aware auto-encoders. Proceedings of the AAAI Conference on Artificial Intelligence, 35(16), 14674–14682. <https://doi.org/10.1609/aaai.v35i16.17724>

Asal, A., et al. (2017). A comprehensive survey on extractive text summarization. American Journal of Engineering Research (AJER), 6(1), 226–230. [http://www.ajer.org/papers/v6(01)/ZH0601226239.pdf](http://www.ajer.org/papers/v6%2801%29/ZH0601226239.pdf)

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 85–92).