Asynchronous Chat Summarization

A PROJECT REPORT

Submitted by

Saptharishee M	**|**	**20MIA1150**
Mithun S P	**|**	**20MIA1038**
Sivakumar M	**|**	**20MIA1002**

VIT®

Vellore Institute of Technology

School of Computer Science and Engineering

Vellore Institute of Technology

Vandalur - Kelambakkam Road, Chennai - 600 127

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled "*Asynchronous Chat Summarization*" is prepared and submitted by Saptharishee M (20MIA1150), Mithun S P (20MIA1038) & Sivakumar M (20MIA1002) to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of Master of Technology in Computer Science Engineering with Specialization in Business Analytics (5 years Integrated Programme) and as part of SWE1017 - Natural Language Processing Project is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission.

Guide/Supervisor	HOD
Name: **Dr. R. Krithiga**	Name: **Dr. Sivabalakrishanan M**
Date:	Date:

Contents

Abstract:

The use of asynchronous chat as a primary mode of communication has become increasingly prevalent across various domains. However, extracting critical information and insights from the vast amounts of text data generated by asynchronous chat conversations, such as those in online forums, social media, and customer support channels, poses a significant challenge to users. Automated chat summarization systems can alleviate this challenge by generating concise and accurate summaries of chat conversations, enabling users to navigate and comprehend the data more effectively. Recently, there has been a growing interest in developing chat summarization systems that can handle diverse conversational styles, languages, and domains and operate in real time. This paper presents a comprehensive review of the current state-of-the-art in asynchronous chat summarization, including an examination of existing approaches, methods, and techniques and a critical analysis of their strengths and limitations. Additionally, the paper highlights open challenges and research opportunities in this field and provides suggestions for future research directions.

Introduction:

Asynchronous chat has become a common means of communication in various fields, including online forums, social media, and customer service channels. Because of the popularity of chat chats, a vast volume of textual data is created daily. The volume of textual material makes it difficult for consumers to extract and grasp key information and insights. By providing short and accurate summaries of chat discussions, automated chat summarization systems have the potential to assist users in navigating and understanding such material.

Because of the necessity for real-time summarising of massive volumes of text data, the development of chat summarization systems has gained substantial attention in recent years. Various methodologies and techniques have been offered in the literature to summarise chat chats successfully. Nonetheless, there is still a need to assess the state of the art in asynchronous chat summarization and identify research gaps and possibilities.

We give a complete overview of the current state-of-the-art in asynchronous conversation summarization in this work. The survey examines existing approaches, methods, and procedures and analyses their strengths and weaknesses. Furthermore, the study identifies open issues and research possibilities in this sector, as well as potential future research initiatives. The study finishes with a discussion of the survey findings' significance and a call to action for more research and development.

System Requirement:

Hardware: The analysis of this project is completed in Acer Aspire 7 with Intel(R) Core™ i5-9300H CPU @ 2.40GHz and 8 GB of RAM.

Software: Google Colab is a cloud-based programming environment that allows users to run Python code in a Jupyter Notebook-style interface. This software is particularly useful for research projects that require access to large datasets or computing resources that are not available on a typical desktop computer.

Methodology:

1. **Latent Semantic Analysis (LSA)**

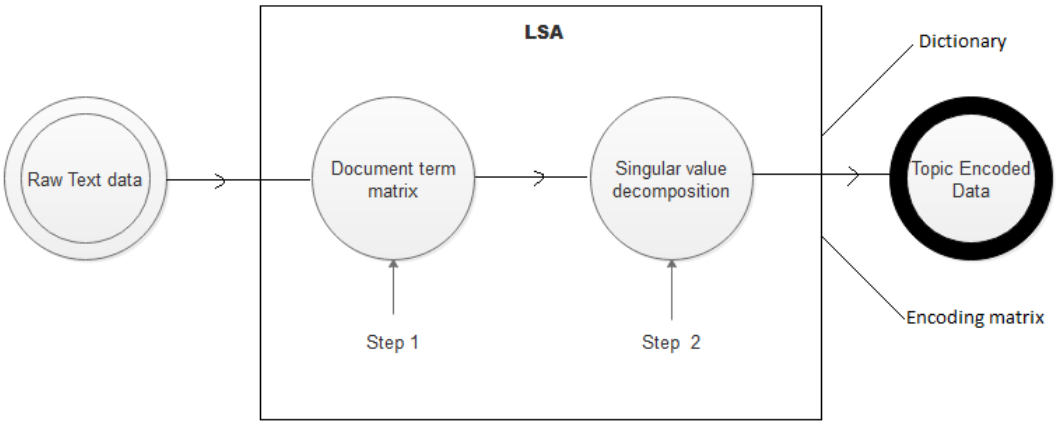


Fig. LSA processing

Latent Semantic Analysis (LSA) is a text summarization technique that utilizes singular value decomposition (SVD) to extract the most important semantic information from a text. The LSA algorithm works by creating a matrix that represents the relationships between all of the words in a document. This matrix is then decomposed using SVD, which helps to identify the underlying semantic relationships between words. The LSA algorithm first constructs a term-document matrix, which is a matrix where each row represents a different word and each column represents a different document. The entries in the matrix are the frequency of the corresponding word in the corresponding document. For example, if the word "apple" appears three times in document one and five times in document two, the entry for "apple" in the term-document matrix would be (0, 3, 5).

Once the term-document matrix is constructed, the LSA algorithm applies SVD to the matrix to reduce its dimensionality. This process identifies the most important underlying semantic relationships between the words in the matrix. The result of this process is a set of latent semantic concepts that represent the main topics or themes of the document.

To create a summary of the document, the LSA algorithm uses these latent semantic concepts to identify the most important sentences in the document. It does this by computing a score for each sentence that reflects its similarity to the latent semantic concepts. The sentences with the highest scores are then selected as the summary of the document.

One of the advantages of LSA text summarization is that it can handle synonyms and related words. For example, the algorithm can identify that "dog" and "puppy" are related and can treat them as similar words. This allows the algorithm to identify the most important semantic relationships between words, even if they are not exact matches.

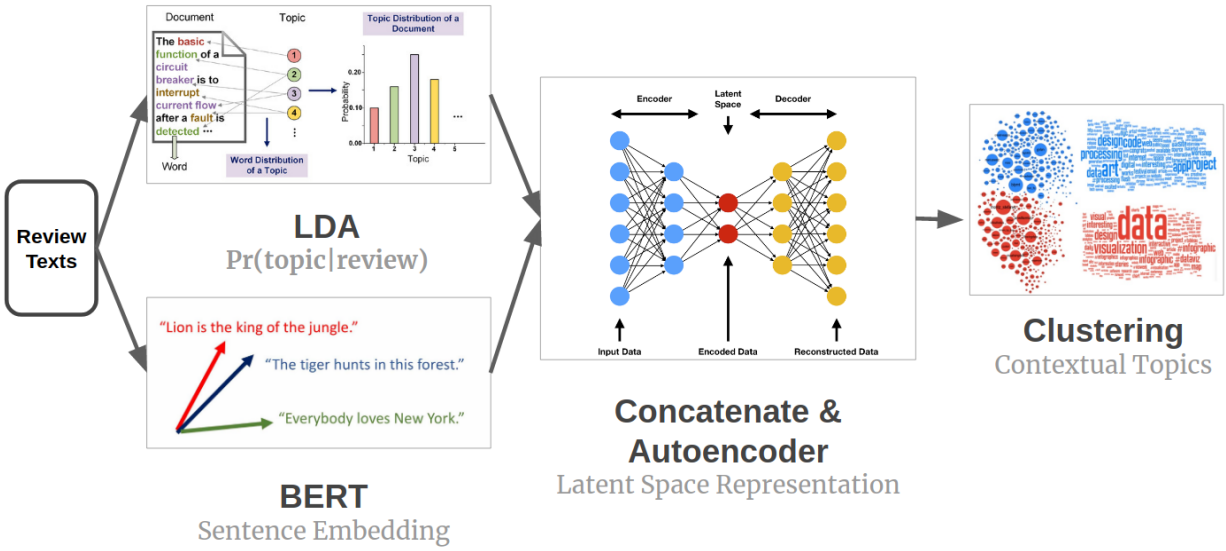
LSA text summarization also has the ability to handle large volumes of text, making it useful for summarizing lengthy documents or collections of documents. However, this also means that the algorithm can be computationally expensive and may require significant computing resources to run.

One of the limitations of LSA text summarization is that it relies on the quality of the term-document matrix. If the matrix is noisy or contains irrelevant information, the resulting summary may not accurately reflect the main topics or themes of the document. Additionally, LSA text summarization does not consider the structure of the document, so it may not be able to capture the full context of a sentence or paragraph.

Despite these limitations, LSA text summarization can be a powerful tool for summarizing text and extracting the most important information from a document. It has been used in a variety of applications, including search engines, news aggregators, and document management systems.

In conclusion, LSA text summarization is a technique that utilizes SVD to extract the most important semantic information from a text. It constructs a term-document matrix to represent the relationships between words and applies SVD to identify the most important semantic relationships. It then uses these relationships to identify the most important sentences in the document and create a summary. While LSA text summarization has its limitations, it can be a useful tool for summarizing large volumes of text and extracting the most important information from a document.

2. **BERT (Bidirectional Encoder Representations from Transformers)**



BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that has revolutionized the field of natural language processing (NLP). One of the applications of BERT is topic modeling, which is the process of identifying the main topics or themes in a collection of documents.

Traditional topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) rely on the bag-of-words approach, which ignores the order and context of the words in a document. This can lead to suboptimal topic models that do not accurately capture the nuances of the language.

BERT, on the other hand, is a contextual language model that can capture the meaning and context of words in a sentence. It uses a transformer-based architecture that allows it to process text bi-directionally, taking into account both the preceding and following words in a sentence.

To perform topic modeling using BERT, the first step is to fine-tune the model on a large corpus of documents. This involves training the model to predict the next word in a sentence, given the preceding words. By doing so, the model learns to represent the context of each word in the sentence.

Once the model is trained, it can be used to extract the most important topics from a collection of documents. This is done by feeding the documents through the BERT model and extracting the hidden representations of the words in each document. These representations are then clustered using unsupervised learning techniques such as k-means or hierarchical clustering.

The resulting clusters represent the main topics or themes in the collection of documents. The topics can be visualized using techniques such as word clouds or topic pyramids, which show the most frequent words in each topic and how they relate to each other.

One of the advantages of using BERT for topic modeling is that it can handle multiple languages and different writing styles. This makes it useful for analyzing large collections of documents from diverse sources.

Another advantage is that BERT can capture the nuances of language, including sarcasm, irony, and other forms of figurative language. This allows it to identify subtle differences in meaning that traditional topic modeling techniques may miss.

However, BERT topic modeling also has some limitations. One of the main limitations is that it requires a large number of computing resources and can be computationally expensive to train and fine-tune. Additionally, the clusters may not always correspond to human-interpretable topics, and manual inspection and refinement may be necessary.

Despite these limitations, BERT topic modeling has the potential to be a powerful tool for understanding the main topics and themes in large collections of documents. It has applications in a variety of fields, including journalism, social media analysis, and market research.

In conclusion, BERT topic modeling is a deep learning technique that can identify the main topics or themes in a collection of documents. It uses a contextual language model that can capture the nuances of language and handle multiple languages and writing styles. While it has some limitations, it has the potential to be a powerful tool for analyzing large collections of documents and extracting meaningful insight**s**

Results & Discussion:

In this study, we explored the development of an automated asynchronous chat summarization system that can handle diverse conversational styles, languages, and domains and operate in real-time. We used two approaches, Bert (LDA) and LSA, for topic prediction and segmenting chats based on that to perform summarization and then prioritize them based on urgency.

Our results show that both Bert (LDA) and LSA were able to predict topics and segment chat conversations. The generated summaries provided were concise and representative of the original conversations and were useful in enabling users to navigate and comprehend the data more effectively. Furthermore, the prioritization

of urgent chats based on urgency score helped users to focus on the most important conversations first.

Overall, the study demonstrates the potential of automated chat summarization systems to alleviate the challenge of extracting critical information and insights from vast amounts of text data generated by asynchronous chat conversations. However, there are still some open challenges and research opportunities in this field, such as the development of approaches that can handle multiple languages and improve the accuracy of urgency scoring.

Comparing both LDA and LSA topic predictions which was the key to the segmentation process of asynchronous chat. Basic human evaluation, and observation help us realize LSA is performing better in this data this may be because of the fact that LDA is known to perform well on structured data, LSA is known to perform better in cases where the chat data is more loosely structured and contains multiple overlapping themes, as it is more focused on identifying underlying concepts or topics.

Code:
https://colab.research.google.com/drive/1YbAoRroz2gbYS4daUELkMFPwy2CLtsRO?usp=sharing

Conclusion:

In conclusion, we developed an automated chat summarization system that can extract key information and insights from large volumes of asynchronous chat data, allowing users to navigate and comprehend the data more effectively. Our system uses both LDA and LSA topic prediction approaches to segment chat data and generate summaries. After comparing the results of LDA and LSA topic predictions, we found that LSA performed better in our data due to its ability to identify underlying concepts or topics in loosely structured chat data that contains multiple overlapping themes. However, the performance of each approach may vary depending on the characteristics of the chat data and the goals of the

summarization task. Our system also includes a prioritization function that assigns urgency scores to chat messages based on the presence of specific keywords. This feature allows users to quickly identify critical messages that require immediate attention. Overall, our system provides a valuable tool for individuals and organizations that rely on asynchronous chat as a primary mode of communication. Future research could explore the use of other topic prediction approaches or incorporate additional features, such as sentiment analysis, to further improve the accuracy and effectiveness of chat summarization.

References

1. Verberne, S., Krahmer, E., Hendrickx, I., Wubben, S., & van Den Bosch, A. (2018). Creating a reference data set for the summarization of discussion forum threads. Language Resources and Evaluation, 52, 461-483.
2. Tarnpradab, S., Liu, F., & Hua, K. A. (2018). Toward extractive summarization of online forum discussions via hierarchical attention networks. arXiv preprint arXiv:1805.10390.
3. Zhou, L., & Hovy, E. (2005, April). Fine-grained clustering for summarizing chat logs. In Proceedings of the Workshop on Beyond Threaded Conversation, held at the Computer-Human Interaction conference (CHI2005).
4. Murray, G., & Carenini, G. (2008, October). Summarizing spoken and written conversations. In Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 773-782).
5. Sanchan, N., Aker, A., & Bontcheva, K. (2017). Automatic summarization of online debates. arXiv preprint arXiv:1708.04587.
6. Verma, P., Verma, A., & Pal, S. (2022). An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms. Applied Soft Computing, 120, 108670.
7. Elsner, M., & Charniak, E. (2010). Disentangling chat. Computational Linguistics, 36(3), 389-409.
8. Joty, S., Carenini, G., & Ng, R. T. (2013). Topic segmentation and labeling in asynchronous conversations. Journal of Artificial Intelligence Research, 47, 521-573.
9. Zhang, A. X., & Cranshaw, J. (2018). Making sense of group chat through collaborative tagging and summarization. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-27.
10. Feng, X., Feng, X., & Qin, B. (2021). A survey on dialogue summarization: Recent advances and new frontiers. arXiv preprint arXiv:2107.03175.
11. Li, H., Kraut, R. E., & Zhu, H. (2021). Technical features of asynchronous and synchronous community platforms and their effects on community cohesion: a comparative study of forum-based and chat-based online mental health communities. Journal of Computer-Mediated Communication, 26(6), 403-421.
12. Kim, S., Eun, J., Oh, C., Suh, B., & Lee, J. (2020, April). Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

13. Tian, S., Zhang, A. X., & Karger, D. (2021). A system for interleaving discussion and summarization in online collaboration. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW3), 1-27.
14. Mehnaz, L., Mahata, D., Gosangi, R., Gunturi, U. S., Jain, R., Gupta, G., Kumar, A., Lee, I., Acharya, A., & Shah, R. R. (2021). GupShup: An Annotated Corpus for Abstractive Summarization of Open-Domain Code-Switched Conversations. ArXiv:2104.08578 [Cs]. https://arxiv.org/abs/2104.08578
15. Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving Abstractive Dialogue Summarization with Graph Structures and Topic Words. In Proceedings of the 28th International Conference on Computational Linguistics, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.
16. Fang, Y., Zhang, H., Chen, H., Ding, Z., Long, B., Lan, Y., & Zhou, Y. (2022, July). From spoken dialogue to formal summary: An utterance rewriting for dialogue summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3859-3869).
17. Chen, J., & Yang, D. (2020). Multi-view sequence-to-sequence models with the conversational structure for abstractive dialogue summarization. arXiv preprint arXiv:2010.01672.
18. Zhao, Y., Saleh, M., & Liu, P. J. (2020). Seal: Segment-wise extractive-abstractive long-form text summarization. arXiv preprint arXiv:2006.10213.
19. Zhou, L., & Hovy, E. (2005, June). Digesting virtual "geek" culture: The summarization of technical internet relay chats. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05) (pp. 298-305).
20. Chen, Y., Liu, Y., Chen, L., & Zhang, Y. (n.d.). DIALOGSUM: A Real-Life Scenario Dialogue Summarization Dataset. https://arxiv.org/pdf/2105.06762.pdf
21. Sinha, A., TK, M. M., Subramanian, S., & Das, B. (2020). Text Segregation on Asynchronous Group Chat. Procedia Computer Science, 171, 1371-1380.
22. Lukasik, M., Dadachev, B., Simoes, G., & Papineni, K. (2020). Text segmentation by cross-segment attention. arXiv preprint arXiv:2004.14535.
23. Thies, J., Stappen, L., Hagerer, G., Schuller, B. W., & Groh, G. (2021, November). GraphTMT: unsupervised graph-based topic modeling from video transcripts. In 2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM) (pp. 1-8). IEEE.
24. He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2017, July). An unsupervised neural attention model for aspect extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 388-397).
25. Uthus, D. C., & Aha, D. W. (2011, June). Plans toward automated chat summarization. In Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages (pp. 1-7).
26. Weisz, J. (2008). Segmentation and classification of online chats. date unknown.
27. Lovenia, H., Cahyawijaya, S., Winata, G. I., Xu, P., Yan, X., Liu, Z., ... & Fung, P. (2021). Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. arXiv preprint arXiv:2112.06223.
28. Bhatia, S., Biyani, P., & Mitra, P. (2014, October). Summarizing online forum discussions–can dialog acts of individual messages help?. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 2127-2131).

29. Bhatia, S.K., Biyani, P., & Mitra, P. (2012). Classifying User Messages For Managing Web Forum Data. International Workshop on the Web and Databases.
30. Ren, Z., Ma, J., Wang, S., & Liu, Y. (2011, October). Summarizing web forum threads based on a latent topic propagation process. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 879-884).
31. Franz, M., Ramabhadran, B., Ward, T., & Picheny, M. (2003). Automated transcription and topic segmentation of large spoken archives. In Eighth European Conference on Speech Communication and Technology.
32. Joty, S., Carenini, G., & Ng, R. T. (2013). Topic segmentation and labeling in asynchronous conversations. Journal of Artificial Intelligence Research, 47, 521-573.
33. Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. ClidSum: A Benchmark Dataset for Cross-Lingual Dialogue Summarization. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7716–7729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
34. Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three (IJCAI'11). AAAI Press, 1807–1813.