

## LIMITATIONS

### Limitations

Limitations associated with the approach documented in this dissertation can be classified into four distinct categories: technical, data, model, and financial. Technical limitations include compute requirements for processing libraries required for sentiment analysis and accuracy measurement, the reliance on Python code which even when provided in complete and ready-to-execute form requires some basic understanding of computer programming and the Python language, the use of open-source libraries like BERTopic and BERTScore which are not guaranteed to have long term support, and the use of AutoTrain from Hugging Face which is a trade-off between optimal model performance tuning which requires deep machine learning knowledge and the convenience of generating a good model without that expertise.

Data limitations include the availability of data to extract from either push shift or academic torrents and the population represented by the data, which is solely limited to those who choose to engage in the online community. Any insights regarding a community such as TFCC are limited only to contact center agents who choose to post on this forum and may not be generalizable across all contact center agents. Furthermore, subreddit data in particular is primarily posts from English-speaking countries with 71% of Reddit traffic originating from the United States (51.5%), the United Kingdom (8.4%), Canada (8.3%), and Australia (4.5%) (d\_mystery, 2021).

Model limitations include hallucination and input/output sequence size limitations. As demonstrated in this dissertation, the readability and apparently high quality of abstractive summarizations generated from LLMs are extremely high. The readability of the summaries can be misleading however as the phenomenon of hallucination from LLMs is well documented and, as of the time of writing, there are no accuracy measures available to reliably identify hallucination. Any insights gained from the outputs generated by LLMs should be validated using quantitative methods. The input/output sequence size limitation puts a cap on the size of the data to be summarized. For an online community like a subreddit, this limitation is not insurmountable as most submissions fell within the token restriction selected, but for other scenarios such as abstractive summarization of interview transcripts or responses to open-ended questions in surveys, it may pose a problem. Overcoming the limitation is doable by breaking up data into smaller chunks and then re-combining, but this increases the complexity of the code.

Depending on the dataset, costs can become an issue. The costs associated with the use of the ChatGPT API for this dissertation were \$110.80. If the InstructGPT model from OpenAI were used instead of the ChatGPT API, those costs would have been \$1,108.00, and if the latest model from OpenAI, GPT-4 (OpenAI, 2023) (which was released during the writing of this dissertation) had been used the costs would have been \$3,324. In addition to API fees, there are costs associated with acquiring the equipment required to run libraries such as BERTScore and BERTopic in a reasonable amount of time. A laptop with 32GB of RAM and an 8GB GPU is recommended, which can cost \$3,000 - \$4,000.

Cloud-based Jupyter Notebook environments such as Google Colaboratory (<https://colab.research.google.com/>), Kaggle (<http://kaggle.com>), and Paperspace (<http://paperspace.com>) are potential alternatives but in certain cases, data privacy requirements may limit the use of these environments. The cloud-based alternatives for programming Jupyter Notebooks also require reliable internet connections. These cost limitations, when combined, create barriers for researchers in developing countries with limited financial resources.