

# WhatsApp Chat Summarization

Andrea Auletta

`andrea.auletta@studenti.unipd.it`

Marco Bernardi

`marco.bernardi.11@studenti.unipd.it`

Davide Baggio

`davide.baggio.1@studenti.unipd.it`

Marco Brigo

`marco.brigo@studenti.unipd.it`

Sebastiano Sanson

`sebastiano.sanson@studenti.unipd.it`

## Abstract

*This research project in Natural Language Processing (NLP) focuses on the development of an automated system for the summarization of chat messages, applicable to both group and individual conversations. The objective is to generate concise summaries of conversations, thereby obviating the need for users to review all messages individually. This application is especially pertinent in contexts where chat logs proliferate quickly, offering significant benefits in terms of time savings and cognitive load reduction for users.*

## 1. Introduction

In the era of digital communication, chat platforms have become fundamentals in our daily interactions, whether they are personal or professional. This new type of transmission favors real-time messages exchange which allows users to engage in conversations that can quickly accumulate extensive chat logs. The huge volume of messages, especially in active group chats formed by dozens if not hundreds of users, often makes it difficult for users to stay updated without dedicating substantial time to review each message. This challenge highlights the need for an efficient method that enables users to quickly extract essential information from chat conversations.

This research project addresses this need by developing an automated system for summarizing chat messages with the goal to generate concise summaries, reducing the need to sort through all messages. By doing so, users can quickly obtain the main points of discussions, significantly saving time and reducing cognitive load.

The significance of this application is evident in various contexts: for instance, in corporate environments, project teams often heavily rely on group chats to coordinate tasks,

share updates and exchange miscellaneous informations; similarly, social groups frequently engage in dynamic conversations. In both scenarios, an automated summarization system can enhance productivity and ensure important information is not overlooked.

This research aims to explore and implement state of the art NLP techniques to create an effective summarization system, focusing on the datasets used, the manipulation of them in order to obtain always a coherent data structure, how we obtained new examples to test the performances and the results we got after all.

TO DO: overview of our main results!

## 2. Datasets

In this section, we explore datasets essential for training and evaluating automated summarization systems for chat messages: an ideal dataset should have a variety of conversation styles and contexts to ensure the robustness and versatility of the summarization models. The datasets should contain real-life or realistic dialogues annotated with concise and accurate summaries. Additionally, the data should reflect diverse communication patterns, including informal and formal registers, and incorporate common conversational elements such as slang, emojis, eventual typos, the use of voice messages and images. We present below two notable datasets, SAMSum and DialogSum (both obtained by Hugging Face repository), which fulfill these criteria and provide rich resources for advancing the field of conversational AI.

### 2.1. SAMSum

The SAMSum [2] dataset contains about 16,000 messenger-like conversations with summaries splitted in:

- training: 14,732;
- validation: 818;

- test: 819.

Conversations were created and written down by linguists fluent in English. Linguists were asked to create conversations similar to those they write on a daily basis, reflecting the proportion of topics of their real-life messenger conversations. The style and register are diversified - conversations could be informal, semi-formal or formal, they may contain slang words, emoticons and typos. Then, the conversations were annotated with summaries. It was assumed that summaries should be a concise brief of what people talked about in the conversation in third person. The SAMSum dataset was prepared by Samsung R&D Institute Poland and is distributed for research purposes.

## 2.2. Dialogsum

DialogSum [1] is an extensive dialogue summarization dataset comprising 13,460 dialogues splitted in:

- training: 12,460;
- validation: 500;
- test: 1,500.

supplemented by an additional 100 holdout dialogues designated for topic generation, which contain the id, dialogue, and topic fields. Each dialogue is paired with manually annotated summaries and topics. The dataset is exclusively in English and includes four data fields: the text of the dialogue, a human-written summary of the dialogue, a human-written topic or one-liner of the dialogue, and a unique identifier for each example.

DialogSum distinguishes itself from previous datasets by incorporating dialogues under rich real-life scenarios, including a wider array of task-oriented contexts. The dialogues exhibit clear communication patterns and intents, making them suitable for summarization.

## 2.3. Custom Dataset

In order to obtain a dataset with an increased volume of data, a broader spectrum of conversation scenarios and an enhanced evaluation, we decided to concatenate the previously mentioned datasets into a bigger one. The splits of the new dataset are as follows:

- training: 27,192;
- validation: 1,318;
- test: 2,319.

To obtain it we had to preprocess the Dialogsum dataset and... TO DO

## 3. Models Involved

### 3.1. BART

TO DO

### 3.2. Florence-2 Large

To ensure that the images were also processed by the summarization model we had to find a way for converting them into a detailed description of their content. For this specific task we used Florence-2 [TO DO: CITE <https://arxiv.org/abs/2311.06242>] by Microsoft, which is an advanced vision foundation model that uses a prompt-based approach to handle a wide range of vision and vision-language tasks. It was designed to take text-prompt as task instructions and generate desirable results in text forms, whether it be captioning, object detection, grounding or segmentation. This model leverages on the FLD-5B dataset, containing 5.4 billion annotations across 126 million images, to master multi-task learning. The model's sequence-to-sequence architecture enables it to excel in both zero-shot and fine-tuned settings, proving to be a competitive vision foundation model.

### 3.3. Whisper

As for the images, we had also to convert the voice messages to text only. To do this task we used Whisper by OpenAI [TO DO: CITE <https://cdn.openai.com/papers/whisper.pdf>], an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The use of such a large and diverse dataset lead to improved robustness to accents, background noise and technical language, which is really helpful in our use case scenario since most of the times the voice messages are not in top notch quality. Moreover, it enables transcription in multiple languages, as well as translation from those languages into English.

## 4. Fine Tuning

## 5. Testing

## 6. Results

## References

- [1] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online, Aug. 2021. Association for Computational Linguistics.
- [2] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated

dialogue dataset for abstractive summarization. *CoRR*,  
abs/1911.12237, 2019.