# WhatsApp chat summarization

Andrea Auletta
andrea.auletta@studenti.unipd.it

Davide Baggio
davide.baggio@studenti.unipd.it

Marco Bernardi
marco.bernardi.11@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 group chat messages. 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

Please follow the guidelines that have been uploaded on Moodle.

## 2. Dataset

### 2.1. SAMSum

The SAMSum [2] dataset contains about 16k messenger-like conversations with summaries. 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 convesations. 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, supplemented by an additional 100 holdout dialogues designated for topic generation. 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. The data splits are as follows: 12,460 dialogues for training, 500 dialogues for validation, 1,500 dialogues for testing, and 100 holdout dialogues containing only the id, dialogue, and topic fields.

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

## 3. Models Involved

## 4. Fine Tuning

## 5. Testing

## 6. Results

### 6.1. Language

All manuscripts must be in English.

## 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.