EEE486 - Natural Language Processing **Group-11**

Name of The Project: Sentiment Analysis and Semantic Compression of Emails

Group Members:
- Harun Can Surav 21903452 - Övgüm Can Sezen 21902418 CS Department
- Doruk Karaman 21901507 - Berk Özçelik 21901424 EEE Department

Project Purpose:
Currently, communication systems turn texts with semantic meanings into bitwise encoded messages and compress them to send to the destination where they get decoded. However, as the Shannon Limit is being tested with current technology [2], it is essential to further optimize these messages. One idea is to use text summarizers and extenders [1] to compress and expand sent messages. The initial text can be summarized with semantic encoders while still holding its semantic value before going into the usual encryption-decryption process and later be extended with semantic decoders to end up with the same message but with less data transmitted in the process. We believe that a tool like this can be implemented to email services to reduce transmission time and increase the message sent per bit ratio.

Inspiration:
Topic suggests an improvement on an area which can be thought of as a crossing point of electronics and computer science, as a group of 2 EEE students and 2 CS students we find this fitting for us [4]. Also, it is always interesting to see how advances in software affect physical technology and vice-versa. As a field all four of us were excited in terms of the possible technical problems about the topic.

Task and Approach:
This project aims to reduce the total amount of bits(symbols) transmitted through emails. Similar to methodology proposed in [2], this project will implement an auto-encoder using transformer architecture. Upon encoding, it is planned to implement a sentimental analyzer to further protect the sentimental meaning. Messages will later be bitwise encoded to further reduce the transmitted package's size. After package transmission, messages will be bitwise decoded; followed by semantic decoding. After semantic decoding, messages will get processed by a transformer which will apply the sentiment of the initial message. To measure performance, a combination of BLEU, BERT and human testing will be conducted. It is important to clarify that the proposed methodology might change as further research is conducted in the survey report.

Literature Review:
Upon reading the suggested papers we became more informed about possible answers to and the existence of questions like "What is the problem definition of this topic?", "What are some proposed solutions to the defined question?", "Are there any solution approaches which may be considered best-practice?". Also, some information regarding the architecture of the system and the models proposed and possible problems regarding the field and technology relevant to the problem definition are also acquired. Further papers may be selected to guide us on the topics regarding the implementation like the models for encoding and decoding process and some general architectural decisions.

Data Description:

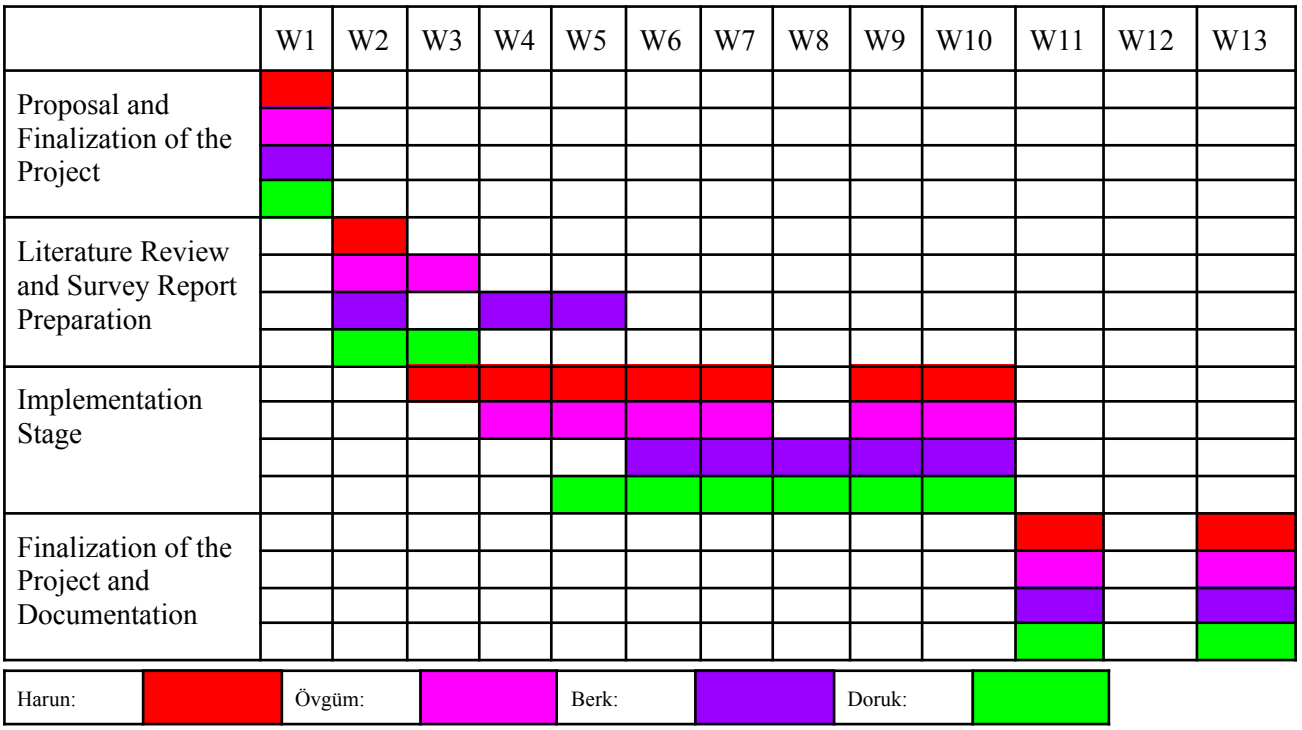
Dataset that is currently planned to be used is the internal mails of Enron Corporation, generated by the employees [5]. It consists of 517.401 mails in .csv format. The layout of the dataset is as follows:

- file (string): refers to the name of the sent file
- messages (string): contains the sent message with technical identifiers and content

The reason behind why this data set is chosen is that it is harder to analyze informal messages, so using business mails decreases the possibility of errors. In addition, since these messages were sent in a business environment, it is safe to assume that they are grammatically correct, which further decreases the amount of pre-processing that is required. The dataset is currently tentative, but the general characteristics of any chosen set shall be consistent with this one.

Gantt Chart:

Table 1: Gantt Chart (Tentative due to school semester not being definite)



References:

[1] H. Xie and Z. Qin, “A Lite Distributed Semantic Communication System for Internet of Things,” IEEE JSAC, vol. 39, no. 1, Jan. 2021, pp. 142–53.

[2] H. Xie, Z. Qin, G. Y. Li and B. -H. Juang, "Deep Learning Enabled Semantic Communication Systems," in IEEE Transactions on Signal Processing, vol. 69, pp. 2663-2675, 2021

[3] Z. Weng, Z. Qin and G. Y. Li, "Semantic Communications for Speech Signals," ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6

[4] X. Luo, H. -H. Chen and Q. Guo, "Semantic Communications: Overview, Open Issues, and Future Research Directions," in IEEE Wireless Communications, vol. 29, no. 1, pp. 210-219, February 2022

[5] W. Cukierski, “The Enron email dataset,” Kaggle, 16-Jun-2016. [Online]. Available: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset. [Accessed: 03-Mar-2023].