University of Ljubljana
Faculty of Computer and
Information Science

FRI Natural language processing course 2021
Project report

Qualitative Research on Discussions - text categorization

Krištof Zupan, Gal Menaše, and Nejc Mušič

Abstract

Advisors: Slavko Žitnik

Introduction

Qualitative discourse analysis is a method used by social scientists for studying human interactions within textual data, involving understanding the underlying meanings, context, and perspectives expressed within conversations or written communication. Since this task is demanding for human ”coders,” we will replace them with the use of large language models. Our goal is to develop a highly reliable language model capable of categorizing postings in online discussions. We will be using a provided corpus, including an online discussion about the story ”The Lady, or the Tiger?” Our aim is also to ensure that the model is capable of performing this task on other online discussions.

To achieve this, we came up with the following plan. Firstly, we will use traditional analysis techniques such as TF-IDF and bag-of-words models. These methods will serve as our baseline models, allowing us to establish a foundation for comparison. Next, we'll move on to using more advanced models like BERT for sequence classification. Additionally, we'll fine-tune BERT for our task, tweaking its settings to improve how it categorizes online discussions. This way, we aim to make our language model better for analyzing conversations.

Literature review

First we skimmed through James Paul Gee's book ”An Introduction to Discourse Analysis: Theory and Method” [1]. The book provides a fairly detailed description of the field of discourse analysis. We find chapters such as the seven building tasks of language, the distinction between situated meanings and discourse models, discourse analysis, and a detailed example of discourse analysis. However, we did not come across any practices or methods related to NLP. We will use it when encountering any unfamiliar definitions in the future.

In ”ChatGPT in education: a discourse analysis of worries and concerns on social media” publication by Lingyao Li, Zihui Ma, Lizhou Fan, Sanggyu Lee, Huizi Yu, and Libby Hemphill [2] we found promising methodology that will be the base for our project. With the use of RoBERTa transformer model they performed sentiment classification on twitter discourse in order to get tweets with negative sentiment. On these tweets BERTopic model was used to get sentence embeddings in latent space. This latent space consists of high-dimensional vectors so to battle the ”curse od dimensionality” UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction method was used. After the k-means clustering algorithm was used on embeddings to extract topic information from similar embeddings. On tweets in the same cluster the Term Frequency-Inverse Document Frequency method (c-TF-IDF) was used to extract the topical keywords and representative tweets from each cluster for different topic categorization.

There was also ”TopicGPT: A Prompt-based Topic Modeling Framework” article by Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer [3] that heavely relied on prompts to the TopicGPT model for topic categorization. Due to the fact that the GPT API service is not wallet-friendly we won't be using openAI embeddings. Nevertheless the article gave us another promising idea. Instead of manual creation of topic category names we will use chatGPT prompts with keywords for each category in order to automize the process.

We also analyzed the work that was done in a paper from the university of Birmingham. ”Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis” [4], the paper focuses on using LLM GPT models to annotate and analyse discourse. In this paper the proposed methedology is using prompts to analyse text. The results of the paper shows that in the specific task the GPT-4 model outperforms the GPT-3.5 model. While our goal in the project is not completly connected to prompt engineering it is valuble information to know how a generic GPT model performs in such a task. Also as we mentioned in the previous paragraph we will be using a GPT model for topic category creation.

Methods

Data preprocessing

For this task, a large number of labeled discourse examples are essential. Initially, the original Excel tab named ”Combined Discussions” contained 612 examples. To extend our dataset, we added all unique examples from other Excel tabs, resulting in a total of 755 unique labeled discourse examples. Despite this effort, we still belive that this number is considerably low for such model training.

We extracted the message and the R2 discussion type column for each discourse example. Additionally, we conducted data cleaning by removing all rows containing any NaN (missing) values. The preprocessed data was then saved to a file named ”final_data.csv”. These preprocessing steps were performed within the ”data_preprocessing.ipynb” notebook.

Implementation with analysis

In this section we will look at text categorization task that was solved with tf-idf and bag-of-words representation which we used to train different machine learning models. After preprocessing all available data we used messages from different forums and their human annotated labels. Every message was converted to vector depending on chosen representation and forwarded to machine learning models with associated label. The models we used are XGBoost, Logistic Regression, Random Forest, and SVM with different kernels (linear, RBF and polynomial). In table x we can see the results for every technique.

Table 1. Results of different ML models using tf-idf representation.

XGB	LR	RF	SVM-lin	SVM-rbf	SVM-poly
0.5927	0.6095	0.5492	0.5351	0.5674	0.6025

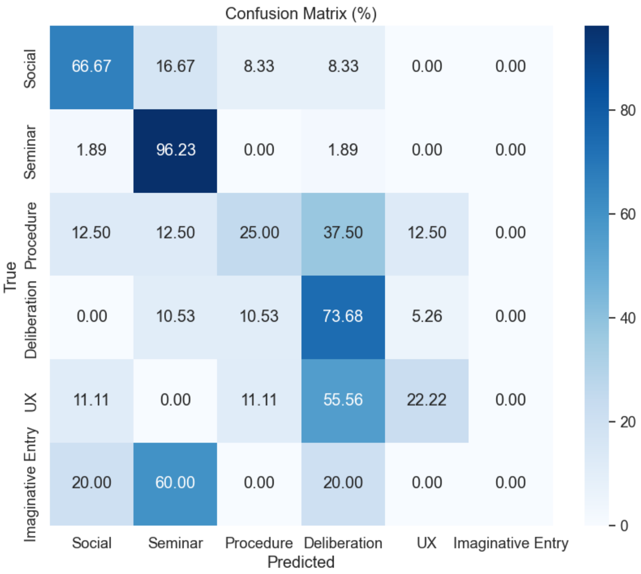
Table 2. Results of different ML models using Bag-of-words representation.

XGB	LR	RF	SVM-lin	SVM-rbf	SVM-poly
0.6062	0.5878	0.5666	0.5099	0.6203	0.6246

For evaluation we used 5-fold cross validation where we computed f1 score with micro averaging. From results we can observe that there is not one universal ML model that will perform best on different representation so it is important to fine tune the classificator in our LLM implementation.

Implementation with BERT

We started this implementation the same way we started in the previous section by importing the preprocessed data. The first task was preparing the data to be accepted by the BERT model. First we made mappings for converstion from labels to ids, which are used by the model. Afterwards we tokenized the messages to the vocabulary that the model uses. This was accomplished using datasets and a Bert Tokenizer. We also split the data into the train (70%), validation (15%) and test (15%) set. For the first test we used a model for sequence classification using the bert model (more specifically the bert base cased). As in the previous tasks we decided to look at the f1 micro score, but we also added a confusion matrix for the results to better understand the model. For the training part we decided to run the learning proces for 10 epochs. After training the neural model we got a classification accuracy on the validation set of 0.733 and on the test set we got a classification accuracy of 0.726. Below we can see the confusion matrix for the test set.



The next step was trying different bert models and adapters to improve the results... tbd

Implementation with custom neural model using BERT embeddings

TO DO

Results

Conclusion

Acknowledgments

References

[1] James Paul Gee. *An introduction to discourse analysis: Theory and method*. routledge, 2014.

[2] L Li, Z Ma, L Fan, S Lee, H Yu, and L Hemphill. Chatgpt in education: A discourse analysis of worries and concerns on social media. arxiv 2023. *arXiv preprint arXiv:2305.02201*.

[3] Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*, 2023.

[4] Li L Su H Fuoli M Yu, D. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*, 2023.