FRI Natural language processing course 2024
Project report

University *of Ljubljana*
Faculty *of Computer and Information Science*

Qualitative Research on Discussions - text categorization

Jakob Mrak, Enei Sluga, and Lan Vukušič

Abstract

This project aims to develop a highly reliable language model for qualitative discourse analysis, a critical method in social science research for understanding human interaction. The task involves categorizing postings in online discussions, such as those about the story ”The Lady, or the Tiger?” using a coded dataset with high inter-rater reliability and a comprehensive codebook. The project addresses the challenge of achieving high inter-rater reliability, which is crucial for ensuring consistency and accuracy in qualitative research. This is particularly important in qualitative discourse analysis, where the nuanced understanding of human interaction requires a deep comprehension of the discussion context, each participant's perspective, and the intertextual relationships between sentences. The project leverages large language models (LLMs) to automate this labour-intensive task, aiming to generalize the model to other online discussions. This approach not only enhances the efficiency of qualitative discourse analysis but also contributes to the broader field of social science research by providing a scalable solution for analyzing complex human interactions in digital spaces.

Keywords
LLM, discourse

Advisors: Slavko Žitnik

Introduction

In this report, we will delve into the significance and objectives of our project, which aims to leverage different language processing techniques for qualitative discourse analysis, a critical method in social science research. This project addresses the challenge of achieving high inter-rater reliability in the categorization of postings in online discussions, such as those about the story ”The Lady, or the Tiger?” using a coded dataset with high inter-rater reliability and a comprehensive codebook [1]. The relevance of our work lies in its potential to enhance the efficiency of qualitative discourse analysis, a traditionally labour-intensive process requiring significant time and resources. By addressing the limitations and variability of existing LLMs, we aim to contribute to the broader field of social science research by providing a robust solution for qualitative discourse analysis. This is crucial for advancing our understanding of human interaction in digital spaces and enhancing qualitative research's efficiency and accuracy.

0.1 ChatGPT in education

The work ChatGPT in education: A discourse analysis of worries and concerns on social media [1] employs a comprehensive research framework for analyzing discourse surrounding ChatGPT in the realm of education on Twitter. Their study begins with data collection, focusing on tweets related to ChatGPT and education. They then utilize sentiment analysis, leveraging a sentiment model based on the RoBERTa architecture, to classify the sentiment of collected tweets. This sentiment analysis enables them to discern the attitudes and opinions expressed towards ChatGPT within educational contexts. Furthermore, they employ BERT-based topic modelling techniques to uncover semantic themes within the collected tweets. This approach involves utilizing BERT embedding to extract semantically relevant sentence embeddings, followed by dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) to construct coherent topic clusters. Finally, they utilize K-Means clustering and class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) to represent and interpret these topics, facilitating a deeper understanding of the discourse landscape surrounding ChatGPT in educational settings.

While this work focused on sentiment analysis and topic modelling within the context of ChatGPT in education, our

[1] Codebook is a structured document or dictionary that contains mappings between symbols or codes and their corresponding linguistic features or meanings

project aims to develop a highly reliable language model for qualitative discourse analysis across diverse online discussion topics. By leveraging large language models (LLMs) and a coded dataset with high inter-rater reliability, we seek to automate the process of categorizing postings in online discussions, thereby contributing to a scalable solution for analyzing complex human interactions in digital spaces. Thus, while *ChatGPT in education* [1] provides valuable insights into the discourse surrounding ChatGPT in education, our work extends beyond sentiment analysis and topic modelling to address the broader challenge of qualitative discourse analysis in social science research.

0.2 TopicGPT

In the article [2] they introduce TopicGPT, a novel framework that leverages large language models (LLMs) to enhance the process of topic modelling in text corpora. Traditional topic models, such as Latent Dirichlet Allocation (LDA), often represent topics as bags of words, which can be difficult to interpret and offer limited semantic control. TopicGPT addresses these limitations by using LLMs to uncover latent topics within a text collection, producing topics that align better with human categorizations. Its strengths lie in generating topics based on seen textual examples. This allows this approach to refine the set of topics to a more coarse or fine set, depending on the analyzed conversation. Additionally, TopicGPT is adaptable, allowing users to specify constraints and modify topics without retraining the model.

This work is relevant to our project as it demonstrates the potential of LLMs in improving the efficiency and interpretability of text analysis tasks, which is a key goal of our project to develop a highly reliable language model for qualitative discourse analysis. The adaptability and semantic understanding of TopicGPT could provide valuable insights for developing a language model that not only categorizes text effectively but also offers a level of semantic understanding and interpretability that aligns with the goals of qualitative discourse analysis.

Data

The dataset utilized in our study originates from an online discussion centered around the narrative of ”The Lady, or the Tiger?” Each message within this discussion has been systematically categorized into distinct discussion types, enabling us to leverage supervised learning techniques for classification purposes. Given the relatively small size of the dataset, comprising approximately 400 messages, we will adopt few-shot learning strategies to effectively train our model. Additionally, the dataset includes supplementary annotations such as dialogue spells and pivot points, which could offer valuable insights if predicted accurately. However, these annotations present challenges due to their inconsistent, sparse nature, necessitating additional preprocessing efforts to render them usable.

Methods

We have focused on topic prediction as our main task. The approaches we've tried are described in the following sections.

0.3 Sentence embeddings

Our following approaches heavily rely on the usage of embedding vectors for further processing. We theorize, that given the size of the dataset, which is between 500 and 1000 samples, an approach using general pre-trained sentence embedings is the only viable approach. Finetuning or training a whole LLM [2] would be impractical and prone to severe overfitting.

We chose a variety of pretrained sentence embedings, from the Huggingace's Sentence-transformers class of models. [3]. This is due to their ease of use, which allowed us to perform more tests.The models can also be trained an run locally, even on machines with no GPU, with reasonable inference speed [4] Model used for final outputs is the **snowflake-arctic-embed-m** [3].

0.4 Clustering

Our goal at first was to try to use sentence transformers to embed the messages from our dataset and then use clustering to group them. This way we could have a better understanding of the data and see if there are any patterns in the data and if they correlate to the provided ground truth labels. We chose to use the sentence transformers library because it is a very powerful tool for embedding sentences and it is very easy to use. We selected four different sentence transformers models to embed the data and then we used the KMeans and Agglomerative clustering to cluster the data. We used TSNE to visualize the clusters and see if there are any patterns in the data.

0.4.1 K-Means Clustering

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n-observations into k-clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. At first our aim was to define the number of clusters as the number of unique labels in the dataset, but we later realised that this is not the best approach because some classes are not well represented in the dataset and the model could not identify the patterns in the data well. We found that the best number of clusters is 6, which is the number of the most represented classes in the dataset. On 1 we can see the clusters that the KMeans model identified for embeddings from the transformer all-mpnet-base-v2, which was the best performing model for the clustering task. It is visible that the model identified some patterns in the data, but the clusters are not

[2] Large language model

[3] Sentance transformers is a library that allows easy use of models pretrained for sentence similarity and semantic search, available here https://huggingface.co/sentence-transformers

[4] model is trained in less than minute and inference is below one second for a batch of 100 sentances.

well seperated, which was expected given the nature of the provided data.

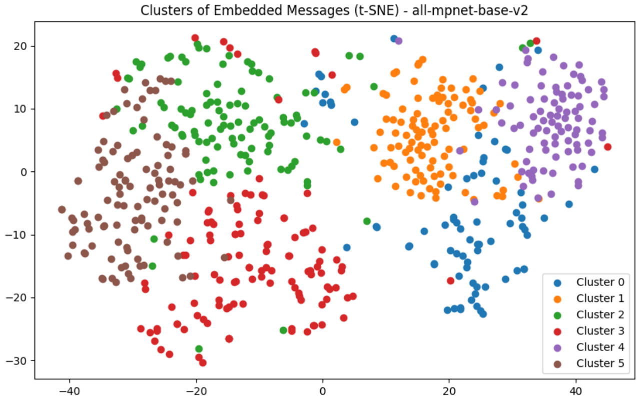

Figure 1. The graph shows the clusters that the KMeans model identified for transformer all-mpnet-base-v2.

In table 1 we look at each cluster in more detail to see which classes are represented in each cluster. At first glance we can see that in most clusters, discourse type ”Seminar” is in majority, which was expected since 54.7% of all data is of class ”Seminar”. In case of cluster 0, we can see that we have a majority of class ”Social” and in cluster 5 we have a majority of class ”Deliberation”. This shows that the model was able to identify some patterns in the data that correlate to the provided labels, but mostly the clusters do not represent the given labels well.

Clusters	Seminar	Social	Procedure	UX	Deliberation	Imaginative e.	Rest of
Cluster 0	2.94%	**47.06%**	27.45%	12.75%	8.82%	0%	0.98%
Cluster 1	**76.81%**	0%	0%	0%	5.80%	17.39%	0%
Cluster 2	**94.57%**	0%	0%	0%	2.17%	1.09%	2.17%
Cluster 3	**45.07%**	10.56%	5.63%	11.97%	19.01%	2.11%	5.65%
Cluster 4	**96.23%**	0%	0%	0%	0.94%	1.89%	0.94%
Cluster 5	24.00%	6.00%	10.00%	17.00%	**42.00%**	0%	1%

Table 1. Cluster composition percentages using K-means with embeddings from transformer all-mpnet-base-v2.

0.4.2 Agglomerative Clustering

Agglomerative clustering is a type of hierarchical clustering that builds a hierarchy of clusters. The model starts with each point as a separate cluster and then merges the closest pairs of clusters until only one cluster remains. We tested different linkage methods, the most promising one was ’complete’ which uses the maximum distances between all observations of the two sets. We used the same number of clusters as in the KMeans model, which is 6. We can observe the results in table 2, where we can see the clusters that the Agglomerative model identified for embeddings from the transformer all-mpnet-base-v2. In cluster 0, we can see that the model correctly clustered 90.04% of the data as class ”Seminar”, which is expected, but it still shows that the model was able to identify some patterns in the data that correlate to the provided labels. The difference between the KMeans and Agglomerative clustering is that the Agglomerative model was able to cluster the data somewhat better, but the clusters are still not well separated. In the case of agglomerative clustering, the clusters exhibit distinct compositions. Notably, there are three clusters (Cluster 1, Cluster 3, and Cluster 4) where the majority of the data does not belong to the ”Seminar” class, which is the most represented class in the dataset. This indicates that these clusters contain data points that are characterized by different attributes compared to the majority of the dataset. This shows some potential for the model to identify patterns in the data that correlate to the provided labels, but ultimately the clusters do not represent the given labels well enough to be able to predict the discourse type of the given data.

Clusters	Seminar	Social	Procedure	UX	Deliberation	Imaginative e.	Rest of
Cluster 0	**90.04%**	0%	0%	0.99%	3.98%	3.98%	1.01%
Cluster 1	22.09%	4.65%	15.11%	12.79%	**43.02%**	1.16%	0.18%
Cluster 2	**29.54%**	22.73%	0%	15.90%	17.04%	0%	14.79%
Cluster 3	0%	0%	5.56%	**61.11%**	33.33%	0%	0%
Cluster 4	4.7%	**47.05%**	25.88%	7.05%	14.11%	0%	1.21%
Cluster 5	**77.44%**	3.76%	1.5%	2.25%	5.26%	7.51%	2.28%

Table 2. Cluster composition percentages using agglomerative clustering with embeddings from transformer all-mpnet-base-v2.

0.5 Classification model

In this approach, all input sentences have been individually embedded using the previously mentioned sentence embedding techniques 0.3. This is the input for the model, that needs to predict in which of the 8 classes does the text belong to.

Compound classes (entries labeled, belonging to more classes) are removed, due to their low occurrence (1 or 2). Here is the table with classes and their respected counts.

As seen classes are unbalanced. To mitigate this, cross entropy loss is weighted by the count of samples in the predicted class, so that minority classes don’t get ignored.

A shallow 2-layer 32 hidden neuron perceptron is used, to predict the classes. Larger networks were too prone to overfitting.

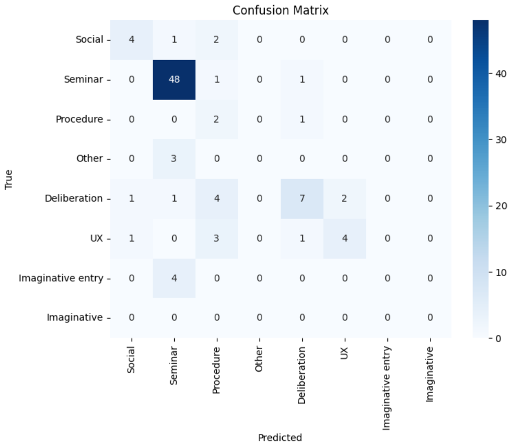

Figure 2. Model confusion matrix on validation data shows, that model correctly predicts the majority class but struggles with less used ones.

Testing the model in interactive mode gave some convincing results. We tried some made-up sentences a concluded that the model predicts **Seminar** and **Social** classes with high certainty but confuses **Deliberation** and **Procedure**.
We believe that part of the problem is that some text samples are difficult and sometimes deceiving for a human, and we (three team members) could not unanimously decide on the correct label.

Overall, we consider this approach successful. In the future, better results can be achieved by providing additional context to the model if parts of the conversation are tempoarly correlated. We believe, that Domain-specific or context-aware embeddings would also greatly improve model performance.

0.6 TF-IDF embeddings

Next we decided to try a classical approach by using TF-IDF embeddings. We used the TfidfVectorizer from the scikit-learn library to transform the text data into numerical data. First we preprocessed the text data by removing stopwords, punctuation, converting the text to lowercase and lemmatizing the text. Then we constructed the TF-IDF matrix with the TfidfVectorizer. We then decided to try three different models to classify the data: logistic regression, support vector machine and random forest classification. We used the scikit-learn library to implement the models, TF-IDF matrix as input data, the provided labels as target data and 5-fold cross-validation to evaluate the models. The results are shown in the table 3.

Model	Accuracy	Standard deviation
Logistic regression	0.57	0.01
Support vector machine	0.56	0.01
Random forest	**0.61**	**0.02**

Table 3. Results of the models using TF-IDF embeddings.

As we can see, the random forest model performed the best with an accuracy of 61%. Given the fact that our dataset has 6 classes that have less then 1% of the data since they are a combination of the majority classes, we decided to either merge them into one class or remove them from the dataset. After removing the minority classes, we tried the same models again and the results are shown in the table 4.

Model	Accuracy	Standard deviation
Logistic regression	0.59	0.01
Support vector machine	0.57	0.01
Random forest	**0.64**	**0.05**

Table 4. Results of the models using TF-IDF embeddings without minority classes.

As we can see, the performance of the models improved after removing the minority classes. The random forest model performed the best with an accuracy of 0.64. Given the fact that the dataset is small and the classes are unbalanced, the results are satisfactory. The majority class 'Seminar' has 54.5% of the data (55.6% after removal of minority classes), so the models are biased towards this class. This means that the Random forest model has about 10% better accuracy than a model that would predict only the majority class. Given the nature of the dataset, the results are satisfactory.

0.7 Prompt based approach

We also experimented with prompting an LLM to identify the type of discourse in a message. For this approach, we used the LLaMA 3 model as described in the original paper [4]. We provided the LLM with context on how the classes are defined and strict guidelines on how it should respond. Despite some incorrect outputs where it failed to recognize the message as valid, it correctly classified the discourse type most of the time, achieving around 65% accuracy on our dataset. This performance is comparable to other methods. However, this approach is not the most practical, as it requires more computational resources than other methods while delivering similar results. Nonetheless, it works without requiring training on the data, making it a viable option when data is not available.

0.8 Experimenting with LLaMA embeddings

0.8.1 Dimensionality reduction

We tried using the embeddings we got from the prompt based approach, to see if they would be better suited for classification. We tried using UMAP [5] for dimensionality reduction. The results reveal 4 distinct groups, but these groups are not at all correlated with the given groups. The results can be seen on figure 3

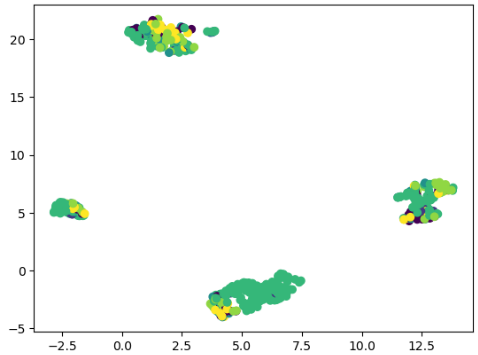

Figure 3. LLaMA embeddings after dimensionality reduction using the UMAP approach. The colors represent our classes.

We then tried using t-SNE [6] to reduce the dimensionality. When using t-SNE we passed in our labels so that it could try and cluster them together. On figure 4 we see that this approach also results in four groups that do not correlate with our labels.

0.8.2 Classification with LLaMA embeddings

Even though the dimensionality reduction methods didn't seem promising we tried classification using these vectors. Using 5 fold cross validation we tested numerous classification models. The results can be seen on table 5.

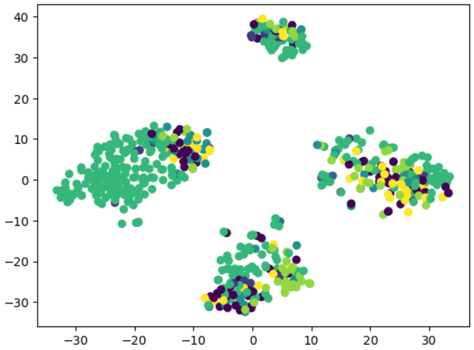

Figure 4. LLaMA embeddings reduced to 2D using t-SNE.

Model	Accuracy	Std. dev.
25 Nearest neighbours	0.62	0.01
Ridge classifier	0.67	0.04
AdaBoost	0.57	0.04
Neural network (500, 100)	0.68	0.05

Table 5. Results of the models using TF-IDF embeddings without minority classes.

The neural network classified the embeddings the best.

0.8.3 Classification with large language models

After finding success with a neural network trained on a general text corpus [7], we decided to fine-tune a **RoBERTa** model on our dataset.
We hypothesized that the dataset consisting 600 examples (400 after train test split) will not suffice for efficient LLM fine-tuning and would severely overfit the model.

Results have shown that after training for only 1 epoch on a whole batch, the model started overfitting. We could not get better results with prolonged training.

Epoch	Training Loss	Validation Loss	Accuracy
1	2.2779	2.583391	0.065041
2	1.9629	3.857165	0.065041
3	0.8381	4.522352	0.097561

After multiple training reruns, the model achieved an average of **65% accuracy**, which is worse than the naive TF-IDF approach 3.

Discussion

We have concluded that several methods that we tried to perform reasonably well on predicting the classes which have a meaningful presence in the dataset.

We believe having a larger balanced dataset would result in significantly better classification accuracy, especially for minority classes, which were often grouped with the prevalent ”seminar” class 2.

Given the dataset limitations, we conclude that statistical methods (TF-IDF) or pre-trained models perform better than custom ones, due to a lack of diverse class examples.

References

[1] Lingyao Li, Zihui Ma, Lizhou Fan, Sanggyu Lee, Huizi Yu, and Libby Hemphill. Chatgpt in education: A discourse analysis of worries and concerns on social media, 2023.

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

[3] Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models, 2024.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[5] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[6] T. Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data, 2022.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.