FRI Natural language processing course 2021
Project report

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

Natural language processing course Latex Template

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 large language models (LLMs) 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]

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

[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.

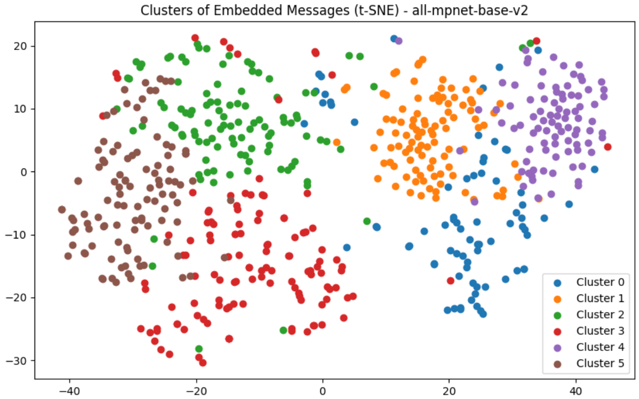

Figure 1. The graph shows the clusters that the KMeans model identified for transformer all-mpnet-base-v2.

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.

0.4.3 DBSCAN Clustering

TODO: če bo treba lahko še tole opišem sam rezultati so še bol šit

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.

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.

Results

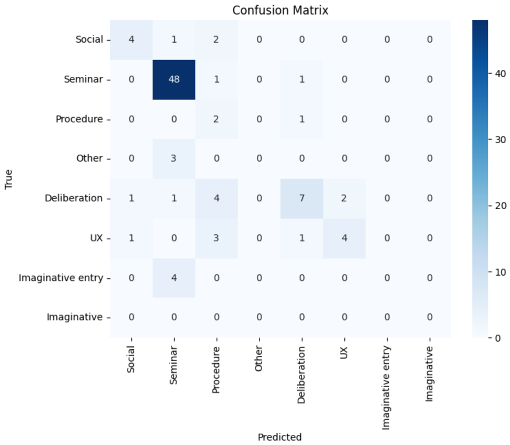

Figure 2. Model confusion matrix on validation data shows, that model correctly predicts the majority class but struggles with less used ones.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

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.