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

Qualitative Research on Discussions - text categorization

Krištof Zupan, Gal Menaše, and Nejc Mušič

Abstract

The abstract goes here. The abstract goes here.

Keywords

Keyword1, Keyword2, Keyword3 ...

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.

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 heavily 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 methodology 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 completely 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

Use the Methods section to describe what you did an how you did it – in what way did you prepare the data, what algorithms did you use, how did you test various solutions ... Provide all the required details for a reproduction of your work.

Below are LaTeX examples of some common elements that you will probably need when writing your report (e.g. figures, equations, lists, code examples ...).

Equations

You can write equations inline, e.g. $\cos \pi = -1$, $E = m \cdot c^2$ and α, or you can include them as separate objects. The Bayes's rule is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are some events. You can also reference it – the equation 1 describes the Bayes's rule.

Lists

We can insert numbered and bullet lists:

1. First item in the list.
2. Second item in the list.
3. Third item in the list.

- First item in the list.
- Second item in the list.
- Third item in the list.

We can use the description environment to define or describe key terms and phrases.

Word What is a word?.

Concept What is a concept?

Idea What is an idea?

Random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis

felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Figures

You can insert figures that span over the whole page, or over just a single column. The first one, Figure 1, is an example of a figure that spans only across one of the two columns in the report.
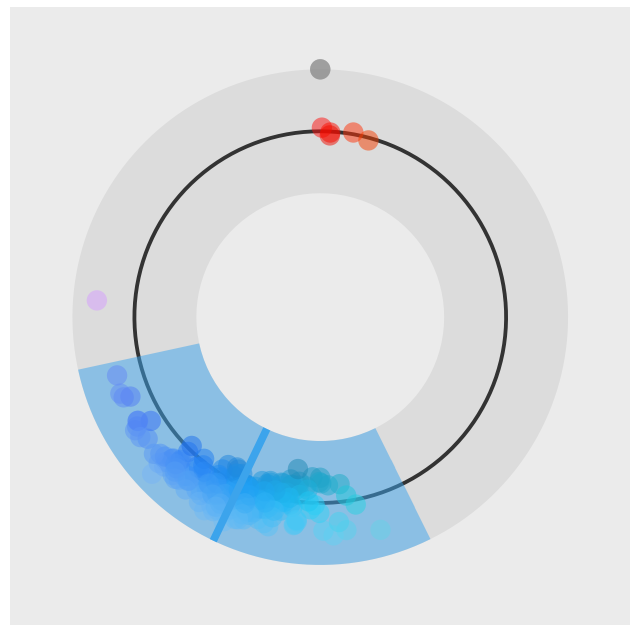
Figure 1. A random visualization. This is an example of a figure that spans only across one of the two columns.
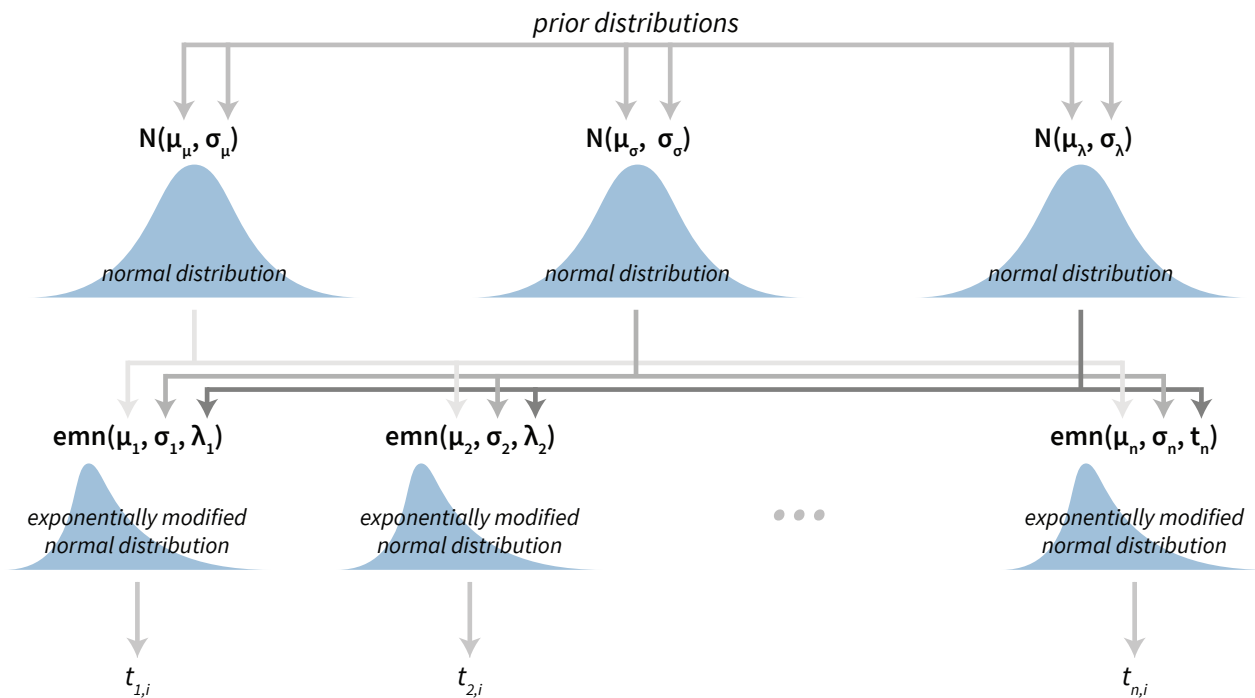
On the other hand, Figure 2 is an example of a figure that spans across the whole page (across both columns) of the report.

Tables

Use the table environment to insert tables.

Code examples

You can also insert short code examples. You can specify them manually, or insert a whole file with code. Please avoid inserting long code snippets, advisors will have access to your repositories and can take a look at your code there. If

prior distributions

$N(\mu_\mu, \sigma_\mu)$

normal distribution

$N(\mu_\sigma, \sigma_\sigma)$

normal distribution

$N(\mu_\lambda, \sigma_\lambda)$

normal distribution

$emn(\mu_1, \sigma_1, \lambda_1)$

exponentially modified normal distribution

$emn(\mu_2, \sigma_2, \lambda_2)$

exponentially modified normal distribution

$\bullet \bullet \bullet$

$emn(\mu_n, \sigma_n, t_n)$

exponentially modified normal distribution

$t_{1,i}$

$t_{2,i}$

$t_{n,i}$

Figure 2. Visualization of a Bayesian hierarchical model. This is an example of a figure that spans the whole width of the report.

Table 1. Table of grades.

Name		
First name	Last Name	Grade
John	Doe	7.5
Jane	Doe	10
Mike	Smith	8

necessary, you can use this technique to insert code (or pseudo code) of short algorithms that are crucial for the understanding of the manuscript.

Listing 1. Insert code directly from a file.

```python
import os
import time
import random

fruits = ["apple", "banana", "cherry"]
for x in fruits:
  print(x)
```

Listing 2. Write the code you want to insert.

```r
import(dplyr)
import(ggplot)

ggplot(diamonds,
        aes(x=carat, y=price, color=cut)) +
  geom_point() +
  geom_smooth()
```

Results

Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem,

vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

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.

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.