Qualitative Research on Discussions - text categorization

Cristian Bassotto, Nikolay Kormushev, and Ahmet Çalış

Abstract

Qualitative discourse analysis is an important way social scientists research human interaction. Large language models (LLMs) offer potential for tasks like qualitative discourse analysis, which demand a high level of inter-rater reliability among human "coders" (i.e., qualitative research categorizers). This is an exceedingly labor-intensive task, requiring human coders to fully understand the discussion context, consider each participant's perspective, and comprehend the sentence's associations with the previous discussion, as well as shared general knowledge. In this task, you create a model to categorize postings in online discussions, such as in a corpus — an online discussion about the story, "The Lady, or the Tiger?". We provide a coded dataset with a high inter-rater reliability and a codebook including definitions of each category with examples. Your task is building and training a highly reliable language model for this coding task that generalizes to other online discussions.

Keywords
NLP, LLM, Discourse analysis

Advisors: prof. Slavko

Introduction

Discourse can be defined as particular way of talking about and understanding the world (or an aspect of the world) [1]. On the other hand, qualititive discourse analysis is an attempt to understand how meaning is formed. While it traditionally has been conducted manually by linguists and researchers, the increasing volume and complexity of textual data call for automated solutions to expedite the process and enable analysis at scale as mentioned here [2].

With the rise of LLMs many solutions appeared that fine-tune these models for specific use cases. This requires a programmer to sit down and understand the domain for which he needs to fine-tune the model. Even if we just take into account only the read time this process can be quite time consuming [3] and with this paper we aim to develop a solution that can generalize and generate discourse analysis on multiple domains.

To achieve our goal of classification we will try three approaches: topic modelling for classification, transformers-based models and prompt engeneering using Mistralis LLM. By comparing these approaches, we seek to identify the method that generalizes the best. For us it is also important our analysis to be explainable and so we use an LLM for explaining our results.

Related Works

In the field of text categorization, researchers have explored various methods to better understand and classify written data. One key study [4] looked at different ways to classify texts using supervised machine learning techniques like Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, and AdaBoost. These approaches represent the state of the art in text categorization and have historically served as foundational pillars in the field.

However, it's important to note that while these traditional methods remain relevant and are still utilized in various contexts, emerging techniques have demonstrated superior performance in certain scenarios. Another important review [5] surveyed the techniques used for text classification across different areas and over the years. They classified documents based on their content and language, using the models that we already discussed, but reaching also the most recent transformers-based models.

Recent research has focused on advanced models like BERT (Bidirectional Encoder Representations from Transformers). These models have shown promise in improving text classification accuracy. For example, one study compared BERT to another model called XLNet to see which one worked better for text classification tasks [6].

In the past few years the focuse changed to fine-tuning

these models by either adding extra layers or adjusting hyperparameters and the corpus used. Despite these advancements, there have been instances where research didn't result in better outcomes. For example, in some cases, adding more layers actually led to poorer performance in specific studies [7]. It's important to note that there isn't a one-size-fits-all solution in text categorization. What works well for one dataset may not work as effectively for another. This study [8] focused on this and tried to better find specific BERT-based models for different specifics dataset, finding out that the best model is not always the most complex one.

Research in text categorization is constantly evolving, with new methods and technologies being developed to better understand and classify written texts. New approches focused on Large Language Models (LLMs) such as GPT-3, but according to the following research [9] their performances still significantly underperform fine-tuned models in this task. In this paper we also want to compare this different approches.

Initial Ideas

Preprocessing

General

In our investigation, we intend to explore diverse preprocessing techniques and embeddings to evaluate their impact on classification performance. Alongside classic methods such as word2vec embeddings, we will examine alternatives including BERT embeddings and fastText. Additionally, we will consider various preprocessing strategies such as tokenization, stemming, and lemmatization, aiming to optimize the quality and representational power of the input data for each embedding method.

Topic Modelling

Topic modelling deals with generating topics from a corpus of documents and then assigning them to a document. As a topic in this context we mean a set of key words that were extracted from the document. These topics can be generated using models like BERTopic [10] or LLMs [11]. In [12] we see that supplying the topic to the model used for classification increases the resulting accuracy so we wanted to try to see how big of an improvement we could get for both approaches. First we will generate the topics for our input text and then we supply it as part of the input to our models. In this regards different approches tried to study online discussions throw topic modeling and manual categorization [13] [14], but we want to implement an automated approach.

Models

BERT

In this study, we adopt a simple approach to automate discourse analysis by fine-tuning the BERT (Bidirectional Encoder Representations from Transformers) model on our dataset. By establishing this baseline, we aim to assess BERT's effectiveness in capturing discourse patterns and set a benchmark for evaluating alternative methodologies.

Prompt Engineering

Inspired by [15] we decided also to try to see how simple prompt engineering compares to classification with BERT. In this case we plan to use Mistralis and send it prompts asking it to classify our sentences.

In the topic modeling paper they achieved a much better performance comapred to solutions like BERTopic which shows it can be a valid approach. The benefits of this approach is that we do not require to fine-tune our model. We just run it and ask it questions. Althought this is a trade off because we need to find appropriate prompts to classify with a good accuracy. Here we need to be carefull since Mistralis is also trained on a bigger data corpus which means it is likely it will perform much better when generalising.

Evaluation

To evaluate how our model generalises we will scrape a small sample from forum sites like Reddit or Quora and evaluate the results ourselves and decided. This is not optimal but we do not have an already labeled dataset we can use for the task so the alternative is to create one which requires an expert and time which we do not have.

Explainability

With the development of AI explainability of models has become a huge topic. It is important that our non-expert lingiust users can understand why a specific category was assigned to that specific input. To make it as easy as possible to use we plan to generate textual explanations using an LLM like Mistralis or ChatGPT by feeding it our inputs and outputs and asking to explain our models conclusions. Using an LLM for explainability can enrich the experience of our end users by making the results much more understandale but it can als mean it can also confuse them because it is suceptible to hallucinations [16]. Because of this we will ask the LLM to return a section of our input based on which it made its conclusions.

We are also planning to try to use topic modelling to aid explainability. By genereting the keywords from the text we classify we believe it can become clearer to the user why a classfication decision was made.

References

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

[2] Thomas Jacobs and Robin Tschötschel. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 22(5):469–485, 2019.

[3] Rolf A Zwaan, Gabriel A Radvansky, Amy E Hilliard, and Jacqueline M Curiel. Constructing multidimensional situation models during reading. *Scientific studies of reading*, 2(3):199–220, 1998.

[4] Ishaan Dawar, Narendra Kumar, Sakshi Negi, Sayeedakhanum Pathan, and Shirshendu Layek. Text categorization using supervised machine learning techniques. In *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, pages 185–190. IEEE, 2023.

[5] Enock Osoro Omayio, Indu Sreedevi, and Jeebananda Panda. Language-based text categorization: A survey. *Digital Techniques for Heritage Presentation and Preservation*, pages 11–36, 2021.

[6] Parsa Sai Tejaswi, Saranam Venkata Amruth, Prakya Tummala, and M Suneetha. Automatic documents categorization using nlp. In *ICT Infrastructure and Computing: Proceedings of ICT4SD 2022*, pages 215–225. Springer, 2022.

[7] Samin Mohammadi and Mathieu Chapon. Investigating the performance of fine-tuned text classification models based-on bert. In *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1252–1257. IEEE, 2020.

[8] Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, and Abdulwahab Ali Almazroi. A fine-tuned bert-based transfer learning approach for text classification. *Journal of healthcare engineering*, 2022, 2022.

[9] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

[10] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[11] Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. Re-visiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*, 2023.

[12] A Glazkova. Using topic modeling to improve the quality of age-based text classification. In *CEUR Workshop Proceedings*, pages 92–97, 2021.

[13] 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. *Education and Information Technologies*, pages 1–34, 2023.

[14] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*, 2023.

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

[16] Patrick Huber and Giuseppe Carenini. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *arXiv preprint arXiv:2204.04289*, 2022.