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

Qualitative Research on Discussions - text categorization

Sanil Safić, Miha Šircelj, Jan Topolovec

Abstract

This paper presents a method to analyze discussions on ”The Lady, or the Tiger?” using Large Language Models (LLMs) for text categorization. Our approach involves exploring a coded dataset, refining models on a HPC infrastructure, and iteratively improving performance compared to human coders. Additionally, we employ separately fine-tuned LLMs to explain identified categories, enhancing interpretability. By implementing this method, we aim to develop reliable models for accurately categorizing discussions, contributing to a better understanding of discourse dynamics.

Keywords

Text categorization, LLMs, Natural language processing

Advisors: Slavko Žitnik

Introduction

Our goal for the seminar project is to leverage large language models (LLMs) for qualitative discourse analysis, specifically focusing on text categorization. We will achieve this through a methodical approach involving literature review, data exploration, model fine-tuning, and performance evaluation. We aim to develop highly reliable models that perform well on a given dataset and generalize to other online discussions. To accomplish this, we will evaluate our model against human-annotated labels after each iteration step.

Related work

Numerous studies have delved into research methods for analyzing discussions and categorizing text. Researchers have explored different approaches like thematic analysis and machine learning algorithms to better understand how discussions evolve and how text can be categorized effectively.

In this section, we will discuss three related articles. The first article concentrates on supervised and traditional term weighting methods for automatic text categorization, while the other two articles focus on the utilization of GPT engines for text analysis and categorization.

The first article [1] investigates various term weighting methods to enhance automatic text categorization performance. It evaluates both supervised and traditional (unsupervised) methods on benchmark datasets like Reuters-21578 and 20 Newsgroups. Introducing a new supervised term weighting method called *tf.rf*, which combines term frequency with relevance frequency, the study aims to improve term discriminating power by considering the distribution of relevant documents in the collection. Results indicate that while traditional methods like *tf.idf* showed mixed performance, the proposed *tf.rf* method consistently outperformed others across different datasets and algorithms, suggesting its effectiveness in enhancing text categorization by better capturing term importance based on document distribution in categories.

The second article [2] introduces TopicGPT, a prompt-based framework that uses LLMs to find topics in a collection of documents. They address the limitation of using bag-of-words topics in other topic modeling methods. TopicGPT operates in two main stages: topic generation and topic assignment. The first one iteratively uses the LLM to generate descriptive topics based on the dataset, refining and merging them for coherence. In the second stage, the LLM assigns topics to documents using the generated topic list, ensuring accuracy through verifiable quotes and self-correction for errors. Results show that TopicGPT aligns substantially better with human-annotated labels than baselines, LDA and BERTopic. The dataset used for this study includes Wikipedia articles and Congressional bills.

The article [3] undertakes an analysis of Twitter posts to identify some concerns about using Generative AI models, particularly ChatGPT, in education. The authors collected data by finding mentions of ChatGPT using Twitter API - they collected a total of 247,484 tweets. For sentiment analysis, they used fine-tuned RoBERTa model which showed better performance than SVM-based models. The RoBERTa

model categorized tweets as positive, neutral and negative, which where subjected to further analysis. They employed BERTopic tool to cluster negative tweets into 200 topics to better expose concerns. The sentiment analysis shows that the majority of users expressed positive or neutral attitude towards ChatGPT in education. Among those expressing reservations about its use, there were some notable topics, such as: Academic integrity, impact on learning outcomes and workforce challenges, specifically expressed by individuals from education and tech.

Data Exploration

In the project, where we try to create a model for text categorization, we will be referencing to the data presented in the form of discussions based on the story "The Lady, or the Tiger?". The author of story does not define its ending and consequently leaves room for many interpretations by readers.

The given data includes discussions that are already appropriately categorized according to category definitions. Each comment is initially categorized according to their type - such as "Seminar" (discussion on story interpretation) or "Social" (discussion between users that is not related to the interpretation). Some of these questions also leave room for further interpretation. In the case of questions, they are generally divided into open and close-ended types, while certain discussions can also be labeled as explanations or agreement of previous arguments.

Proposed implementation

The first step involves thoroughly exploring the provided dataset of coded discussions. This includes examining the dataset's structure, studying the distribution of categories, and understanding the details of the coded discussions.

After exploring the data, we will start building and refining Large Language Models (LLMs) for text categorization. To handle the computational demands of training large-scale models, we'll use a HPC infrastructure, ensuring efficient model development.

Performance evaluation will be a key component of our implementation strategy. Our approach will include iteratively comparing and revising our model's performance against that of human coders (categorizers).

We will use a separately fine-tuned LLM to generate explanations for the identified categories. These explanations will undergo qualitative assessment to ensure clarity and trustworthiness, improving the interpretability of our models' predictions.

First implementation

The first version of our project was implemented using a pre-trained BERT model. Initially, we read the prepared data from the classroom (The Lady Or The Tiger) and preprocessed it - tokenization, converting to lowercase, removing stopwords, etc. Then we selected a pre-trained model suitable for classification. Next, we divided the acquired data into training and testing sets, using an 80:20 training and testing ratio. Later, we created data loaders to iterate over the batches during both training and validation phases. We performed a grid search to optimize hyperparameters. We used ARNES HPC for training our model, which expedited the training process. At the end, we evaluated the performance of the fine-tuned model by testing it on the validation dataset. We measured many metrics such as accuracy, precision, recall, F1-score, confusion, confusion matrix and also made a classification report to assess the effectiveness of the model. These are the best results we achieved - Accuracy: 0.739, Precision: 0.756, Recall: 0.739 and F1 Score: 0.730.

Future directions and ideas

In the future, we plan to improve our initial results by finding better approaches to text categorization. Some of the ideas we are considering include:

- Exploring other pre-trained models like RoBERTa or ALBERT. These models might offer different benefits for the task. For instance, RoBERTa modifies BERT's training process by removing the next sentence prediction objective and training with much larger mini-batches and learning rates.
- Active Learning: Instead of using a static train-test split, we could implement an active learning approach. The model could be initially trained on a small subset of the data, and then iteratively retrained on the most difficult instances.
- Transfer Learning from Related Tasks: Text categorization is related to many other NLP tasks, such as sentiment analysis and topic modeling. By training the model on these related tasks, it could potentially learn features that are useful for text categorization.

References

[1] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):721–735, 2009.

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

[3] 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*, Oct 2023.