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

Qualitative Research on Discussions - text categorization

Ana Petrova, Žan Korošak and Bojan Ilić

Abstract

Text categorization of online discussions involves classifying user-generated content into predefined categories to enhance understanding and analysis. This project focuses on a dataset sourced from an online discussion about ”The Lady, or the Tiger?”. Each entry is labeled according to discussion type, such as Seminar, Deliberation, Social, and others. We employed several different AI models and approaches for text classification and conducted an in-depth analysis of incorrect predictions to understand the challenges that this task presents. First we take a look at traditional approaches to text classification, such as Multinomial Naive Bayes, Random forest, Support Vector Machine and others. BERT was then utilized for its advanced capability in understanding context and classifying text. Lastly, we implemented the concept of prompt engineering using the LLAMA2 model.

Keywords

discourse analysis, text categorization, language model

Advisors: Slavko Žitnik

Introduction

In the evolving field of social science research, qualitative discourse analysis stands out as a key methodology for understanding the complexities inherent in human interactions. This process involves meticulously categorizing text within discussions to capture nuances of context and participant perspectives, traditionally performed by human coders.

This paper presents a novel exploration into automating the categorization of text within online discussions, focusing on the narrative ”The Lady, or the Tiger?”. As digital forums grow, understanding these interactions becomes crucial, yet traditional manual coding is labor-intensive and subject to human error. By leveraging the latest advancements in natural language processing (NLP), particularly large language models (LLMs), we aim to streamline and enhance the reliability of text categorization.

Our research leverages AI-driven models like BERT and LLAMA2, known for their adeptness in deep contextual understanding and flexibility across various discussion contexts. Utilizing a dataset validated with high inter-rater reliability, we aim to refine and apply these models to achieve and potentially exceed the traditional coding tasks performed by human coders. A detailed analysis of model outputs alongside human coding helps pinpoint where AI can not only match but advance the field of discourse analysis. Initiating with a thorough literature review, we anchor our methods in well-established discourse and dialogic analysis frameworks. This review shapes our approach to the analysis and subsequent phases of model development and tuning. Ultimately, the study seeks to shed light on the potential of modern NLP techniques in qualitative text analysis, comparing how our AI models stack up against traditional human coders and existing computational models in terms of performance.

The findings of this study are intended to contribute to the ongoing dialogue about the integration of AI in qualitative research, providing insights into the potential for more efficient and accurate analysis of complex textual data within online platforms. Through rigorous evaluation and comparison, we aim to demonstrate the practical applications and limitations of employing LLMs in the realm of social science research.

Related works

The article by Devlin et al. [1] introduces BERT, a novel language representation model acronym for Bidirectional Encoder Representations from Transformers. Unlike previous models, BERT is uniquely designed to pre-train deep bidirectional representations from unlabeled text, capturing both left and right context in all layers. This allows for fine-tuning with minimal modifications to achieve state-of-the-art performance on various tasks, including question answering and language inference. BERT exhibits simplicity in concept and remarkable empirical effectiveness, evidenced by significant improvements on eleven natural language processing tasks.

In the study ”How to Fine-Tune BERT for Text Classi-

fication" by Sun et al. [2], the researchers examine various methods for optimizing the BERT model to enhance its effectiveness in text classification tasks. They perform detailed experiments across several well-known datasets, establishing new performance benchmarks. Their proposed methodology includes a pre-fine-tuning phase using multitask learning and targeted pre-training with in-domain data. This preparatory step is followed by specific fine-tuning adjustments tailored to the text classification task at hand. The paper offers a comprehensive exploration of factors such as the selection of neural network layers, addressing catastrophic forgetting, and managing long text inputs, all of which significantly influence the performance of the model. This approach highlights the adaptability of BERT in managing a range of text classification problems and provides a robust framework for leveraging deep learning models in natural language processing tasks.

Ye et al. [3] focused on prompt engineering, a critical task for optimizing the performance of large language models on customized tasks. It highlights the complexity involved in analyzing model errors, identifying deficiencies in prompts, and articulating tasks clearly. While existing research suggests that large language models can be meta-prompted for automatic prompt engineering, the authors argue that this approach is limited by a lack of guidance for nuanced reasoning. To address this limitation, the authors propose PE2, a method that enhances the meta-prompt with detailed descriptions, context specification, and a step-by-step reasoning template. PE2 demonstrates remarkable versatility across various language tasks, surpassing competitive baselines on tasks such as MultiArith and GSM8K. Additionally, the method excels in making targeted prompt edits, rectifying erroneous prompts, and generating multi-step plans for complex tasks.

Corpus analysis

The main dataset, sourced from an online discussion on "The Lady, or the Tiger," is accessible on our Github repository [1]. Each entry includes the user's pseudonym and message, categorized with high inter-rater reliability. Entries were labeled primarily by *Discussion type*, aided by provided definitions and examples. Coders also marked messages with classifications like *Dialogic spell*, *Uptake*, *Question*, and *Pivot*. As a primary classification class, we used column *R2 Discussion Type*. The distribution graph (Figure 1) depicts the popularity of each discussion type. Notably, *Seminar* dominates, defined as discussions on content meaning or interpretation, followed by *Deliberation*, focusing on content-related decision-making.

An example of the most popular discussion type, Seminar, would be: *"Do you think the lady is more beautiful than the princess?"* where the main subject is clearly the story itself. An example of Deliberation would be: *"Yes, let's do that"* where users are discussing a decision or action among themselves.

[1] https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-azb

Data preprocessing involved standardizing the classification labels by merging similar categories, such as combining 'Imaginative' and 'Imaginative Entry'. Additionally, we addressed instances where multiple discussion types were assigned to a single message (e.g., 'Seminar, Deliberation, Social') by retaining only the first listed type for consistency.

Due to the definitions of types and subjective human rating, some examples could be classified into multiple categories. For instance, *<username>, those are some good thoughts!* could be seen as a mix of Social, Deliberation, and Seminar.

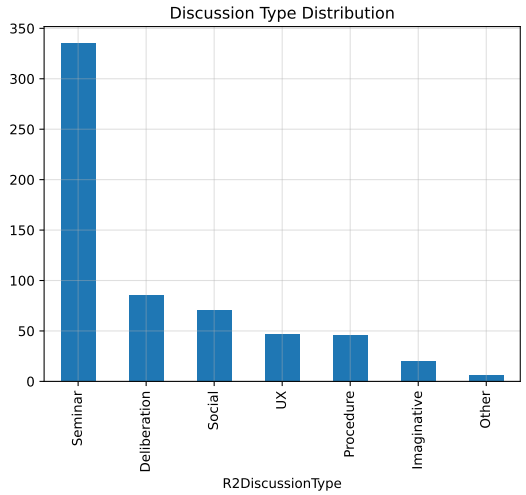

Figure 1. Distribution of different discussion types. We can see that 'Seminar' is the predominant type, which is expected from the definition and the type of discourse. The least common type is "Imaginative" which "places the learner in the discussion as an active participant".

Methods

The main focus of our project is addressing the classification challenge. We chose the *R2DiscussionType* column as our target column as it consistently defines the decision task and serves as the primary class for classification by test subjects. As a benchmark, we employed a pre-trained BERT model (*bert-base-cased*) to classify Discussion types within our dataset. Implementation was carried out using the PyTorch library in Python, which gave us access to Trainer API that we used to fine-tune the pre-trained BERT model. Baseline BERT model was trained with 3 epochs, utilizing a batch size of 8, and employing the standard learning rate of 2×10^{-5}.

We divided our dataset into three parts: train, test and validation. In the training dataset we have 427 samples or 70% of the dataset, and the test and validation datasets both have 92 samples or 15% of the dataset each. Due to the class imbalance in our dataset, certain less represented classes have only a few samples in the divided datasets. The test dataset is used as an evaluation dataset during the training phase, and we use the validation dataset as a final score of the model.

Using the Trainer API, we also fine-tuned the BERT model by increasing number of epochs. To setup optimizers, we used Adam optimizer by PyTorch and tested a scheduler with implemented *ReduceLROnPlateau*, to reduce learning rate when a metric stops improving.

We also tried using DistilBERT[4], which is a lightweight BERT model fine-tuned for text classification.

We also evaluated several traditional AI models: Multinomial Naive Bayes, Random Forest Classifier, Support Vector Machine (SVM) and XGBoost. We used TF-IDF as the embedding technique to transform the input.

In our exploration of advanced NLP techniques for text categorization, we implemented various methodologies centered around the LLAMA2 model [5] and prompt engineering, aiming to enhance the model's ability to perform zero-shot and few-shot prompting. We decided to use LLAMA2, because of it's ability to process 4096 tokens. The LLAMA2 model is highly effective for prompting and text categorization due to several key features. Its fine-tuning with reinforcement learning from human feedback (RLHF) ensures it generates relevant responses based on context. Optimized for dialogue, it handles conversational inputs well, producing coherent replies. With 13 billion parameters, it has a substantial capacity to understand and categorize diverse text accurately. Additionally, its training on a vast dataset of 2 trillion tokens enhances its generalization across topics. Evaluations confirm its performance in terms of safety and helpfulness, making it a reliable choice for various applications.

We chose to perform zero-shot prompting to demonstrate how highly specific text categorization can be achieved using only basic context provided to the LLAMA2 model. In our case, this context consists of category descriptions (Table 4) derived from a thorough review of the dataset. This approach offers valuable insights into the model's capability to categorize text accurately without extensive pre-training on the specific task.

To select relevant prompts for few-shot prompting, we reviewed the dataset to extract a representative sample of prompts for each category. We ensured the sample proportions matched the actual distribution of categories in the dataset. This approach tests the model's ability to categorize text using examples from the dataset. Additionally, we enhanced few-shot prompting by providing both examples and descriptions of the categories, aiming to improve the model's categorization performance.

Results

For evaluating our BERT models, we will use micro-F1, or average accuracy, as an evaluation metric to determine how well the model is performing.

The best performing model is the large BERT model (*bert-large-uncased*), which has the best accuracy score on the validation dataset. The accuracy is greatly affected by the batch size and the number of epochs in which the model was trained.

Table 1. Results of different BERT models with varying epoch number and batch size. We can see that in general bert-large-uncased performs slightly better than others with highest test and validation accuracy.

Model	Epochs	Batch size	Test accuracy	Validation accuracy
bert-base-cased	3	4	0.728	0.652
bert-base-cased	3	8	0.75	0.685
bert-base-cased	10	8	0.77	0.696
distil-base-uncased	10	8	0.728	0.641
distil-base-uncased	50	16	0.77	0.641
bert-large-uncased	10	8	0.72	0.72
bert-large-uncased	20	16	0.75	0.72

Table 2. Results of traditional AI approaches

Model	Test accuracy	Validation accuracy
Multinomial Naive Bayes	0.739	0.707
Random Forest Classifier	0.685	0.641
Support Vector Machine	0.739	0.641
XGBoost	0.641	0.641

From the traditional AI models we evaluated, Multinomial Naive Bayes performed best on both the test and validation datasets. Similarly, SVM achieved very good results on the test dataset, but the performance dropped sharply on the validation dataset.

Table 3. Results of prompting

Approach	Test accuracy	Validation accuracy
Zero-shot	0.467	0.391
Few-shot	0.130	0.152
Few-shot without category descriptions	0.152	0.174

In Table 3 we can see that the best performing prompting approach is zero-shot prompting, while few-shot prompting performs the worst among the models tested in this paper.

Discussion

We can analyze robustness, accuracy, and performance of a specific model through an in-depth examination of instances where the model fails.

For analysis of the BERT model approach, we used the *bert-base-cased* model, as presented in Table 1, due to its relatively high accuracy and smaller size. We focused on the incorrect predictions made on the test and validation datasets, which together represent 30% of the total data. There were a total of 50 erroneous predictions in this subset. The text length of these misclassified instances is visualized in Figure 2. Notably, 9 errors occurred in examples where the content length was 20 characters or less. Short examples are considerably harder for BERT to classify correctly due to the limited context available for accurate interpretation.

In Table 5, we present several examples with their corresponding true and predicted labels. Upon inspection, it becomes evident that many of these examples pose significant challenges for classification, indicating the need for additional contextual information to make accurate predictions.

To see which discussion types are more prone to errors, we can analyse a confusion matrix of errors presented in Figure 3.

Matrix shows a high amount of examples wrongly predicted as seminar, which may be explained due to large proportion of examples being that type. We can also see a high amount of *UX* examples being classed as *Deliberation*. These errors are often due to the deliberate manner of speech, even though the focus was on UX.

As for the most common misclassification when it comes to the traditional AI models we evaluated, they all most commonly mistook an example which should have been classified as *Deliberation* for *Seminar*. A common pattern was misclassifying examples as *Seminar* or *Deliberation*. Random Forest Classifier, SVM and XGBoost all had similar distributions of wrongly classified examples, whereas Multinomial Naive Bayes mixed up *Procedure* and *Deliberation*, as well as *UX* with *Social*.

Zero-shot prompting performed well on the test set compared to other prompting approaches, due to a relatively good description of the task and category descriptions. However, it shows a larger difference between accuracy on the test and validation sets compared to other models, although all models underperform on the validation set. This can suggest that the performance is dependent on the testing examples. There are numerous instances where LLAMA2 fails to return a category, particularly on the test set. This indicates that LLAMA2 couldn't determine a category based on the given descriptions, as it responded with a plain dash. This suggests that the model struggled with certain inputs (predominantely *Seminar*) and was unable to categorize them effectively using the provided information.

The misclassifications observed in the LLAMA2 zero-shot prompting task can be attributed to several key factors. Primarily, the overlap in category definitions creates significant challenges, as categories such as "Seminar" and "Social" exhibit overlapping characteristics, leading to frequent misclassifications (Figure 4). Furthermore, the presence of ambiguous language within the posts exacerbates classification difficulties, as contextually similar language can be interpreted in multiple ways by the model. Some of ambiguous examples are shown in Table 6. Additionally, the model's tendency to generalize based on frequent patterns encountered during training contributes to misclassification, often defaulting to common categories like "Social" when uncertain. The low precision and recall observed for certain categories indicate that the model struggles to accurately identify and differentiate these categories.

Few-shot prompting has shown limited effectiveness in classifying text into the correct categories. This might be due to our prompts not being distinct enough for LLAMA2 to differentiate between categories. Additionally, even when we included category descriptions and examples, the accuracy did not improve significantly. This suggests that the examples may have introduced noise, making the categories less distinguishable. However, in the case of few-shot prompting, LLAMA2 always returns one of the categories, whereas in the case of zero-shot prompting, it sometimes returns none of the categories.

Conclusion

In this project, we evaluated several text categorization methods, with our primary focus being qualitative discourse analysis. Our dataset was particularly grounded in reality, as it was comprised of forum posts discussing the story "The Lady, or the Tiger?".

Due to the nature of the dataset and the overall lack of clear boundaries between the target categories, the models we evaluated struggled with achieving high results. It's noteworthy to mention however, that even traditional machine learning models like Multinomial Naive Bayes and SVM achieved results comparable to our best performing BERT model. While the experiments we did with prompt engineering and LLMs did not achieve the best results, we find there is a lot of potential and future work to be done in this direction. It was with LLAMA where we had the most interesting "errors", whereas the other models had quite similar misclassification distributions.

Addressing these challenges requires enhanced prompt design with explicit instructions and refined category definitions. Moreover, fine-tuning the model with a domain-specific labeled dataset would enable it to better learn the nuances of each category, thereby reducing misclassification rates. Improving model sensitivity to subtle linguistic cues and reducing misclassification by enhancing training datasets could also help address these issues.

In conclusion, our exploration into the automated categorization of text in online discussions not only advances our understanding of discourse analysis using NLP but also sets the stage for future work to refine these techniques. As AI continues to evolve, its integration into social science research promises to enhance the efficiency and accuracy of qualitative analyses, thereby enriching our understanding of human interaction in digital contexts.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification?, 2020.

[3] Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. Prompt engineering a prompt engineer, 2024.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Hatem Babaei, Nicolas Ballas, Jason Carreira, Mojtaba Hejrati, Max Henaff, Aurelia Guy,

Melanie Kunc, Zhiqing Sun, Anmol Gulati, Ross Wightman, Alykhan Tejani, Myle Ott, Nicolas Usunier, Michal Drozdzal, and Armand Joulin. Llama 2: Open foundation and fine-tuned chat models, 2023.

Appendix

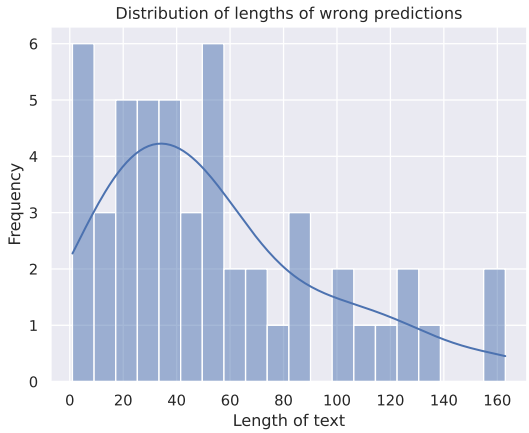

Figure 2. Histogram showing the distribution of the lengths of incorrectly predicted text. Numerous examples have a length of less than 20 characters, posing a challenge for BERT.

Table 4. Descriptions of each category

Category	Description
Seminar	*Posts discussing the deeper meanings of content, encouraging analysis and interpretation.*
Deliberation	*Posts about making decisions, often involving questions and considerations for future actions.*
Social	*Posts to establish or maintain relationships, often casual and friendly in nature.*
UX	*Posts discussing the user experience, including issues and feedback about interfaces and usability.*
Procedure	*Posts about accomplishing a task, often with step-by-step instructions or suggestions for organizing activities.*
Imaginative	*Posts about imaginative content, often involving hypothetical scenarios or creative storytelling.*
Other	*Posts that do not fit into any of the above categories.*

Table 5. Examples of short text misclassifications with true and predicted labels.

Text	True Label	Predicted Label
I like it	Deliberation	Seminar
Submitted	Deliberation	Social
w	Other	Seminar
And that's true!	Imaginative entry	Seminar
Good idea.	Seminar	Social
I am on,	Procedure	Deliberation
I couldn't either.	Procedure	Seminar
it's wonderful!	Seminar	Social
ok	Deliberation	Social
ooooh	UX	Social

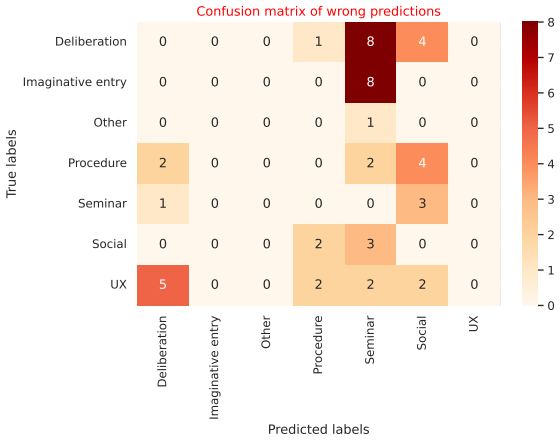

Figure 3. Confusion matrix representing true labels on horizontal and predicted labels on vertical lines. Unsurprisingly, many text cases are classified as seminar, due to its predominance in the dataset.

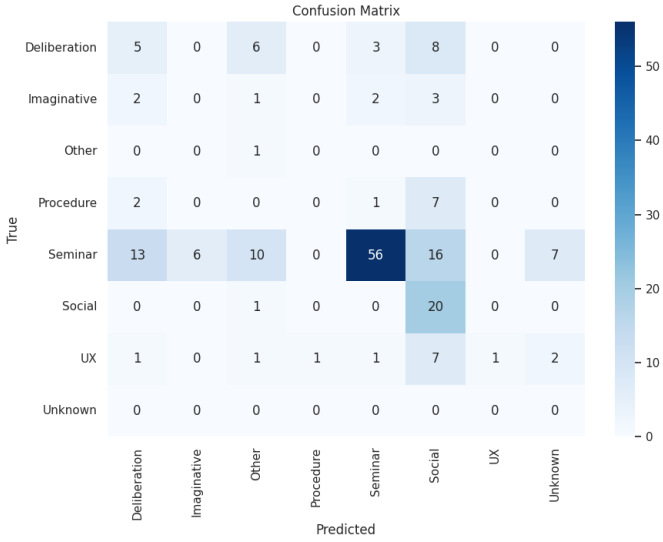

Figure 4. Confusion matrix of classification by zero-shot prompting on combined set of test and validation set. Many text of category *Seminar* cases are classified as *Social* and *Deliberation*, due to overlap in category definitions.

Table 6. Examples of short text misclassifications with true and predicted labels with prompting.

Text	True Label	Predicted Label
Did you guys already read the story? It won't move past the part of four questions for me.	UX	Social
And that's true!	Imaginative	Social
Oh that would be a good punishment for her!	Seminar	Social
Would the king even entertain that idea given it is his daughter?	Seminar	Deliberation
What would happen then?	Seminar	Imaginative
I'm not sure. They're probably equally beautiful and just jealous :)	Seminar	Social