Discourse Parsing Classification Task Using Pre-Trained BERT Model on Close Tasks

Amit Sultan 205975444
Efrat Cohen 207783150
Liat Cohen 205595283
Liel Serfaty 312350622

1 Introduction

In recent years discourse parsing in Natural Language Processing (NLP) has been a topic of interest, given its potential to extract important information and relationships between texts. The growth of online argumentation platforms has given researchers the chance to create computational approaches for a broader-scale investigation of the key elements of persuasiveness, such as language use (Hidey et al., 2017; Tan and Lee, 2016; Zhang et al., 2016), audience traits (such as demographics, preconceived notions) (Durmus and Cardie, 2019b,a) and social interactions (Durmus and Cardie, 2019c). The recognition that discourse units (at their most basic, clauses) are connected to one another in principled ways, and that the conjunction of two units produces a combined meaning greater than each unit's alone, forms the basis of discourse structure (Li et al., 2014).

In this project, we focus on analyzing discourse in the "Change My View" subreddit, where users present arguments for their perspective and engage in discussions to challenge and potentially change their viewpoints (Tan and Lee, 2016; Li et al., 2020). Only the original poster (i.e., if they modify their position) evaluates whether a post is persuasive in CMV (Li et al., 2020). The discussion trees are of high quality since moderators keep an eye on them, but they are less organized because after the original post is made, any member of the subreddit can engage in the debates (Tan and Lee, 2016). The ability to examine the discourse structure of arguments and counterarguments, as well as how users build upon and reply to each other's comments, is provided by this setting's singularity.

Our research asks the question of whether fine-tuning a pre-trained large language model such as BERT (Devlin et al., 2018) on similar texts with different classification tasks, can help improve performance on discourse parsing tasks. The fine-tuning approach has been shown to improve performance on various NLP tasks such as text classification, named entity recognition, and question answering. We examine this theory using two different datasets of the same nature. The first dataset aimed at promoting meaningful and constructive online discussions by identifying constructive comments among online comments. The second dataset contains argumentative essays written by students and the classification task is to detect the presence of various discourse elements commonly found in argumentative writing (see **??**). We found a small improvement in the classification ability of the fine-tuned model.

2 Related Work

The goal of discourse parsing is to identify the discourse elements and their relationships in a text, such as topics, arguments, and perspectives, as well as how they are organized and connected to form a coherent whole. Discourse parsing approaches can be rule-based, statistical, or machine-learning-based, and we will focus on machine-learning-based methods. Rutherford et al. proposed a Naive Bayes classifier using a large amount of unlabeled training data (Rutherford and Xue, 2015) . They gathered clear argument pairs with freely omissible discourse connectives that may be discarded regardless of context without affecting how the discourse connection is understood. Wu et al. present discourse-specific word embeddings (Wu et al., 2017). Discourse connectives from several explicit discourse argument pairs were used to categorize embeddings. Recently, pre-trained transformer encoders such as BERT (Devlin et al., 2019) and SpanBERT (Joshi et al., 2020) have been demonstrated to significantly increase discourse parsing accuracy (Guz and Carenini, 2020; Koto et al., 2021)

Model adaptation is a common approach to leveraging pre-trained models like BERT (Devlin et al.,

2018), as it enables the use of a large pre-trained model without having to train it from scratch on the target task. In his GitHub repository, (Devlin et al., 2018) suggest that if a large corpus of domain-specific data is available for the downstream task, it will probably be advantageous to execute extra pretraining procedures. Qasim et al. proposed a study that used fine-tuning the BRET model on two datasets of COVID-19 fake news to evaluate this approach (Qasim et al., 2022). The BERT models were fine-tuned on a third dataset and tested on the first two. In another study (Kishimoto et al., 2020), further pre-training was performed on text specifically for discourse classification to enhance classification capabilities. The findings from these studies demonstrate the effectiveness of fine-tuning BERT.

3 Data

In this chapter, we will describe the databases we used during the work. Initially, we will present the data which we fine-tune a pre-trained BERT model with, and then the data of the main task of the project.

3.1 Constructive Comments Corpus (C3)

This project [1] aimed at promoting meaningful and constructive online discussions by identifying construed comments among online comments on news articles published online. The data for the project is a subset of comments from the SFU Opinion and Comments Corpus[2], called the Constructive Comments Corpus (C3), consisting of 12,000 comments annotated for constructiveness and its characteristics.
Each comment, written by a certain user on some article, includes several features, such as the level of constructiveness or toxicity of the comment judged by the crowd or an expert, and finally binary labeling of whether the comment is constructive or not. The average number of words per comment is 71.

3.2 Feedback

The data from the Kaggle competition Prize[3] contains argumentative essays written by students in grades 6-12 in the United States. Expert raters

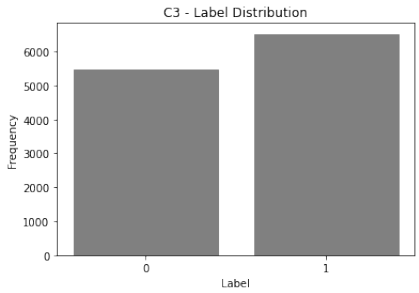

Figure 1: Distribution of the target variable in C3 data - a binary task

have annotated these essays for various discourse elements commonly found in argumentative writing, including the lead, position, claim, counterclaim, rebuttal, evidence, and concluding statement. The original task of the competition is to predict the quality rating of each discourse element as assigned by human readers, which are rated as either "Ineffective", "Adequate", or "Effective" but we used it to predict the discourse element of the paragraph itself. The data contains approximately

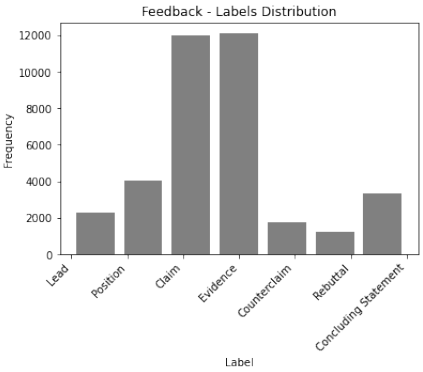

Figure 2: Distribution of the new target variable in Feedback data - a multiclass task

36.7K tagged sentences, with an average length of 45 tokens per paragraph.

3.3 Reddit ChangeMyView

The forum "Change My View" on Reddit[4] provides a space for meaningful debates on contentious topics. Users initiate discussions by presenting their viewpoints and supporting evidence, and then the community evaluates the arguments and discusses on the given issue. The conversations take place in a tree format, where a comment to a single post can lead to multiple branches. The data present in (Zakharov et al., 2021) collected from 16,000

[1] https://www.kaggle.com/datasets/mtaboada/c3-constructive-comments-corpusl
[2] https://github.com/sfu-discourse-lab/SOCC
[3] https://www.kaggle.com/competitions/feedback-prize-effectiveness/data
[4] https://www.reddit.com/r/changemyview

discussions between April 2017 to January 2018 on Reddit's subreddit and transformed into conversation trees. Out of these, 1,946 branches from 101 trees were annotated by professional annotators. The tags are categorized into four main categories, including responsiveness quality and knowledge exchange, style and tone (positive or negative attitudes), and disagreement strategies. The annotated dataset includes 9,620 posts and 17,964 tags (31 unique tags) when posts can be tagged with several tags.

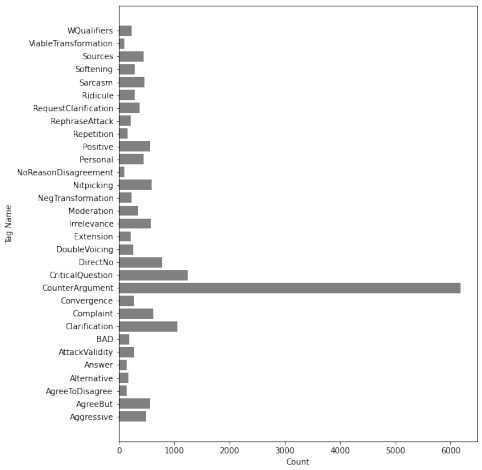

Figure 3: The count of each tag in the discourse data

4 Computational approach

4.1 Software packages and tools

The code developed during this project was implemented as a Python package running on a Linux platform and employs transformers and data analysis packages.

4.2 Task Definition

Discourse parsing involves the analysis and comprehension of text structures beyond the sentence level. It aims to identify the relationships and connections between sentences, paragraphs, and sections of a document or text to understand its overall meaning and structure. To enhance the discourse parsing classification task, we performed model adaption. Model adaptation of the BERT model for the discourse parsing task refers to the process of adapting the pre-trained BERT model to the specific requirements of the discourse parsing task. The goal of model adaptation is to improve the performance of the BERT model on the discourse parsing task by updating its parameters with task-specific information. A common approach to achieve that is fine-tuning the pre-trained BERT model on a smaller task-specific dataset. Fine-tuning adds a task-specific layer or layers to the pre-trained BERT model and trains the combined model end-to-end on the target task data, allowing the model to learn task-specific representations while retaining its pre-trained language understanding. Our approach involved fine-tuning the BERT model to similar task datasets in order to enhance its performance on the discourse parsing task. This approach leverages the pre-trained language understanding ability of the BERT model and updates its parameters with task-specific information to improve its performance on the target task. To achieve that, we first fine-tuned all the layers of the BERT model on two tasks that were similar close to the discourse task (as described in Model adaption). This resulted in two re-trained BERT models. The next step in our approach was to further optimize the two retrained BERT models for the specific requirements of the discourse parsing task. This was done to see if the adaptation of the BERT model to similar tasks would improve its performance on the target task. The two options we evaluated were: one model per tag and one model for all tags. Once the models were trained, we compared their performance on the discourse parsing task to the baseline BERT model, which had only been fine-tuned on the discourse task (as described in Classification Models). The aim of this comparison was to determine the effectiveness of our approach to fine-tuning the models on similar tasks.

4.3 Model adaption

In Task Definition, we discussed the adaptation of the BERT model to different datasets that are closely related to the discourse parsing task. The first dataset is the Constructive Comments Corpus (C3) dataset, where the objective is to classify comments as constructive or non-constructive. Both the C3 dataset and the discourse parsing task share similarities in that they both require the analysis of relationships within the text, such as between words, sentences, and paragraphs. To classify a comment, the model must understand the meaning of words, the relationships between them, and the overall tone of the comment. Similarly, to perform discourse parsing, the model must understand the relationships within the text and the overall discourse structure. The tone of a comment is particularly relevant to discourse parsing, as it gives

important context and insight into the writer's intention. For example, a positive tone may indicate a constructive and solution-focused attitude, while a negative tone may indicate frustration or disagreement. Recognizing the tone of comments is crucial for understanding their purpose and meaning within the larger discourse. By fine-tuning the BERT model on the "Constructive Comments" dataset, the model can learn to recognize the tone of comments and improve its performance on the discourse parsing task. The fine-tuning process provides the model with more data and examples to recognize the relationships necessary for discourse parsing, leading to improved generalization and more accurate predictions. The second dataset, Feedback, aims to predict various discourse elements found in argumentative writing such as the lead, position, claim, counterclaim, rebuttal, evidence, and concluding statement in feedback messages. This task shares similarities with the discourse parsing task in terms of the skills required from the model, as both involve the analysis and categorization of text, requiring the model to understand relationships between elements of the text. Similar to C3, by fine-tuning the BERT model on this dataset, the model can learn to identify the relationships between words, sentences, and paragraphs in feedback messages, which are also useful for the discourse parsing task. The advantage of this dataset is its size, which can lead to the model learning more robust language representations and improving its performance on the discourse parsing task. Furthermore, the dataset has a large amount of text data related to discourse elements in argumentative writing, allowing the BERT model to gain the ability to identify task-specific features upon fine-tuning and improve its performance on the discourse parsing task. In conclusion, fine-tuning the BERT model on these datasets can help the model learn relevant skills, robust language representations, and task-specific features, leading to improved performance on the discourse parsing task.

The input for a fine-tuning BERT model is typically the text of the argumentative essays or comments. The BERT model has a set limit of 512 tokens for its input sequence. Despite the limit, most of the text in the C3 and Feedback Prize dataset is still shorter than 512 tokens, making it suitable for processing by the standard BERT model.

4.4 Classification Models

The discourse parsing task is a complex and challenging problem, as it involves assigning multiple class labels to a single input sample. To address this multi-class problem, we explored and evaluated two different approaches: one model per tag and one model for all tags. The classification of the sample was determined by predicted probabilities exceeding a specified threshold. The sample classification encompasses all the labels that had predicted probabilities equal to or above the established threshold value. **Model for each Tag:** In this approach, one binary classifier is trained for each class. Each classifier is trained to predict whether an input sample belongs to its class or not. We created 31 models as the number of tags. **Model for all Tags:** In this approach, a single model is trained to predict the likelihood of each class. This model is usually a multiclass classification model. The input for a BERT model is the text of the post. Since the limitation of the input for the BERT model, in the discourse parsing dataset, a small percentage of the posts were truncated. Furthermore, an additional step is implemented to integrate crucial information into the BERT model by incorporating the labels of the previous two posts in the same branch, similar to the sequence input presented in (Zakharov et al., 2021). For example: "previous tags: tag Counter Argument Sources tag Agree But Counter Argument Critical Question, current post: I guess I agree with you..." This structure input sentence allows for the retention of important information, despite the limitation of the input sequence.

5 Results and analysis

To evaluate our approach, 85% of the trees were designated for training and the remaining were utilized for testing the various models. This partitioning of the data into training and test sets based on trees is crucial as it ensures the preservation of the relationships between the posts, represented by the tree structure, for a more accurate evaluation of the model. The BERT model was fine-tuned on the discourse parsing dataset by training it for 5 epochs, with the aim of improving its performance for discourse parsing tasks. Additionally, for other adaptation tasks, the model was trained for 3 epochs (the learning results of each epoch are presented in Table 1).

Model for all tags: As described above, we eval-

	Train Loss	Precision	Recall	F1
	C3 data			
Epoch 1	0.218	0.9525	0.9459	1.0
Epoch 2	0.1726	0.9209	0.9547	0.947
	Feedback data			
Epoch 1	0.9028	0.7793	0.3896	0.5194
Epoch 2	0.642	0.7544	0.3095	0.438

Table 1: Results of different metrics, on the first two epochs for each dataset. The C3 achieved better results with the binary task

uated utilizing in a single model for all tags and employing a BERT model that was pre-trained on a diverse dataset. The pre-trained BERT model on the Feedback dataset achieved the highest F1-Score of 0.46. This is a remarkable accomplishment, especially considering the significant number of annotations. Table 2 illustrates that this approach outperformed the baseline model, which achieved a score of 0.419, by a substantial margin. These results demonstrate the effectiveness of pre-training the model on diverse data in comparison to classification with a model that pre-trained on only the discourse data.

	F1	Precision	Recall	Accuracy
Base-line	0.4193	**0.667**	0.305	0.152
Pre-trained C3	**0.436**	0.591	**0.345**	0.148
Pre-trained Feedback	**0.46**	0.655	**0.337**	**0.160**

Table 2: **Model for all tags:** in most metrics, the pre-training on other data achieved better results than the baseline

Model for each tag: In our examination of the second method, we utilized a distinct classification model for each annotation tag. However, due to the limited occurrence of some tags, the individual performance was subpar. As a result, we presented an average score for each category by summing the scores of all annotations within the category and dividing them by the number of tags. These results are documented in Table 3.

It was observed that the best performance was achieved when the model was pre-trained on external data, especially on the Feedback dataset, which resulted in the most optimal results across all categories. The results indicate that the "Promoting Discussion" category demonstrated superior performance, a trend that can be attributed to the prevalence of the "CounterArgument" tag within the dataset (figure 3). It is reasonable to expect that a high frequency of a particular tag within the dataset would positively impact the performance of the category to which it belongs, as the model would have more opportunities to learn from and make predictions about this tag.

		Precision	Recall	F1
Promoting discussion	Base-line	0.559	0.529	0.543
	C3	0.566	0.531	0.548
	Feedback	**0.578**	**0.541**	**0.559**
Low responsiveness	Base-line	0.354	0.308	0.329
	C3	0.367	0.31	0.3361
	Feedback	**0.393**	**0.35**	**0.37**
Tone and style	Base-line	0.121	0.31	0.174
	C3	0.122	0.311	0.175
	Feedback	**0.156**	**0.322**	**0.21**
Disagreement	Base-line	0.291	0.539	0.377
	C3	0.283	0.532	0.369
	Feedback	**0.304**	**0.541**	**0.389**

Table 3: **Model for each tag:** micro-average of the classification results of the tags in each category, according to the different metrics.

6 Conclusion

In conclusion, this study proposed a fine-tuned BERT-based approach to solving discourse parsing tasks, by performing fine-tuning on two close task datasets. The results showed a slight improvement in classification performance compared to the baseline models (fine-tuned BERT on the original dataset). These results indicate the potential of fine-tuned BERT models for discourse parsing, however, the improvement was not statistically significant. This finding highlights the need for further exploration and refinement of the proposed approach. Additionally, the results of the experiment demonstrate that the Feedback dataset outperforms the baseline in F1 score. This improvement can be attributed to the fine-tuning of BERT on the Feedback dataset, which allowed the model to better comprehend the discourse parsing task. The close similarity of the task to the target task, the large size of the dataset, and the multi-class nature of the task are crucial factors that contribute to the success of the model.

Future studies may aim to boost the effectiveness of fine-tuned BERT models by incorporating extra elements or models such as graph representations or attention mechanisms. To refine the fine-tuned BERT model, modifications to its architecture could also be made, such as incorporating a task-specific layer atop pre-trained BERT representations to enhance task fit. Furthermore, applying different learning rates to various layers during fine-

tuning can improve performance by allowing the model to maintain pre-trained representations while still adapting to the new task. Overall, this study advances our understanding of NLP discourse parsing and provides a step forward in developing more effective and sophisticated methods for analyzing the discourse structure of text data. The findings of this study contribute to the ongoing research in this field.

References

Jacob Devlin, Ming-Wei Chang, and Kenton Lee. 2019. Google, kt, language, ai: Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esin Durmus and Claire Cardie. 2019a. A corpus for modeling user and language effects in argumentation on online debating. *arXiv preprint arXiv:1906.11310*.

Esin Durmus and Claire Cardie. 2019b. Exploring the role of prior beliefs for argument persuasion. *arXiv preprint arXiv:1906.11301*.

Esin Durmus and Claire Cardie. 2019c. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, pages 2701–2707.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. *arXiv preprint arXiv:2102.02080*.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. *arXiv preprint arXiv:2010.03538*.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

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

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *HLT-NAACL*, pages 799–808.

Chenhao Tan and Lillian Lee. 2016. Talk it up or play it down?(un) expected correlations between (de-) emphasis and recurrence of discussion points in consequential us economic policy meetings. *arXiv preprint arXiv:1612.06391*.

Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. 2017. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 269–274.

Stepan Zakharov, Omri Hadar, Tovit Hakak, Dina Grossman, Yifat Ben-David Kolikant, and Oren Tsur. 2021. Discourse parsing for contentious, non-convergent online discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 853–864.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*.