Unleashing the Power of Discourse-Enhanced Transformers for Propaganda Detection

Alexander Chernyavskiy[1,2], **Dmitry Ilvovsky**[1], **and Preslav Nakov**[3]
[1] HSE University, Russia
[2] Sber, Russia
[3] Mohamed bin Zayed University of Artificial Intelligence, UAE
alschernyavskiy@gmail.com, dilvovsky@hse.ru
preslav.nakov@mbzuai.ac.ae

Abstract

The prevalence of information manipulation online has created a need for propaganda detection systems. Such systems have typically focused on the surface words, ignoring the linguistic structure. Here we aim to bridge this gap. In particular, we present the first attempt at using discourse analysis for the task. We consider both paragraph-level and token-level classification and we propose a discourse-aware Transformer architecture. Our experiments on English and Russian demonstrate sizeable performance gains compared to a number of baselines. Moreover, our ablation study emphasizes the importance of specific types of discourse features, and our in-depth analysis reveals a strong correlation between propaganda instances and discourse spans.

1 Introduction

The widespread of disinformation and information manipulation in various domains, such as politics, economics, and health (e.g., COVID-19), has led to an increased demand for fact-checking and propaganda detection. To tackle this, several datasets have been created to analyze online media and assist in system development (Martino et al., 2020; Maarouf et al., 2023). One of the recent competitions in this field is Semeval 2023 Task 3 (Piskorski et al., 2023), which introduces a multilingual dataset with six languages. This paper specifically addresses the most challenging task in the competition, namely persuasion techniques detection.

For propaganda detection, the most recent effective approaches employ encoder-based Transformers as their backbone models, with some minor task-specific or dataset-specific modifications (Jurkiewicz et al., 2020; Liao et al., 2023; Wu et al., 2023). However, these approaches have aimed mostly to achieve the highest quality or competition scores.

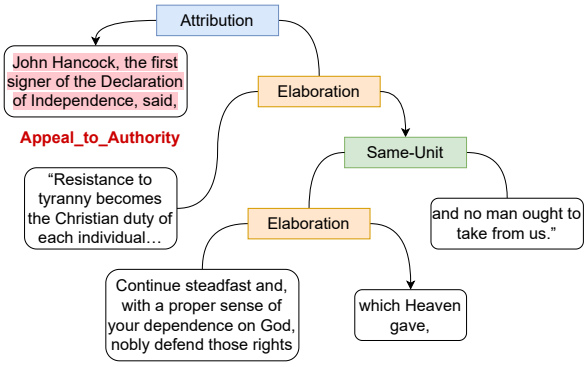

Figure 1: A discourse tree for a piece of text annotated with propaganda labels. The arrows point from nuclei to satellites.

In contrast, we adopt a broader perspective and aim not only to develop a qualitative approach, but also to gain interpretation and understanding that can facilitate progress in propaganda detection. To this end, we analyze the linguistic structure of the texts by examining the discourse features.

Among the various types of discourse representations available, the Rhetorical Structure Theory (RST) suggested by Mann and Thompson (1988) was selected for our analysis due to its widespread usage and availability of high-quality open-source parsers for multiple languages. According to this theory, a text can be represented as a tree structure, where the nodes correspond to text spans and are connected by discourse relations such as Elaboration, Joint, and Condition. Figure 1 illustrates an example of a discourse tree constructed for a real media text, with propaganda spans annotated. Notably, the "Attribution" discourse span aligns precisely with the "Appeal to Authority" propaganda span. It is reasonable to observe this alignment, and further examples can be found in the dataset. However, establishing a definitive set of rules for matching between discourse and propaganda is quite challenging.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1452–1462

The primary objective of our research is to enhance the effectiveness of a Transformer-based approach for propaganda detection by integrating discourse information. This allows not only to improve the quality of the model, but also to evaluate the relationship between discourse and propaganda.

Our contributions can be summarized as follows:

- We modify a Transformer-based architecture in order to integrate discourse features.
- The proposed approach greatly improves the performance of the base model in text classification and token classification tasks for the SemEval-2023 dataset in both English and Russian languages.
- An ablation study is conducted to evaluate the importance of specific discourse features.
- An in-depth analysis is conducted to examine the errors made by the discourse-based model and to investigate the actual correlations between discourse and propaganda.

2 Related Work

General View A fine-grained propaganda analysis was proposed by Martino et al. (2019), who developed a corpus of news articles annotated with 18 propaganda techniques, considering separately the task of technique spans detection and classification. Subsequently, the dataset and its extentsions were employed in several shared tasks, such as SemEval-2020 (Martino et al., 2020) and SemEval-2023 (Piskorski et al., 2023).

Transformer-based approaches have become common in solving the propaganda detection task, treating it as a token- or text classification problem (Jurkiewicz et al., 2020; Wu et al., 2023). Recent studies have utilized BERT-based models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), ALBERT (Lan et al., 2020), or their ensemble (Purificato and Navigli, 2023). Liao et al. (2023) considered XLM-RoBERTa and extended the base approach by using a contrastive formulation of the loss function. Another approach proposed by Baraniak and Sydow (2023) introduced a BERT-based hierarchical model that combines token classification and multilabel token classification tasks.

We also focus on Transformer-based approaches, but our analysis goes beyond the typical modifications associated with task formulation or base dataset characteristics.

Instead, we additionally investigate the nature of propaganda in terms of discourse structure in order to enhance the interpretability of our approach.

Discourse In previous studies, the effectiveness of discourse-aware measures has been demonstrated in evaluating the quality of machine translation (Joty et al., 2017). We also consider some discourse-based characteristics, but incorporate them as features in our neural approach. Similarly, Xu et al. (2019) enhanced extractive summarization by combining discourse-based representations with BERT embeddings.

The fact-checking task is closely related to propaganda detection, as both involve analyzing the reliability of information. Previous research has shown that discourse integration techniques have been effective for this task. For instance, Karimi and Tang (2019) developed the multitask model that incorporated discourse tree reconstruction as an auxiliary loss. Chernyavskiy and Ilvovsky (2020b,a) utilized pre-constructed discourse trees and encoded them using a recursive neural network.

Regarding interpretable approaches, Yu et al. (2021) conducted a study on classification-based methods for propaganda detection. Nevertheless, their analysis primarily concentrated on syntactic and sentiment features, neglecting discourse features. To fill this gap, our work investigates the relationship between discourse trees and propaganda spans, and underscores the importance of specific discourse features in enhancing the efficacy of neural approaches. Finally, an analogy can be drawn with the study conducted by Rodríguez et al. (2023), which suggested multi-task learning with propaganda identification as the main task and metaphor detection as an auxiliary task.

3 Methods

3.1 Preliminaries: RST

Rhetorical Structure Theory (RST) was proposed by Mann and Thompson (1988). It posits that each text can be represented as a tree structure, which is constructed incrementally from the bottom-up. The first step in RST analysis involves identifying and segmenting the text into elementary discourse units (EDUs), which are indivisible coherent units of thought. These EDUs serve as the leaves of the tree structure. Once the EDUs are identified, the text spans are connected recursively using discourse relations, such as “Summary”, “Attribution”, and “Condition”.

RST categorizes vertices into two types: "Nucleus" and "Satellite". Nucleus vertices contain essential information, while Satellite vertices provide additional details. Certain relations, such as "Joint" and "Same-Unit", can be multi-nuclear.

Figure 1 demonstrates an example of a discourse tree for a text from the SemEval-2023 dataset. This tree comprises five EDUs, with the main Nucleus EDU being "Resistance to...", since all other nodes are achievable from this one by arrows.

The theory is language-independent and parsers have been already developed for multiple languages. In the case of English, we utilized the Two-Stage discourse parser (Wang et al., 2017) for its public availability and its proven state-of-the-art performance in discourse parsing. Consequently, we employed the list of discourse relations provided by this parser. Among the languages considered for the SemEval-2023 competition, we have identified a parser for the Russian language, as proposed by Chistova et al. (2020). Although it has slightly lower quality compared to the English parser, it covers a similar set of discourse relations.

3.2 Discourse Features

Considering the multicomponent nature of the discourse structure, we distinguish several types of discourse features.

EDU boundaries (EDUB) This is a per-token binary feature that identifies whether the token represents the beginning or the end of an EDU. For instance, in Figure 1 the token "which" would be assigned the value of 1, indicating that it represents the start of an EDU, while "Heaven" would be assigned the value of 0. This rather simple feature can be valuable as propaganda spans often consist of multiple concatenated EDUs.

Nucleus-Satellite (NucSat) This is a per-EDU binary feature that indicates whether a node in a discourse tree is classified as Nucleus (label 1) or Satellite (label 0). It can be easily projected to the token level, since each token is associated with only one EDU. In Figure 1, all tokens that are pointed to by arrows will be assigned the label 0. This feature is particularly relevant when the specific relation name is not of primary importance, but rather the presence of secondary information is important. For instance, propaganda spans belonging to the "Distraction" class often assume the inclusion of a Satellite either within the phrase itself or in the nearest dependent phrase.

Relations This feature encodes the discourse relations of the corresponding EDUs. Here, we use one-hot encoding to transform the relations into integer vectors. In should be emphasized that relations from trees are assigned only to Satellite nodes, while Nucleus nodes are given a default "span" relation. Thus, the final feature has a size of $N + 1$, where N represents the number of discourse relations in the chosen discourse parser. Again, each token has a vector representation that is equal to the vector representation of the corresponding EDU.

Positions All features described above only consider the entire EDU information and did not take into account the positions of tokens within the tree. To address this limitation, we introduce a discourse-based positional feature, which consists of two parts: absolute position and path-based position.

The absolute position represents the EDU number in the discourse tree constructed for the entire text. This feature is particularly useful when analyzing large texts that are divided into paragraphs and it is needed to consider the relative position of paragraph spans in the overall tree structure.

The path-based position is based on the path from the root to the corresponding EDU leaf in the discourse tree. This path is constructed sequentially, assigning -1 when moving to the left vertex, and 1 otherwise. Therefore, the path can be represented as a binary vector with a length not exceeding the depth of the tree. To facilitate analysis, we truncate the path and retain only the last p values. Moreover, to ensure equal final vector length, we pad shorter vectors with zeros on the left side.

In the given example depicted in Figure 1, the node "which Heaven gave" is assigned an absolute position of 4. Additionally, its path-based position is represented as $(0, 0, 0, 1, 1, -1, 1)$ for $p = 7$. The dimension of the full position encoding is $p+1$.

Depth Considerations The previously described features only encode the individual leaves (EDUs) and do not consider high-level tree relations that connect spans containing multiple EDUs. In the example in Figure 1, the tokens from the span "and no man ought to take from us" are not only "Same-Unit" tokens, but also "Elaboration" tokens at a lower depth. To incorporate these high-level relations, we expand the NucSat and the Relation representations by concatenating the embeddings for all nodes located at a maximum depth of k from the leaves. We pad with zeros the embeddings of the vertices located at depth less than k.

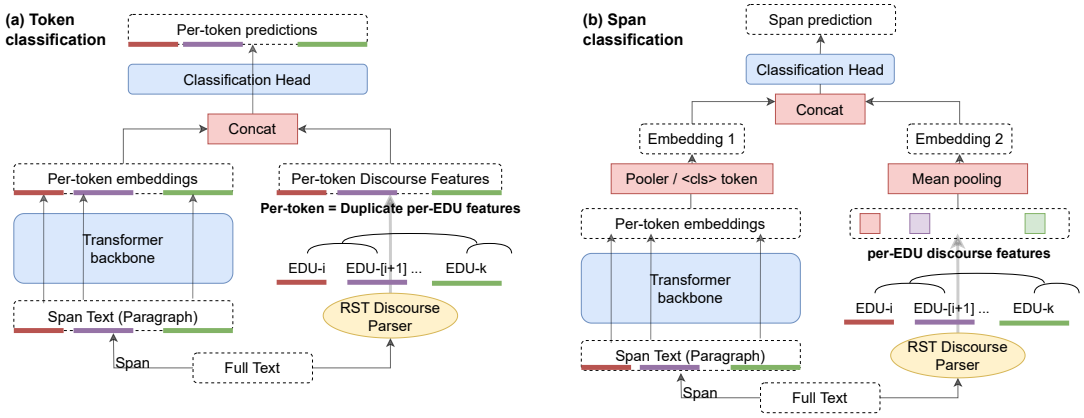

Figure 2: Model architecture for the two tasks: (a) token classification; (b) span classification (paragraph-level). The trainable blocks of the model are indicated in a blue color. The colors in "Span Text" indicate individual EDUs, and only these EDUs are used to calculate discourse features in the entire discourse tree.

For a depth limit of $k = 3$, the span we are considering should be encoded using the following sequence: [(Nucleus, Attribution), (Satellite, Elaboration), (Nucleus, Same-Unit)].

The **final discourse vector representation** is obtained by concatenating the encodings and has a dimension of EDUB + k(NucSat + Relations) + Positions = $1 + k(1 + (N + 1)) + (p + 1)$.

3.3 Model Architecture

The propaganda detection task can be formulated in several ways. In the SemEval-2023 competition (Piskorski et al., 2023), the final quality is assessed at the level of paragraph multilabel classification. At the same time, the competition also provides the token-based markup. Therefore, we consider two task formulations: the span classification and the token classification tasks.

Token Classification In the token classification task, we consider token-level embeddings constructed using two model's branches as depicted in Figure 2 (a).

The first branch is a trainable neural architecture that encodes tokens using a standard Transformer-based backbone. This branch effectively learns representations for tokens by considering their contextual information within the given span or paragraph. The second branch considers the entire text and extracts discourse features, as described in Section 3.2. So, it utilizes the complete discourse tree constructed by the discourse parser and computes non-trainable features for the corresponding tokens located solely within the EDUs of the respective text span.

Hence, even if the span consists of one or two EDUs, the discourse features will consider the higher-level discourse dependencies across the entire text.

The two types of embeddings are combined using a concatenation. Subsequently, each token representation is passed through a trainable classification head to obtain the final per-token predictions. The classification head is comprised of one or two linear layers that are separated by the RELU nonlinearity and a dropout layer.

Span Classification The architecture for the span classification task is presented in Figure 2 (b) and exhibits several distinctions. Firstly, it employs a Transformer backbone to compute the embedding for the entire span. This can be achieved either through a pooling layer, similar to the approach used in original DeBERTa, or by leveraging an embedding solely for the ⟨CLS⟩ token, as done in BERT or RoBERTa.

Additionally, in contrast to projecting EDU embeddings (per-EDU discourse features in Figure) to the token level, we use a mean pooling strategy to compute a full discourse-based representation. Finally, the classification head is solely applied to a concatenated vector that represents the complete text span.

Loss To address an imbalance in the propaganda classes, we utilize a weighted cross-entropy loss function (token-based or span-based). The weights are calculated based on the distribution in the training dataset. To prevent excessive dispersion, we have set a maximum weight limit of 70.

Model	Setting	micro F1	macro F1
DeBERTa	base	0.3287	0.1621
	+ NucSat	0.3423	0.1692
	+ Relations	**0.3751**	**0.1801**
	+ Positions	0.3620	0.1733
Purificato et al. (2023)		**0.3756**	0.1292
Wu et al. (2023)		0.3680	0.1719

Table 1: Performance on paragraph classification for English. The quality is compared to the quality of the top models proposed during the competition. EDUB features are only applicable on the token level. The standard deviation is less than 0.008 in all cases.

Setting	micro F1	macro F1
(1): base	0.1433	0.0908
(2): (1) + EDUB	0.1466	0.0923
(3): (2) + NucSat	0.1569	**0.0991**
(4): (3) + Relations	0.1542	0.0900
(5): (3) + Positions	**0.1596**	0.0956

Table 2: Performance of the DeBERTa-based models on token classification for English. Metrics for ensemble approaches from the competition are not available.

4 Datasets

In our experiments, we utilized the dataset provided by the SemEval-2023 Task 3 Subtask 3 competition (Piskorski et al., 2023), which is devoted to the persuasion technique detection task. This dataset is regarded as the most relevant and comprehensive publicly available resource. It encompasses news articles in six different languages, each of which has been annotated by experts using a set of 19 labels. Articles have been pre-partitioned into train, validation, and test sets in competition. We employed these sets and compared our results with the best approaches suggested during the competition.

To investigate the discourse structure, we employed pre-trained discourse parsers. While this approach has certain limitations, such as potential errors in the parsers, it provides a universally applicable approach that does not depend on human resources. We have identified publicly available parsers with *MIT license* for two languages: English (Wang et al., 2017) and Russian Chistova et al. (2020); and compared results for these languages.

In this research, we utilize Transformers, eliminating the need for any specialized preprocessing techniques. The only preprocessing was to remove all non-ascii characters from the articles.

5 Implementation Details

We fine-tuned the base-sized DeBERTa-v2 (He et al., 2021) for English and the base-sized XLM-RoBERTa (Conneau et al., 2020) for Russian. These models have 184M and 125M parameters respectively. The maximum sequence length was set to 256 in all cases: we selected this value by analyzing the training set. The models were trained on batches of size 16, with a learning rate of 3e-5, for 20-40 epochs. For all other hyper-parameters, we used the default values.

Regarding the hyper-parameters related to the discourse features, we selected $k = 2$ and $p = 7$ using grid search on the development set.

We trained each model (setting) on a Tesla V100 32G GPU for approximately two hours.

6 Results

6.1 Experimental Results

In this research, we initially conducted experiments using the English dataset. The results for the paragraph classification and token classification tasks are presented in Table 1 and Table 2 respectively. As in the competition, we focused on the micro-averaged F1 and macro-averaged F1 scores, and specifically regarded micro F1 as the primary quality metric. At the same time, macro F1 assesses the performance of infrequent classes that may have a stronger correlation with discourse, and therefore is indicative in our case. For the token classification task, we employed BIO labeling and measured the performance by considering only the tokens that have a predicted or true tag other than "O".

To evaluate the relative effectiveness of discourse features, we incrementally incorporated these representations into our approach, starting from simple ones and progressing to more complex ones. We did not use EDUB features to classify spans, since they only applicable at the token level. Our results demonstrate that generally discourse features exhibited quality enhancements in both tasks. Notably, the most sizeable improvements were observed when integrating discourse relations into the classification of spans, as well as when incorporating discourse types (Nucleus/Satellite) into the token classification task.

In contrast, the inclusion of positional embeddings resulted in a marginal enhancement, and only in the token classification task. This indicates that these features might have a limited or potentially negative impact on the overall performance.

Model	micro F1	macro F1
XLM-RoBERTa base	0.2411	0.1814
XLM-RoBERTa disco.	0.3120	0.2064
Hromadka et al. (2023)	0.3868	0.1888
Wu et al. (2023)	0.3184	0.2052

Table 3: Performance on paragraph classification for Russian. A standard deviation is less than 0.012.

This could be attributed to the fact that propaganda spans do not necessarily align with the beginning and the end of a text, but rather can be uniformly distributed over the text (discourse tree). However, positions within a sentence, such as indicators of introductory phrases, might still provide valuable insights.

Most of the methods proposed during the competition employed an ensemble of multiple Transformers, potentially with modifications to the loss function, such as the incorporation of weights. At the same time, the primary objective of our research is not to attain state-of-the-art results or construct extensive neural ensembles. Instead, we aim to investigate the impact of discourse on Transformers and propose a universal approach that can be easily integrated with more complex methods. Nevertheless, we also conducted a comparison of our model with the approaches proposed during the official competition. The results in Table 1 demonstrate that the discourse-enhanced DeBERTa achieved the best macro F1 score and almost the best micro F1 score in the paragraph classification task. This indicates that the incorporation of discourse features even into the base Transformer model can result in a substantial quality improvement and outperform complex ensemble approaches.

In order to enhance our findings, we additionally performed experiments on the Russian language using the XLM-RoBERTa model. Table 3 demonstrates the corresponding results. Here, we utilized all embeddings except for the positional ones to train the discourse-based model. It can be seen that the incorporation of discourse information led to a sizeable improvement in the performance metrics of the base model. As a result, the discourse-enhanced XLM-RoBERTa achieved comparable results to the top-performing approaches in the competition, particularly in terms of macro F1. These findings demonstrate the universality of our proposed approach, as it can be effectively applied to different languages and Transformer architectures.

Label	Freq.	F1 disco	F1 base
Loaded Lang.	22.94	**0.594**	0.520
Repetition	11.24	0.037	**0.052**
Exag.-Minim.	10.57	**0.361**	0.269
Flag Waving	4.63	**0.306**	0.212
Slogans	3.69	**0.360**	0.185

Table 4: Macro F1 scores for the base and discourse models in paragraph classification for English. Frequency is shown as a percentage of total paragraphs.

Label	Freq.	F1 disco	F1 base
Loaded Lang.	22.94	**0.254**	0.231
Name Calling	17.19	**0.406**	0.366
Doubt	8.54	**0.159**	0.141
Appeal to Fear	6.78	0.124	**0.166**
Slogans	3.69	**0.303**	0.210

Table 5: Macro F1 differences for token classification.

6.2 Error Analysis

To evaluate the impact of different classes on overall quality, we assessed the quality of each propaganda class (persuasion technique) individually. This evaluation was performed by calculating the binary F1 scores for the base model and the best discourse-enhanced model. Table 4 and Table 5 present the classes that exhibited most indicative differences in the paragraph classification and token classification tasks respectively.

In both cases, the best quality is primarily attained through enhancements in frequency classes, such as "Loaded Language" and "Name Calling". However, we should note that there was a slight decline in the frequent "Repetition" class in the context of span classification. This suggests that, at the paragraph level, discourse features exhibit a relatively weak correlation with the "Repetition" propaganda technique.

The set of classes exhibiting the most substantial improvements differs between the two tasks. Nevertheless, the class "Slogans" is present in both cases, and it demonstrates the highest relative improvement. Furthermore, improvements are also evident for less common classes such as "Exaggeration Minimisation" and "Doubt". In the following section, we endeavor to provide an interpretation for these improvements.

It is important to highlight that despite incorporating external information through discourse and weights in the loss function, the effectiveness in accurately classifying rare propaganda classes still remains negligible or close to zero.

Instances of such classes include "Whataboutism", "Red Herring", and "Appeal to Popularity". Nevertheless, we have established correlations with discourse structure for them (see Section 7) that can enhance the quality.

7 Discussion

This section aims to elucidate the importance of incorporating discourse structure in the propaganda detection task and investigate the interpretable correlation between propaganda spans and specific types of discourse features.

EDU boundaries As in the discourse features construction approach described in Section 3.2, we firstly analyzed the intersection of EDUs and propaganda spans. To investigate intuitive correlations, we utilized token-level labelling to calculate character-based mean Intersection over Union (mIOU) scores. Specifically, we examined the EDUs that included a given propaganda span and calculated the ratio of its length to the total length of these EDUs. This process was performed for all propaganda spans belonging to the selected class, and the resulting values were averaged.

Table 6 shows the obtained mIOU scores and frequencies for each of the 19 persuasion techniques. The correlation between the most frequent classes and EDUs is relatively small, as these classes are more associated with individual words rather than entire spans. In contrast, the rare classes are linked to specific speech patterns and have a strong connection with EDUs. However, due to their infrequency, it is challenging to achieve high recall scores for these classes. Nonetheless, we observe that in 9 out of the 19 propaganda classes, the intersection with the corresponding EDUs is above 80%, indicating a substantial correlation.

Discourse Types Our another objective was to investigate how discourse node types, specifically Nucleus and Satellite features, can improve classification accuracy. To this end, we focused on examining types of EDUs and measured the percentage of propaganda spans that were encompassed by Nucleus leaves.

The results in Table 6 illustrate that the propaganda spans primarily occur within Nucleus EDUs (the proportion exceeds 0.5 in all cases). Notably, in 6 out of the 19 classes, the proportion actually surpasses 0.7.

Label	Cnt.	mIOU	Nuc.
Loaded Language	1671	0.404	**0.71**
Name Calling	887	0.379	**0.75**
Repetition	496	0.389	0.69
Doubt	391	0.895	**0.70**
Exaggeration-Minimiz.	328	0.629	0.67
Appeal to Fear	269	**0.830**	0.67
Flag Waving	239	0.659	0.64
Causal Oversimplif.	179	**0.910**	0.59
Appeal to Authority	129	**0.878**	0.67
Slogans	116	0.661	0.64
False Dilemma	97	**0.882**	0.65
Conversation Killer	73	0.796	**0.72**
Guilt by Association	50	0.760	**0.71**
Red Herring	42	0.681	0.64
Appeal to Hypocrisy	24	**0.880**	0.65
Obfuscation Confusion	15	**0.820**	0.68
Appeal to Popularity	15	**0.916**	**0.89**
Straw Man	12	**0.937**	0.66
Whataboutism	9	**0.933**	0.56

Table 6: mIOU scores (based on propaganda spans and EDUs) and Nucleus-based coverage of propaganda spans for the English dataset. The indicative maximum values are highlighted in bold and the minimum values are underlined.

These classes include the most common ones, such as "Loaded Language" and "Name Calling", and are generally located in the parts of the text containing the main idea. At the same time, some propaganda instances, such as "Causal Oversimplification" and "Whataboutism," can also be frequently found in the Satellites. This suggests that propaganda can be employed to complicate the primary concepts of a text, substantially influencing the structure and content of discourse.

Is should be emphasized that this correlation is not symmetrical. While the majority of propaganda spans are found within Nucleus nodes, only a small percentage of Nucleus words are involved in propaganda spans, typically ranging from 3% to 6%.

Discourse Relations Similarly to the analysis of node types, we considered the coverage of propaganda spans by discourse relations and vice versa. To calculate the coverage of spans A relative to spans B, we divided the sum of the lengths of spans from the intersection of A and B by the sum of the lengths of all spans in A. We performed the summation across all documents. The results are shown in Figure 3. We can see that propaganda spans are frequently observed in the most prevalent relation types: Elaboration, Joint, and Same-Unit.

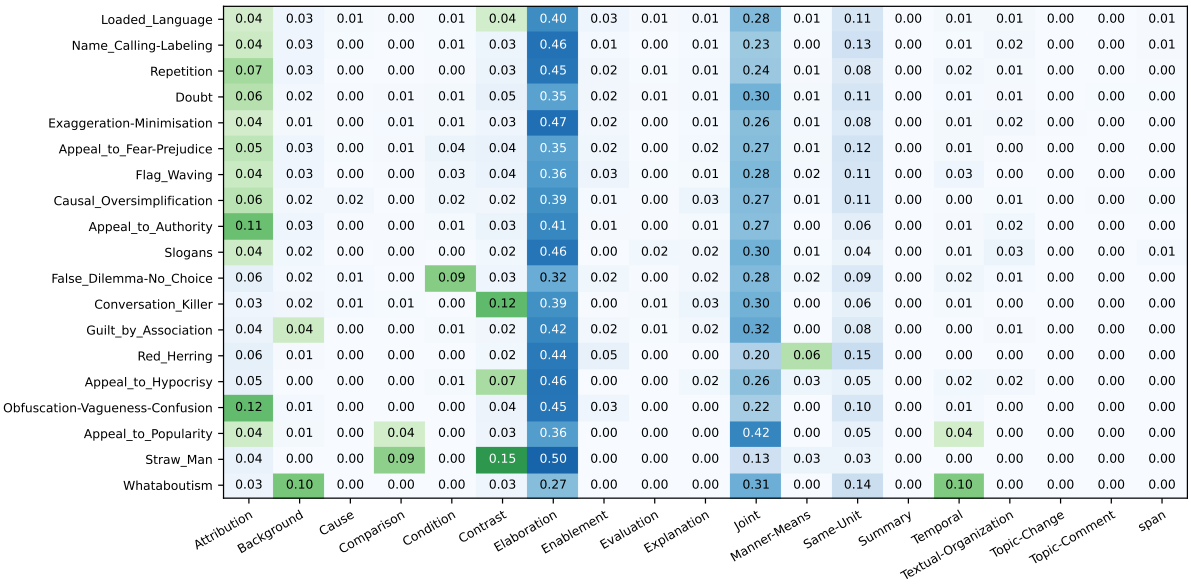

Figure 3: Propaganda spans coverage by discourse relations (character-based proportions). All values have been rounded to the second decimal place. For each propaganda class, the most covered discourse relations are highlighted in green (excluding three default and the most popular relations, namely Elaboration, Joint and Same-Unit).

Nonetheless, the following non-trivial correlations are identified: "Appeal to Authority" and "Obsucfation" tend to contain Attribution; "False Dilemma" - Condition; "Conversation Killer" - Contrast, "Straw Man" - Comparison and Contrast; "Whataboutism" - Background and Temporal.

In this research, we also focus on inverse correlations rather than direct ones, as we incorporate discourse relations as features into our model. We observe that propaganda spans are infrequent, resulting in a considerable number of zeros in the corresponding coverage table (see Appendix A).

We find that 10% of the Summary relations are covered by the "Loaded Language" class; 13% of the Topic-Comment relations are covered by the "Doubt" class; whereas the "Exaggeration-Minimisation" class is primarily associated with the Comparison relation. Additionally, we observe a correlation between the "Slogans" class and the Evaluation relation, as well as between the "Appeal to Fear-Prejudice" class and the Condition relation.

General Summary We can conclude that there are discernible correlations between discourse features and propaganda spans. However, the model did not learned rare classes due to our major optimization of micro F1 and emphasis on the most frequently occurring classes. Besides, there are various ways of methods of feature construction and encoding, and their investigation is one of the directions for further research.

Furthermore, it can be inferred that Transformers exhibit a fundamental understanding of discourse, as evidenced by the fact that EDU boundaries did not result in substantial enhancements in quality.

8 Conclusion and Future Work

In this paper, we investigated the efficacy of discourse-enhanced Transformers in the context of the propaganda detection task. Specifically, we examined two different settings, namely paragraph and token classification, using the English and Russian subsets of the SemEval-2023 dataset.

We suggested a modification of the base Transformer architecture to incorporate discourse features. Our experimental results indicated that discourse information substantially enhances the performance of the base models. We conducted a comprehensive analysis to determine the relative importance of each type of discourse feature. Furthermore, our findings revealed a strong correlation between propaganda instances and discourse spans. We believe that this research contributes to the advancement of propaganda detection algorithms and provides valuable insights into the role of discourse in propagandistic texts.

Future work can focus on investigating additional types of discourse features, neural architecture modifications, as well as exploring the generalizability of the suggested approach.

Limitations

The proposed approach is not limited to the specific languages or specific Transformer approaches. However, there are certain limitations that need to be considered. One limitation is the requirement for annotated data, which can be obtained either through manual annotation or with the assistance of a RST discourse parser. Another limitation is the reliance on the encoder architecture of the Transformer-based approach.

Ethics and Broader Impact

The training of large Transformer-based models has been identified as one of the reasons leading to global warming. Nevertheless, it is worth noting that in our research these models were not trained from scratch but instead underwent a fine-tuning process. Additionally, our focus is primarily on utilizing the base variants of these models, which possess a lower number of trainable parameters.

Acknowledgements

The article was prepared within the framework of the HSE University Basic Research Program. It was also supported in part through the computational resources of HPC facilities at NRU HSE.

The publication was also supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

References

Katarzyna Baraniak and Marcin Sydow. 2023. Kb at semeval-2023 task 3: On multitask hierarchical BERT base neural network for multi-label persuasion techniques detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 1395–1400. Association for Computational Linguistics.

Alexander Chernyavskiy and Dmitry Ilvovsky. 2020a. DSNDM: Deep Siamese neural discourse model with attention for text pairs categorization and ranking. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 76–85, Online. Association for Computational Linguistics.

Alexander Chernyavskiy and Dmitry Ilvovsky. 2020b. Recursive neural text classification using discourse tree structure for argumentation mining and sentiment analysis tasks. In *Foundations of Intelligent Systems - 25th International Symposium, ISMIS 2020, Graz, Austria, September 23-25, 2020, Proceedings*, volume 12117 of *Lecture Notes in Computer Science*, pages 90–101. Springer.

Elena Chistova, Artem Shelmanov, Dina Pisarevskaya, Maria Kobozeva, Vadim Isakov, Alexander Panchenko, Svetlana Toldova, and Ivan Smirnov. 2020. RST discourse parser for Russian: An experimental study of deep learning models. In *Proceedings of Analysis of Images, Social Networks and Texts (AIST)*, pages 105–119.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Timo Hromadka, Timotej Smolen, Tomas Remis, Branislav Pecher, and Ivan Srba. 2023. Kinitveraai at semeval-2023 task 3: Simple yet powerful multilingual fine-tuning for persuasion techniques detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 629–637. Association for Computational Linguistics.

Shafiq R. Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Comput. Linguistics*, 43(4).

Dawid Jurkiewicz, Lukasz Borchmann, Izabela Kosmala, and Filip Gralinski. 2020. Applicaai at semeval-2020 task 11: On roberta-crf, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1415–1424. International Committee for Computational Linguistics.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. *CoRR*, abs/1903.07389.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Qisheng Liao, Meiting Lai, and Preslav Nakov. 2023. Marseclipse at semeval-2023 task 3: Multi-lingual and multi-label framing detection with contrastive learning. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 83–87. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2023. HQP: A human-annotated dataset for detecting online propaganda. *CoRR*, abs/2304.14931.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1377–1414. International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. *CoRR*, abs/1910.02517.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 2343–2361. Association for Computational Linguistics.

Antonio Purificato and Roberto Navigli. 2023. Apatt at semeval-2023 task 3: The sapienza NLP system for ensemble-based multilingual propaganda detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 382–388. Association for Computational Linguistics.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 472–489. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 184–188. Association for Computational Linguistics.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João Augusto Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Sheffieldveraai at semeval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 1995–2008. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. In *Annual Meeting of the Association for Computational Linguistics*.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James R. Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, pages 1597–1605. INCOMA Ltd.

A Discourse Coverage

Figure 4 demonstrates the coverage of EDUs associated with corresponding discourse relations with respect to propaganda spans (similarly to Figure 3). The table contains a considerable number of zero values after rounding due to the infrequency of propaganda spans. The distribution contains a bias towards the most popular classes, since the total length of their spans is higher. However, the distribution is not entirely straightforward and there are notable correlations that have been discussed in Section 7. Even for rare classes there are some interesting findings, e.g., if an EDU span contains a slogan then it is most likely that is an Evaluation span. It is important to highlight that "Repetition" is conceptually linked to words, which explains its limited presence in the text despite its frequent occurrence. Consequently, in the context of span classification, a decline in quality is observed.

Figure 4: Discourse relations coverage by the propaganda spans (character-based proportions).

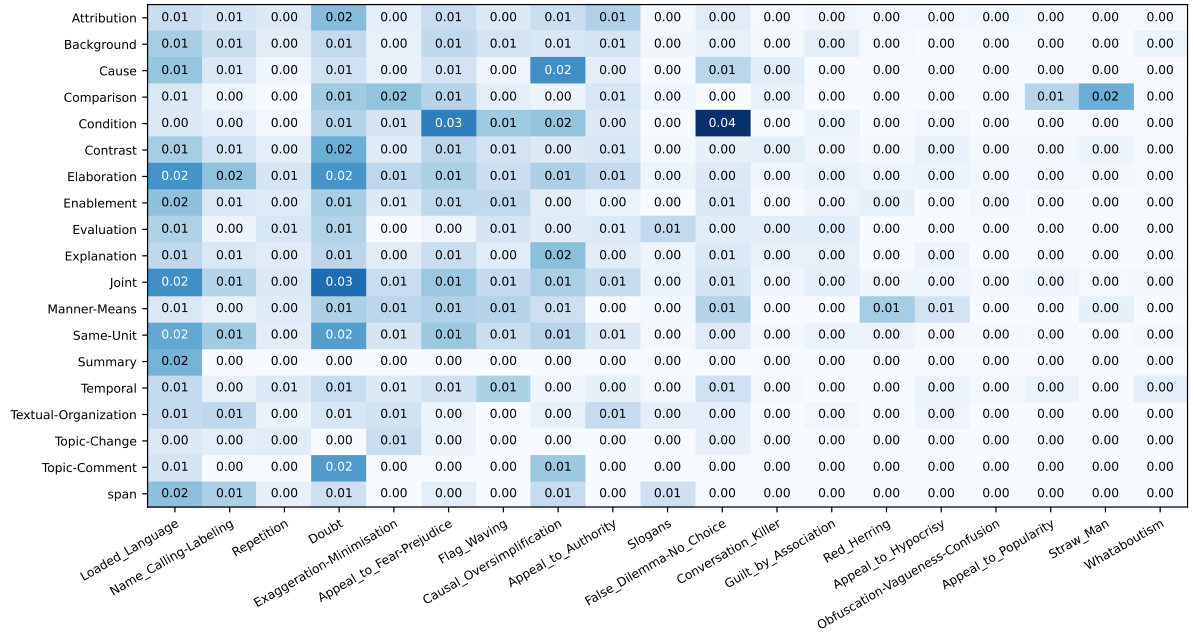

Figure 5: IOU scores calculated based on the overlap between propaganda and discourse spans (character-based proportions).

Figure 5 shows the IoU scores calculated based on the overlap between propaganda and discourse spans and EDUs associated with specific discourse relations (character-based proportions). Generally, the correlations observed in this case align with the aforementioned findings. Here, the alignment of the boundaries for propaganda spans and EDUs also impacts the scores.

Therefore, the intersection between Topic-Comment and "Doubt" spans is lower, whereas between Condition and "False Dilemma No Choice" spans it remains relatively high.

Overall, the identified correlations facilitate the interpretation and analysis. At the same time, the proposed model incorporates features that are not limited to EDUs.