Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

# Discourse-aware text classification for argument mining

**Elena Chistova**
FRC CSC RAS / Moscow, Russia
chistova@isa.ru

**Ivan Smirnov**
FRC CSC RAS / Moscow, Russia
ivs@isa.ru

**Abstract**

We show that using the rhetorical structure automatically generated by the discourse parser is beneficial for paragraph-level argument mining in Russian. First, we improve the structure awareness of the current RST discourse parser for Russian by employing the recent top-down approach for unlabeled tree construction on a paragraph level. Then we demonstrate the utility of this parser in two classification argument mining subtasks of the RuARG-2022 shared task. Our approach leverages a structured LSTM module to compute a text representation that reflects the composition of discourse units in the rhetorical structure. We show that: (i) the inclusion of discourse analysis improves paragraph-level text classification; (ii) a novel TreeLSTM-based approach performs well for the computation of the complex text hidden representation using both a language model and an end-to-end RST parser; (iii) structures predicted by the proposed RST parser reflect the argumentative structures in texts in Russian.

**Keywords:** Discourse parsing, RST, text classification, argumentation mining
**DOI:** 10.28995/2075-7182-2022-21-93-105

# Классификация текстов с учетом дискурсивной структуры для анализа аргументации

**Елена Чистова**
ФИЦ ИУ РАН
Москва, Россия
chistova@isa.ru

**Иван Смирнов**
ФИЦ ИУ РАН
Москва, Россия
ivs@isa.ru

**Аннотация**

В работе демонстрируется эффективность автоматического дискурсивного анализа для анализа аргументации в текстах на русском языке. Улучшенный за счет применения современного метода построения неразмеченных риторических структур метод дискурсивного анализа применяется в классификации документов на примере двух подзадач анализа аргументации в соревновании RuARG-2022. Предлагаемый подход к классификации на основе структурной LSTM предусматривает обучение векторного представления текста, отражающего композицию его фрагментов в дискурсивном дереве. В ходе исследования показано, что: (1) учет предсказанной дискурсивной структуры позволяет улучшить качество классификации текста на уровне абзаца; (2) предложенный подход на основе TreeLSTM эффективен при обучении векторного представления абзаца с использованием языковой модели и автоматического дискурсивного анализатора; (3) предсказанные анализатором риторические структуры в целом отражают аргументативную структуру текстов.

Ключевые слова: Дискурсивный анализ, теория риторических структур, классификация текстов, анализ аргументации

## 1 Introduction

As an attention module of an advanced classification model traverses a complex sentence or a document sequentially, it may become confused as to which phrases pertain to the document's class and which represent a different class from the document's meaning, or whether certain phrases argue for or against the author's position. It is possible to uncover relations between text parts with discourse parsing. The

Rhetorical Structure Theory (RST [22]) is the discourse framework suggesting that texts have a hierarchical, connected structure, with both intra- and inter-sentential relations. A rhetorical tree shows how elementary discourse units (EDUs) and non-elementary units combine to form the overall meaning of a document (see Figure 1).

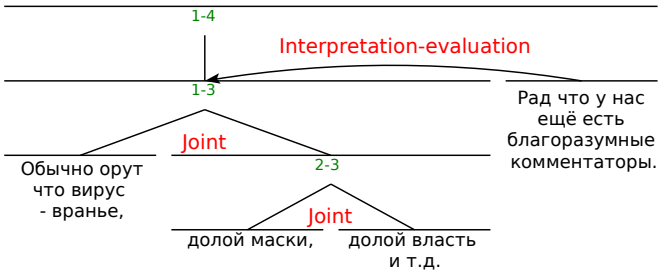



Figure 1: RST parsing result for a short example from RuArg-2022: *[Usually, they yell that the virus is a scam,]*$_1$ *[down with the masks,]*$_2$ *[down with the government, etc.]*$_3$ *[Glad there are still sane commenters out there.]*$_4$. In this example, there is only one mononuclear rhetorical relation, where the left constituent (EDU$_{1-3}$) is the nucleus, and the right constituent (EDU$_4$) is the satellite.

Text classification methods adopt one of two main approaches based on the discourse parsing: either (1) weighting tokens based on their position in an unlabeled discourse tree for lexicon-based analysis [4, 13, 32], or (2) combining phrases based on the discourse structure to determine an overall class score [16, 20]. These methods focus mainly on sentiment analysis; several studies have also identified a connection between the rhetorical and argumentation structures [8, 5, 10].

This paper investigates the impact of discourse parsing on document classification for argument mining in social media texts in Russian. We first improve the performance of the current RST parser for Russian by introducing it to the top-down paragraph parsing. Then, we investigate the text classification applying a Tree LSTM [31] module on the predicted discourse structures. We use this module to correct the predictions of the fine-tuned language model. The classification methods are tested on the RuARG-2022 [17] shared task.

Our contributions can be highlighted as follows:

- To the best of our knowledge, we are the first to explicitly analyze the effect of discourse among opinion mining and argument classification in Russian.
- We achieve a significant improvement in RST discourse parsing for Russian using a top-down algorithm for unlabeled tree construction at the paragraph level.
- We propose a new method to utilize RST discourse structure in Tree LSTM for paragraph-level text representation learning.

Our code is publicly available[1].

## 2 Related work

**Rhetorical structure in text classification:** A number of early studies investigated the possibility of opinion mining using shallow text structures derived from the discourse connectors vocabulary [24, 29] or based on the manual discourse annotations [2]. The development of automatic discourse parsers for English has strengthened research in this area. Finding the most common nucleus of rhetorical relations in each sentence, authors of [32] investigate whether the sentiment lexicon can be weighted based on RST structure. Assuming that the nuclei of each relation encapsulate the general idea of the text, they explain the lack of performance improvement through the poor accuracy of the early RST parser SPADE [30]. However, the results obtained with the SPADE and HILDA [28] parsers in [13] demonstrate an improvement in sentiment classification when weighting words based on the depth of corresponding subtree and nuclearities in it. Markov logic and the sentence-level discourse trees predicted by the

[1] https://github.com/tchewik/discourse-aware-classification

HILDA parser are used in [34] to calculate the sentiment score using information about contrastive (Contrast, Concession) and non-contrastive (the rest of RST-DT relations) rhetorical relations occurring between elementary discourse units.

More recent approaches explore the integration of the explicit discourse structure in deep learning models. In [4], sentiment scores are propagated recursively up the RST tree to the root via a neural network with architecture specific for each parse, and scalar parameters related to particular relations are tuned. They do not construct latent representations of discourse units and also train a simplified version of the DPLP RST parser [15], focusing only on distinguishing contrastive and non-contrastive relations. Method [16] exploits the trainable representations of discourse units at all levels. The authors propose to build a shared text vector representation for a discourse tree node based on the composition of representations of individual EDUs and subtrees. A weighting of the importance of individual discourse units is automated through the attention mechanism. They test the method on multiple text classification tasks. In [9], the authors apply Tree LSTM to the unlabeled sentence-level RST trees with nuclearities. Binary Tree LSTM in their method does not process left and right children of the current node, but rather its predefined nucleus and satellite (or two nuclei). The use of the DPLP parser to construct structural neural networks is also demonstrated in [20]. They construct RecNN [12] and Tree LSTM [31]. To reduce the complexity of the neural network to be constructed, they consider individual sentences rather than EDUs as leaves of the discourse tree. In the representation of each discourse tree node, the text embedding and the rhetorical relation embedding are concatenated; the sentence embeddings are trained independently. In [19] the authors propose a Tree LSTM model similar to the one proposed in [9] with and additional tree nodes augmentation. In their study, they predict the polarity of each EDU using dictionaries and word embeddings and found that incorporating embeddings leads to strong overfitting in the Tree LSTM models.

**Argument mining using RST annotation:** Argument mining is known to benefit from discourse analysis. It has been shown [8] that certain semantic groups of discourse connectors are indicative of either claims or premises and can be used to differentiate between the two. There are certain argumentative relations in RST that represent supportive, incentive, justification, and persuasion arguments, as outlined in [3]. Communicative discourse structure inspired by RST is used in [10] to categorize texts as being either argumentative or non-argumentative. The authors of [5] propose combining a BERT-based classifier with a gradient boosting model based on a rhetorical relation label in the root of the discourse tree. Examples of how the classifier on discourse relations corrects the predictions of BERT are given in order to illustrate how some RST relations, such as Evaluation or Antithesis, correlate with argumentative ones. TreeLSTM over RST structure is probed for argumentation mining in [6]. This module is used to obtain a vector representation of the text (the root of the rhetorical tree) based on EDU embeddings, which are formed by concatenating word, sentence, and part-of-speech tag embeddings.

In this work, we propose a TreeLSTM-based text classification method for argument mining. Current text classification methods using TreeLSTM over RST structures, usually designed for sentiment analysis, are subject to strong overfitting due to the high dimensionality of discourse unit embeddings trained jointly with the recursive neural module. The key difference between our work and previous work is that we do not train TreeLSTM from scratch in conjunction with the text encoder, but instead use the module to refine predictions of a high-performance sequential text classifier on documents with rhetorical structure.

## 3 Improving discourse parsing for Russian

This section describes the end-to-end RST parsing method we later use in text classification. We propose constructing unlabeled trees at the paragraph level by using a top-down approach, which improves the structure awareness of the recent discourse parser for Russian.

**Method:** RST parser for Russian recently proposed in [7] is proven to be highly accurate for relation classification and EDU segmentation, although its greedy bottom-up tree-building algorithm limits its overall performance for document parsing. However, the method takes on the challenge of segmenting

long texts into separate discourse trees despite weakly paragraph related tree boundaries, a feature of the Russian RST corpus RuRSTreebank [27] that disallows direct application of state-of-the-art unlabeled tree construction methods (1 document = 1 tree) developed for other languages. Therefore, for our experiments, we reproduce [7], but replace the sentence- and paragraph-level unlabeled tree construction methods in the parser with the recent top-down parsing approach proposed in [26] under the assumption that each paragraph corresponds to a separate subtree. As opposed to prior top-down discourse parsing methods [21, 33] which considered each span separately at each time step, the novel method allows for comparison of subtree candidates globally at the full-tree level by computing all span boundary representations in text at each time step and using beam search to find the best subtree candidate.

**Data, Results, and Discussion:** We use the standard RuRSTreebank corpus [27] for training and evaluation, focusing on two genres: news and blogs, and selecting 15% of data for the test. Since there is no available language model for long documents in Russian, we rely on character and pretrained word2vec embeddings for the initial representation of the document. For training on gold segmentation, we use the following parameters: beam size = 20, batch size = 4000 tokens. In Table 1, we compare the end-to-end discourse analysis performance at the different granularity levels between the system using greedy bottom-up paragraph parsing [7] and the one proposed in this study using micro-averaged standard Parseval metric [23]. In both cases, we use the same BiLSTM-CRF discourse segmentation model on pretrained ELMo embeddings, achieving 88.4% F1 on the test set. We use word2vec and ELMo pretrained models provided by RusVectores[2]. For our structure-aware classification method, the parser's most important feature is its ability to retrieve discourse structure regardless of labeled relations. The top-down approach improves the unlabeled tree construction (span identification) performance by 10.5% F1 at the sentence level, 10.4% F1 at the paragraph level, and 8.9% F1 at the document level, taking into account that the relations between paragraphs are in both cases detected by applying the same greedy bottom-up algorithm. The full end-to-end parsing performance increases by 10.6%, 7.0%, and 6.2% F1, respectively. We publish the source code for the end-to-end parser used in our experiments[3].

| Method | Sentence level | | | | Paragraph level | | | | Document level | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | span | nuc | rel | full | span | nuc | rel | full | span | nuc | rel | full |
| Greedy [7] | 58.0 | 38.9 | 27.8 | 27.1 | 49.4 | 31.0 | 20.4 | 20.3 | 43.6 | 27.3 | 18.0 | 17.7 |
| Beam search | 68.5 | 50.6 | 38.1 | 37.7 | 59.8 | 38.8 | 27.5 | 27.3 | 52.5 | 34.2 | 24.2 | 23.9 |

Table 1: Performance of end-to-end RST parsing using different paragraph-level unlabeled tree construction methods

In this study, improving parsing performance at both the sentence and paragraph levels is crucial. Analysis of the RuArg-2022 dataset reveals that each example corresponds to a single automatically identifiable sentence. However, the sentence segmenter often fails to segment social media comments properly, because some sentences end with emojis, parentheses, ellipses, or no punctuation at all. In addition, social media users often write extremely long sentences that could be broken down into several grammatically correct shorter ones. Therefore, in some situations, it may be necessary to analyze intersentential discourse relations.

## 4 Discourse-aware classification method

In this section, we detail our proposed method for stance and argument classification, addressing the limitations of unstructured full-text classification methods. We discuss the pipeline-based framework for the classification of texts with or without recognizable rhetorical structure. The first stage involves fine-tuning the sequential model on the dataset including texts of different lengths and complexity. In the second stage, we freeze the base model and then train a discourse-aware neural module on top of it for the classification of texts with discourse structure.

[2] https://rusvectores.org/

[3] https://github.com/tchewik/isanlp_rst/releases/tag/v2.0

### 4.1 BERT

For text classification based on token sequences, we adapt the multitask baseline model architecture proposed by the competition organizers, where two outputs are being trained simultaneously in a classifier based on a language model. We use the DeepPavlov RuBERT Conversational[4] along with BERT pooling to encode the document. This particular language model was chosen because it is pretrained on dialogue and social media texts, so it is well suited for encoding social media comments. The hidden representation is then passed through two fully-connected layers for stance and argument prediction; all parameters are trainable.

The model as it stands is used in the final pipeline for predicting labels for structure-lacking sentences (EDUs). It is also used in the structure-aware model for the initial encoding of discourse tree nodes.

### 4.2 RST-LSTM

RST parsers represent discourse as a binary constituency tree. If the binary discourse tree is traversed from the bottom up, information from the left and right constituents can be combined to represent the tree node at the upper level and all the way up to the root. Our structure representation module is based on the Binary Tree LSTM network [31]. In Binary Tree LSTM, a non-elementary discourse unit's hidden and cell states are determined by the hidden and cell states of its left and right constituents rather than the sequence of words inside it. It allows computation over self-contained phrases within a complex discourse. We draw inspiration from previous work on Tree-LSTM over RST structure for document classification, but instead of classifying each node in an unlabeled RST tree based on text features [9, 14], or dictionary-based class scores [4, 19], we use outputs of a pretrained classifier and a type of rhetorical relation as input features of each node to predict the only label for the rhetorical root of the document. A single overall class label is defined for the entire text in the tasks presented in this paper. Hence, we propose a deep model for aggregation of the class labels predicted for all the discourse units in a document by a sequential text classifier. First of all, this allows for a strong sequential text classification method, one that itself takes into account some aspects of discourse [18]. Additionally, the methods based on training high-dimensional EDU representations simultaneously with Tree-LSTMs are found to be prone to strong overfitting [19]. Therefore, it is important to produce DU representations that are as compact and informative as possible, which the proposed method achieves by encoding them with a pre-trained classifier.

The six types of fine-grained relations in the RuRSTreebank corpus outlined in [27] are used in the initial feature representation of each node. These include Coherence (Background, Elaboration, Restatement, Interpretation-evaluation, Preparation), Causal-argumentative:Contrastive (Concession, Contrast, Comparison), Causal-argumentative:Causal (Purpose, Evidence, Cause-effect), Causal-argumentative:Condition (Condition), Structural (Sequence, Joint, Same-unit), and Attribution.

Considering RST tree $t$ and the current nonterminal node (nonelementary discourse unit) $u_i \in t$, its left and right constituents $u_{i_1}$ and $u_{i_2}$ sharing relation $r_i = (r_{i_1}, r_{i_2})$ (e.g., Attribution_NS $=$ (Attribution_Nucleus, Attribution_Satellite)) are initially encoded into representations $U_{i_1}$ and $U_{i_2}$ as follows:

$$U_{i_j} = [\mathrm{FC}_{stance}(\mathrm{Enc}(u_{i_j})); \mathrm{FC}_{premise}(\mathrm{Enc}(u_{i_j})); r_{i_j}] \text{ for } j = 1, 2. \quad (1)$$

An additional Root relation is introduced to encode a root node that is not a constituent. We derive both the BERT-based text encoder $\mathrm{Enc}$ and the fully-connected layers for preliminary labels predictions $\mathrm{FC}$ from the sequence-level base model with frozen weights. Since all Structural relations are multinuclear and do not have satellites, the one-hot vector $r_{i_j}$ of discourse unit labels (Coherence_Nucleus, Coherence_Satellite, Root, etc.) in our model has a length of 12. Binary Tree LSTM is then applied to these representations $U_k \in t$. The model uses a Tree LSTM hidden representation of the root discourse unit for both stance and argument prediction and has two output feedforward layers as with the BERT model.

[4] https://huggingface.co/DeepPavlov/rubert-base-cased-conversational

## 5 Experiments

### 5.1 Data

The dataset for joint stance and premise classification is provided by the RuArg-2022 competition organizers. The stance label represents the point of view of the author in relation to the given claim. The presence of arguments for, against, or mixed in the text is indicated by the premise (argument) label. There are three claims in the dataset regarding the COVID-19: "Wearing masks is beneficial for society", "Vaccination is beneficial for society", and "The introduction and observance of quarantine is beneficial for society".

Figure 2 illustrates the length distribution of data by elementary discourse units derived with RST parsing. Since the texts in the dataset do not have paragraph breaks, each text is considered to belong to a single tree. Thus, if the text is $l$ elementary discourse units long, its rhetorical structure contains $l - 1$ relations. Each subset of the data contains about 25% simple sentences with no automatically recognizable discourse structure. From this, we hypothesize that for 75% of the data, the classification performance can be improved by analyzing the coherence structure within the text. Most examples are found to have only one discourse relation between two elementary units; in the official test set, this is the case in 35.9% of examples.

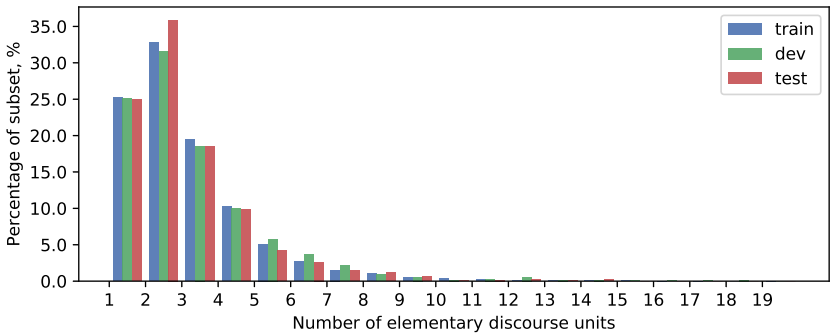
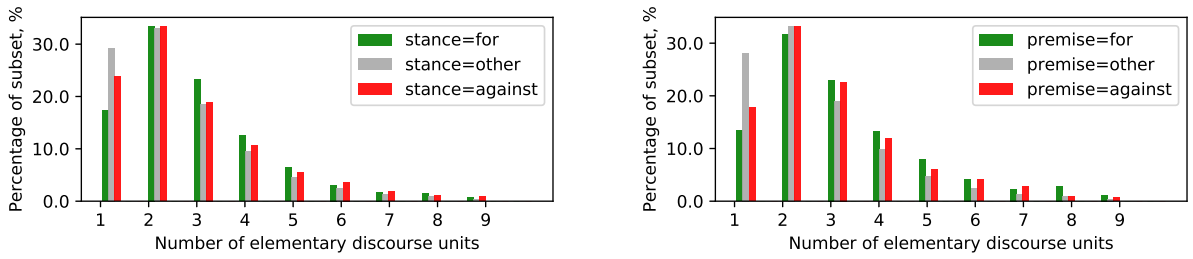



Figure 2: Distribution of text lengths in RuArg-2022

Figure 3 shows the distribution of text lengths in different topic-irrelevant classes across the labeled train set. Complex texts with a rhetorical structure are the most common way in which polar opinions are expressed in the corpus. The simplest sentences are most common among the examples of the mixed class *Other*, with this difference being particularly evident in the premise (argument) classification subtask (Fig. 3b). It demonstrates that most examples in this class lack any argumentation typically [5, 8, 25] expressed by causal, conditional, or any other meaningful discourse relations, which is consistent with the definition of the *Other* in the premise classification subtask description. One interesting observation is that the examples in which the author expresses a positive stance or argument tend to have the most complex structures in the train set.




(a) Stance detection

(b) Premise classification

Figure 3: Distribution of text lengths in the train set

### 5.2 Settings

In the data, the number of examples of various classes is unbalanced, with a major predominance of the *Irrelevant*. We choose to add class weights to the loss in both BERT and RST-LSTM models to prevent unbalanced learning. These weights are adjusted to correspond to the overall class weights in the train data. We use the Optuna optimization framework [1] for automated hyperparameter tuning in both BERT fine-tuning and RST-LSTM training. The optimal hidden size of Tree-LSTM for the three topic-related models is found to be between 50 and 125 units. We use PyTorch and AllenNLP libraries [11] for implementation and a single Nvidia GeForce RTX 2080 Ti GPU. In our experimental setup, RST-LSTM takes on average 2.4 times less time to run one training epoch than BERT, with 2 to 5 epochs total.

### 5.3 Evaluation Procedure

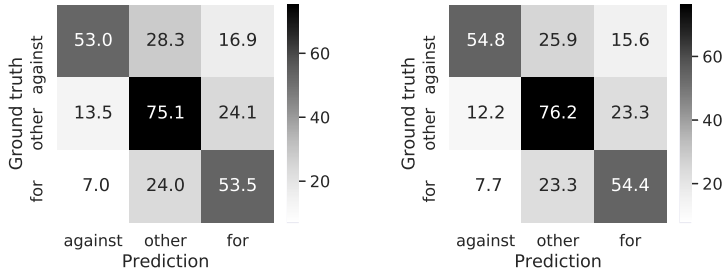
In our evaluation, we use the metric proposed by RuArg-2022: macro F1 excluding the score for label *Irrelevant*. We use a 5-fold cross-validation over the labeled train set to accurately compare approaches that employ and do not employ rhetorical structure. For the official test and development sets, the final predictions are obtained by averaging predictions from five models trained on cross-validation. This is similar to an ensemble, where each model is trained using 80% of the train data.

## 6 Results and Discussion

In Table 2 we compare the results of the model with Tree LSTM over RST structure with the baseline BERT model.

| Non-EDU classification | Performance on non-EDU Masks | Vaccines | Quar. | Mean | Overall performance Masks | Vaccines | Quar. | Mean |
|---|---|---|---|---|---|---|---|---|
| | | | | Stance detection | | | | |
| BERT | 59.8 ± 2.7 | 62.4 ± 3.4 | 54.5 ± 3.4 | 58.9 ± 2.3 | 60.6 ± 2.6 | 64.4 ± 2.2 | 56.4 ± 2.8 | 60.5 ± 1.9 |
| + RST-LSTM | **61.3 ± 2.7** | **63.4 ± 4.2** | **55.6 ± 2.7** | **60.1 ± 2.3** | **61.7 ± 2.6** | **65.1 ± 3.0** | **57.5 ± 2.4** | **61.4 ± 1.8** |
| | | | | Premise classification | | | | |
| BERT | 66.4 ± 2.9 | **61.7 ± 4.3** | 56.4 ± 2.8 | 61.5 ± 2.2 | 66.0 ± 2.4 | **62.6 ± 2.7** | 57.0 ± 2.3 | 61.9 ± 1.6 |
| + RST-LSTM | **68.1 ± 2.1** | 60.4 ± 3.3 | **57.6 ± 2.0** | **62.0 ± 1.3** | **67.5 ± 1.9** | 61.5 ± 2.3 | **58.3 ± 2.1** | **62.4 ± 0.9** |

Table 2: Performance (F1, mean ± std) during cross-validation




(a) BERT

(b) BERT + RST-LSTM

Figure 4: Averaged cross validation confusion matrices for stance detection

**Stance detection:** Introducing discourse structure to this model leads to an improvement across all topics. The method resulted in an averaged 1.2% mean F1 improvement in the classification of texts with discourse structure and a 0.9% mean F1 improvement for all texts. Figure 4 shows the averaged cross-validation confusion matrices for stance detection in the non-elementary samples of the train set, topics not separated, excluding label *Irrelevant*. The model with RST-LSTM shows improved ability to distinguish between *For* and *Against* polar stance labels and the mixed label *Other*. For the sequential

classifier, *Other* is the most challenging label. First, it can mean that the text has examples of both polar classes; second, it is the most frequent. We include the examples where discourse structure helps to differentiate the stance labels in Figures 5 and 6 in Appendix.

**Premise classification:** On average, the classification performance improved by 0.5% both for texts with identifiable rhetorical structure and for all texts. Classification performance improved significantly for the topics Masks (+1.6% F1) and Quarantine (+2.0% F1) and worsened for the Vaccines (-1.3% F1), which may indicate a drawback of the approach in which a single structure representation for two targets is trained simultaneously; it is also worth noting that the scores related to the Vaccines topic have the highest deviation in both the sequential and structural classification methods on complex (non-EDU) examples. The examples where discourse structure helps to differentiate the premise labels are illustrated in Figures 7 and 8 in Appendix.

**Evaluation on the official dev and test sets:** In Table 3 we compare the methods on the official dev and test sets of RuArg-2022. Both sets are treated as unseen, so the official development set was not used for the parameters adjustment. The results confirm that the RST-LSTM is capable of capturing the overall polarity of stance and arguments in a document based on the rhetorical structure. Vaccines-related text classification continues to produce the least stable results. On both dev and test sets, the BERT model enhanced with the RST-LSTM module achieves the best performance for premise classification.

| Non-EDU classification | Dev Masks | Vaccines | Quar. | Mean | Test Masks | Vaccines | Quar. | Mean |
|---|---|---|---|---|---|---|---|---|
| BERT | 66.0 / 66.1 | 66.8 / **58.5** | **58.4** / 58.8 | 63.7 / 61.1 | **70.0** / 76.4 | **68.3** / 63.4 | 61.0 / 71.6 | **66.5** / 70.6 |
| + RST-LSTM | **67.3 / 68.2** | **67.8** / 56.3 | 56.3 / **59.8** | **63.8 / 61.4** | **70.0 / 76.5** | 66.0 / **63.8** | **61.7 / 72.5** | 65.9 / **71.0** |

Table 3: Performance (F1, stance / argument) on dev and test sets of RuArg-2022

The public RuARG-2022 test leaderboard (Table 4) shows only the results of the last model evaluated. In our case, it is a model different from the one evaluated in Tables 2 and 3. Specifically, it is an additional variant of RST-LSTM where nuclei of asymmetric relations are marked as Span (instead of Attribution_Nucleus, Coherence_Nucleus, etc.). As our method aggregates the rhetorical relations with similar semantics and nuclearity discrepancies, such as the causal relations Purpose[5] and Cause-Effect[6], or contrastive Concession[7] and Contrast[8], this idea was later dismissed. However, according to Table 3, our final method described in this paper also ranks 4th in stance prediction (65.9%) and 3rd in argument classification (71.0%) in the competition leaderboard on the official test set.

| Stance detection | | | Premise classification | | |
|---|---|---|---|---|---|
| 1 | camalibi | 69.7 | 1 | camalibi | 74.0 |
| 2 | sevastyanm | 68.2 | 2 | sevastyanm | 72.4 |
| 3 | iamdenay | 66.8 | 3 | ursdth (ours) | 70.6 |
| 4 | ursdth (ours) | 65.7 | 4 | iamdenay | 65.6 |
| 5 | sopilnyak | 56.0 | 5 | dr | 60.4 |
| 6 | kazzand | 55.5 | 6 | kazzand | 56.0 |
| 7 | morty | 53.5 | 7 | morty | 54.5 |
| 8 | invincible | 58.9 | 8 | invincible | 54.3 |
| 9 | dr | 47.5 | 9 | Baseline | 43.6 |
| 10 | Baseline | 41.8 | | | |

Table 4: Public leaderboard of RuArg-2022 (F1)

**Ablation study:** We inspect the importance of the rhetorical relation labels and nuclearities in our method in Table 5. We found that excluding particular relation types individually, i.e. replacing the

[5]Satellite represents the intended *result* behind the situation described in the nucleus.

[6]Nucleus represents the actual *result* after the situation described in the satellite.

[7]Mononuclear relation in which additional information in satellite creates expectations that the situation in the nucleus would be opposite.

[8]Multinuclear relation in which nuclei describe alternative situations.

only type in the trees with a structural type, marginally affects the classification performance on the discourse trees. However, the simultaneous substitution of all semantic relations for a multinuclear relation Structural leads to a 0.2% F1 decrease in the stance identification performance and a 0.6% F1 decrease in argument classification, demonstrating the importance of the features related to the labeled rhetorical structure in argument mining. We note that although the RST parser is far from perfect in recognizing the labeled trees, it is capable of identifying argumentative structures.

| Discourse relations | Stance detection | Argument classification |
|---|---|---|
| All | 60.1 | 62.0 |
| - Coherence | - 0.1 | - 0.1 |
| - Contrastive | - 0.1 | - 0.1 |
| - Causal | - 0.0 | - 0.1 |
| - Condition | - 0.1 | - 0.1 |
| - Attribution | - 0.1 | - 0.0 |
| Only structural | - 0.2 | - 0.6 |

Table 5: Ablation study on the rhetorical relations types during cross-validation (F1, mean)

## 7 Conclusion

Sequential text classifiers typically perform well when applied to short texts, but their performance degrades for longer texts due to the complexity of discourse. In order to form an accurate hidden representation of a complex text, we propose a method leveraging both a pretrained language model and an end-to-end RST parser. Additionally, we improve the rhetorical parsing for Russian using a recent top-down algorithm for paragraph parsing and report fine-grained RST scores for different text granularities. The improved RST parser is used to show the utility of rhetorical parsing in stance detection and premise classification on social media comments.

The architecture we propose shows the effectiveness of a two-step rhetorical-driven approach, where the base text classification method can be any advanced neural network or feature-based machine learning model. Future work should investigate the suitability of discourse parsing in Russian for other tasks requiring argument extraction and processing.

## Acknowledgements

This paper is supported by the Research Program of the National Center for Physics and Mathematics (project no. 9).

## References

[1] Akiba Takuya et al. Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. — 2019. — P. 2623–2631.

[2] Asher Nicholas et al. Appraisal of opinion expressions in discourse // Lingvisticæ Investigationes. — 2009. — Vol. 32, no. 2. — P. 279–292.

[3] Azar Moshe. Argumentative text as rhetorical structure: An application of rhetorical structure theory // Argumentation. — 1999. — Vol. 13, no. 1. — P. 97–114.

[4] Bhatia Parminder et al. Better Document-level Sentiment Analysis from RST Discourse Parsing // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 2212–2218.

[5] Chakrabarty Tuhin et al. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — P. 2933–2943.

[6] Chernyavskiy Alexander and Ilvovsky Dmitry. Recursive Neural Text Classification Using Discourse Tree Structure for Argumentation Mining and Sentiment Analysis Tasks // International Symposium on Methodologies for Intelligent Systems / Springer. — 2020. — P. 90–101.

[7] Chistova Elena et al. RST Discourse Parser for Russian: An Experimental Study of Deep Learning Models // In Proceedings of Analysis of Images, Social Networks and Texts (AIST). — 2020. — P. 105–119.

[8] Eckle-Kohler Judith et al. On the role of discourse markers for discriminating claims and premises in argumentative discourse // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. — 2015. — P. 2236–2242.

[9] Fu Xianghua et al. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis // Asian conference on machine learning / PMLR. — 2016. — P. 17–32.

[10] Galitsky Boris et al. Argumentation in text: discourse structure matters // CICLing 2018. — 2018.

[11] Gardner Matt et al. AllenNLP: A Deep Semantic Natural Language Processing Platform. — 2017. — arXiv:1803.07640.

[12] Goller Christoph and Kuchler Andreas. Learning task-dependent distributed representations by backpropagation through structure // Proceedings of International Conference on Neural Networks (ICNN'96) / IEEE. — 1996. — Vol. 1. — P. 347–352.

[13] Hogenboom Alexander et al. Using rhetorical structure in sentiment analysis // Communications of the ACM. — 2015. — Vol. 58, no. 7. — P. 69–77.

[14] Huber Patrick and Carenini Giuseppe. From Sentiment Annotations to Sentiment Prediction through Discourse Augmentation // Proceedings of the 28th International Conference on Computational Linguistics. — 2020. — P. 185–197.

[15] Ji Yangfeng and Eisenstein Jacob. Representation learning for text-level discourse parsing // Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). — 2014. — P. 13–24.

[16] Ji Yangfeng and Smith Noah A. Neural Discourse Structure for Text Categorization // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2017. — P. 996–1005.

[17] Kotelnikov Evgeny et al. RuArg-2022: Argument Mining Evaluation // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". — 2022.

[18] Koto Fajri et al. Discourse Probing of Pretrained Language Models // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — P. 3849–3864.

[19] Kraus Mathias and Feuerriegel Stefan. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees // Expert Systems with Applications. — 2019. — Vol. 118. — P. 65–79.

[20] Lee Kangwook et al. A discourse-aware neural network-based text model for document-level text classification // Journal of Information Science. — 2018. — Vol. 44, no. 6. — P. 715–735.

[21] Lin Xiang et al. A Unified Linear-Time Framework for Sentence-Level Discourse Parsing // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — 2019. — P. 4190–4200.

[22] Mann William and Thompson Sandra. Rhetorical structure theory: A theory of text organization. — University of Southern California, Information Sciences Institute Los Angeles, 1987.

[23] Morey Mathieu et al. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT // Conference on Empirical Methods on Natural Language Processing (EMNLP 2017). — 2017. — P. pp–1330.

[24] Mukherjee Subhabrata and Bhattacharyya Pushpak. Sentiment analysis in twitter with lightweight discourse analysis // Proceedings of COLING 2012. — 2012. — P. 1847–1864.

[25] Musi Elena et al. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

[26] Nguyen Thanh-Tung et al. RST Parsing from Scratch // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2021. — P. 1613–1625.

[27] Pisarevskaya Dina et al. Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference Dialogue. — 2017. — P. 23.

[28] Prendinger Helmut et al. A novel discourse parser based on support vector machine classification // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 665–673.

[29] Somasundaran Swapna and Wiebe Janyce. Recognizing stances in online debates // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. — 2009. — P. 226–234.

[30] Soricut Radu and Marcu Daniel. Sentence level discourse parsing using syntactic and lexical information // Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. — 2003. — P. 228–235.

[31] Tai Kai Sheng et al. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — 2015. — P. 1556–1566.

[32] Voll Kimberly and Taboada Maite. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance // Australasian Joint Conference on Artificial Intelligence / Springer. — 2007. — P. 337–346.

[33] Zhang Longyin et al. A Top-down Neural Architecture towards Text-level Parsing of Discourse Rhetorical Structure // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 6386–6395.

[34] Zirn Cäcilia et al. Fine-grained sentiment analysis with structural features // Proceedings of 5th International Joint Conference on Natural Language Processing. — 2011. — P. 336–344.

## Appendix

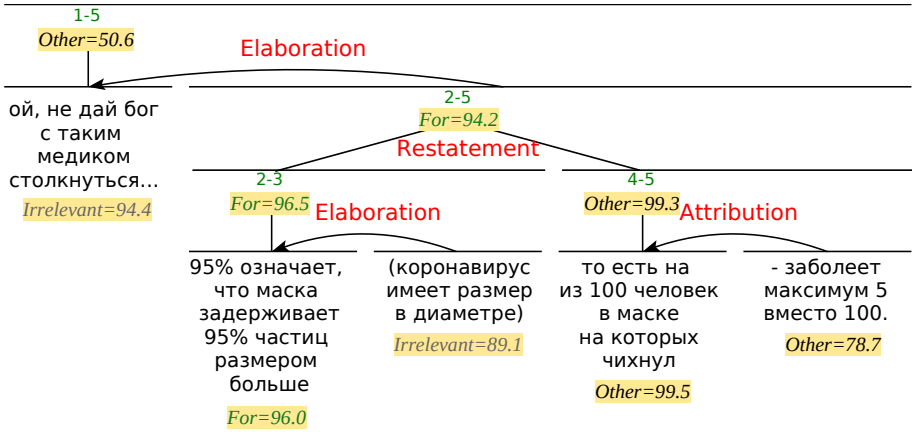



Figure 5: Example of RST parsing for a claim regarding masks from RuArg-2022: *[oh, god forbid anyone encountering such a physician...]*$_1$ *[95% means that the mask blocks 95% of particles larger]*$_2$ *[(coronavirus has a diameter dimension)]*$_3$ *[that is, out of 100 people wearing masks who were sneezed on]*$_4$ *[- at most 5 will get sick instead of 100.]*$_5$. For this example, the BERT-based classifier predicts a false stance label *Other*, but RST-LSTM predicts a true stance label *For*.

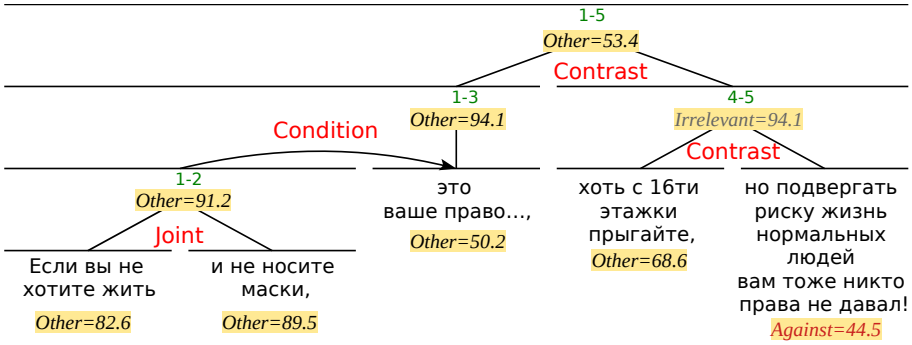



Figure 6: Example of RST parsing for a claim regarding masks from RuArg-2022: *[If you don't want to stay alive]*$_1$ *[and are not wearing masks,]*$_2$ *[that's your right...,]*$_3$ *[you are free to jump off a 16-story building as well,]*$_4$ *[but no one authorized you to endanger normal people's lives!]*$_5$. Yellow boxes indicate BERT stance label predictions and their probabilities. For this example, the BERT-based classifier predicts a false stance label *Other*, but RST-LSTM predicts a true stance label *For*.

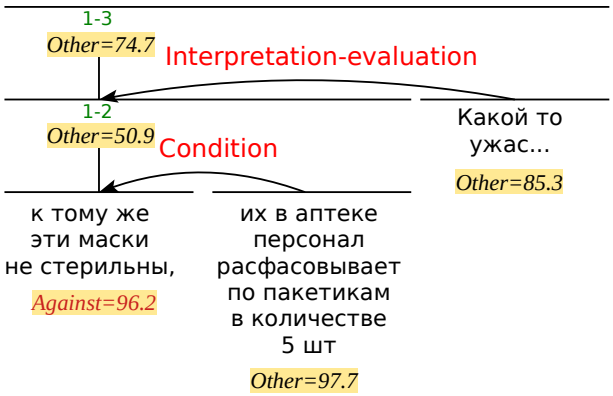



Figure 7: Example of RST parsing for a claim regarding masks from RuArg-2022: *[in addition, these masks are not sterile,]$_1$ [the pharmacy staff packs them in bags of five pieces]$_2$ [What a nightmare...]$_3$*. Yellow boxes indicate BERT premise label predictions and their probabilities. For this example, the BERT-based classifier predicts a false premise label *Other*, but RST-LSTM predicts a true premise label *Against*.

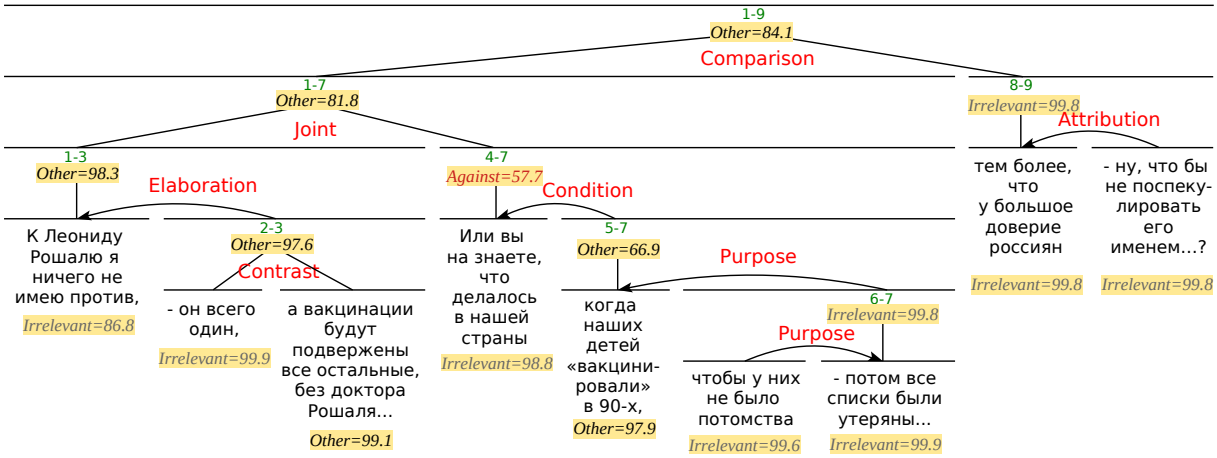



Figure 8: Example of RST parsing for a claim regarding vaccines from RuArg-2022: *[I have nothing against Leonid Roshal]$_1$ [- he is just one person,]$_2$ [and everyone else will be vaccinated, without Dr. Roshal...]$_3$ [Or don't you remember what happened in our country]$_4$ [when they "vaccinated" our children back in the 90s]$_5$ [so that they wouldn't have children]$_6$ [- then all the records went missing...]$_7$ [especially given how much trust Russians have for him]$_8$ [well, why not speculate on his name...?]$_9$* Yellow boxes indicate BERT premise label predictions and their probabilities. For this example, the BERT-based classifier predicts a false premise label *Other*, but RST-LSTM predicts a true premise label *Against*.