RuArg-2022: Argument Mining Evaluation

Evgeny Kotelnikov[1], Natalia Loukachevitch[2], Irina Nikishina[3], and Alexander Panchenko[3],

[1]Vyatka State University, [2]Lomonosov Moscow State University,
[3]Skolkovo Institute of Science and Technology
kotelnikov.ev@gmail.com, louk_nat@mail.ru,
{irina.nikishina, a.panchenko}@skoltech.ru

Abstract

Argumentation analysis is a field of computational linguistics that studies methods for extracting arguments from texts and the relationships between them, as well as building argumentation structure of texts. This paper is a report of the organizers on the first competition of argumentation analysis systems dealing with Russian language texts within the framework of the Dialogue conference. During the competition, the participants were offered two tasks: stance detection and argument classification. A corpus containing 9,550 sentences (comments on social media posts) on three topics related to the COVID-19 pandemic (vaccination, quarantine, and wearing masks) was prepared, annotated, and used for training and testing. The system that won the first place in both tasks used the NLI (Natural Language Inference) variant of the BERT architecture, automatic translation into English to apply a specialized BERT model, retrained on Twitter posts discussing COVID-19, as well as additional masking of target entities. This system showed the following results: for the stance detection task an F1-score of 0.6968, for the argument classification task an F1-score of 0.7404. We hope that the prepared dataset and baselines will help to foster further research on argument mining for the Russian language.

RuArg-2022: соревнование по анализу аргументации

Евгений Котельников[1], Наталья Лукашевич[2], Ирина Никишина[3], и Александр Панченко[1]

[1]Вятский государственный университет,
[2]Московский государственный университет им. М. В. Ломоносова,
[3]Сколковский институт науки и технологий
kotelnikov.ev@gmail.com, louk_nat@mail.ru,
{irina.nikishina, a.panchenko}@skoltech.ru

Аннотация

Анализ аргументации – это область компьютерной лингвистики, в которой исследуются методы извлечения из текстов аргументов и связей между ними, а также построения аргументационной структуры. Настоящая статья представляет собой отчет организаторов о первом соревновании русскоязычных систем анализа аргументации в рамках конференции «Диалог». В ходе соревнования участникам были предложены две задачи: определение позиции автора по заданной теме и классификация доводов. Для обучения и тестирования систем был подготовлен и размечен корпус, содержащий 9,550 предложений (комментариев к постам в социальных медиа) по трем тематикам, связанным с пандемией COVID-19: вакцинация, карантин и ношение масок. Система, занявшая первое место по обеим задачам, использовала NLI вариант (Natural Language Inference – вывод по тексту) применения архитектуры BERT, автоматический перевод на английский язык для использования специализированной модель BERT, дообученной на постах Твиттера, обсуждающих ковид, а также дополнительное маскирование целевых сущностей. Эта система показала следующие результаты: для задачи определения позиции F1-score=0.6968, для задачи классификации доводов F1-score=0.7404. Мы надеемся, что подготовленные наборы данных и методы помогут стимулировать дальнейшие исследования по анализу аргументации для русского языка.

1   Introduction

People have been constantly arguing at all social levels and the *Argumentation Theory* was developed to study and control the process of coming to a conclusion from premises through logical reasoning. According to this theory, an argument must include a *claim* containing a *stance* towards some topic or object, and at least one *premise* ("favor" or "against") of this stance. Often a "premise" is called an "argument" when it is clear from the context which claim it is being referred to.

With the development of intellectual systems and neural networks, arguments can now be both produced and studied automatically. Therefore, the *Computational Argumentation* task arose to address the problem of computational analysis and synthesis of natural language argumentation. In this paper we focus on its branch — *Argument Mining* (or *Argumentation Mining*) — which explores methods for extracting arguments and their relationships from texts, as well as constructing an argumentative structure.

There is a large number of works devoted to this field which are thoroughly reviewed by (Stede and Jodi, 2018; Lawrence and Reed, 2020; Stede, 2020; Vecchi et al., 2021; Schaefer and Stede, 2021). Special attention has been paid to the stance detection as a sub-task of Argument Mining (Küçük and Can, 2020; ALDayel and Magdy, 2021; Küçük and Can, 2021) where the authors describe the proposed approaches so far, descriptions of the relevant datasets and tools, and some other related issues.

The growing interest in the task is justified by the application of the Argument Mining algorithms for argument search, fact checking, automated decision making, argument summarization, writing support and intelligent person assistants. For instance, Args.me[1], ArgumenText[2] and CAM (comparative argumentative machine)[3](Schildwächter et al., 2019) are well known systems widely used for searching arguments.

The main research forum for the task is the Argument Mining workshop series. Since 2014, eight workshops on the analysis of arguments have already been held[4] addressing burning issues like multi-task learning (Tran and Litman, 2021; Putra et al., 2021) and Argumentation Mining in different areas (science (Lauscher et al., 2018; Fergadis et al., 2021), news articles (Bauwelinck and Lefever, 2020) and cross-lingual research (Rocha et al., 2018)). Moreover, there are several shared tasks on the topic adjacent to the Argument Mining: Shared Task on Argumentation Mining in Newspaper Editorials (Kiesel et al., 2015), SemEval-2016 (Stance Detection) (Mohammad et al., 2016), Touché (Argument Retrieval) in 2020 (Bondarenko et al., 2020) and 2021 (Bondarenko et al., 2021).

In this paper, we present RuArg-2022 — the first shared task on Argument Mining for the Russian language. It consists of two sub-tasks: stance detection and premise classification. The first task aims to determine the point of view (stance) of the text's author in relation to the given claim. The second task is devoted to classification of texts according to premises ("for" or "against") to a given claim.

To highlight the differences between the two tasks, consider the following example: *Я против масок, но приходится их носить: мне проще так, чем с кем-то что-то обсуждать и кому-то что-то доказывать* (*I am against masks, but I have to wear them: it's easier for me than to discuss something with someone and prove something to someone*). In this sentence there is an explicit stance against masks but it gives a premise for wearing masks.

The contribution of the current paper is three-fold. First, we prepare a gold standard dataset for stance detection and premise classification. Second, we develop and release a baseline for the argument mining tasks that uses a multi-task multi-label BERT architecture. Third, we compare and analyse the results of the participants of the shared task and propose steps for further improvement of both sub-tasks. All the materials and data could be found on GitHub[5] and CodaLab[6] competition pages.

Thus, our work is the first, to the best of our knowledge, dealing with argument mining task for the Russian language. While our setup is a simple text categorization task, we argue that it may be an

---

[1]https://www.args.me/index.html

[2]https://www.informatik.tu-darmstadt.de/ukp/research_ukp/ukp_research_projects/ukp_project_argumentext/index.en.jsp

[3]http://ltdemos.informatik.uni-hamburg.de/cam/

[4]https://2021.argmining.org

[5]https://github.com/dialogue-evaluation/RuArg

[6]https://codalab.lisn.upsaclay.fr/competitions/786

important building block of larger argument mining pipelines featuring retrieval of arguments from large text collections (Bondarenko et al., 2020).

2   Previous Work

The Argument Mining task (Palau and Moens, 2009) involves the automatic identification of argumentative structures in free text. According to (Cabrio and Villata, 2020), "researchers have investigated argument mining on various registers including legal texts, scientific papers, product reviews, news editorials, Wikipedia articles, persuasive essays, political debates, tweets, and online discussions". A detailed overview of all argument mining related papers is out of the scope of the current work. We refer the reader to the recent surveys on this topic: (Lawrence and Reed, 2020) and (Schaefer and Stede, 2021).

The topic of COVID-19 is nowadays popular not only in the biomedical field, but also in social science and, especially, NLP research (Verspoor et al., 2020). There already exist several datasets on COVID-19 for stance detection (Wührl and Klinger, 2021), argument mining and fact extraction/verification. For instance, in (Beck et al., 2021) the authors collect a dataset from German Twitter on people's attitude towards the government measures. First, they identify relevant tweets for governmental measures and if relevant, detect what stance is expressed. (Menin et al., 2022) create a linked data version of the CORD-19 data set and enriched it via entity linking and argument mining.

Another dataset collected lately (Reddy et al., 2021) comprises the following topics related to COVID-19: origin of the virus, transmission of the virus, cure for the virus and protection from the virus. The authors present a pipeline for detecting claim boundaries and detecting stance. Unlike most stance detection datasets (Hanselowski et al., 2019; Allaway and McKeown, 2020) it involves identifying the claimer's stance within a claim sentence and not the stance for target–context pairs. (Li et al., 2022) follows (Reddy et al., 2021) and identify the stance from the perspective of each claimer, namely whether the claimer affirms or refutes a claim. They finetune a Bart-large model (Lewis et al., 2020) to automatically identify the stance. One more dataset related to the COVID-19 pandemic is collected by seven science teachers through three scenarios (Atabey, 2021). This dataset contains not only stances about vaccination, curfew and distance education, but also arguments and supporting reasons that might construct an argument mining dataset.

Most of the above mentioned publications on COVID-19 stance detection refer to the FEVER-like dataset COVID-Fact (Saakyan et al., 2021) of 4,086 claims concerning the COVID-19 pandemic. It could be also applied for the argument mining needs. Another dataset for COVID-19 fact checking is presented in (Liu et al., 2020) which also could be reformatted for the argument mining tasks.

As regards argumentation mining for the Russian language, there are not so many studies and datasets on the topic. (Fishcheva and Kotelnikov, 2019) translated into Russian and researched the English language Argumentative Microtext Corpus (ArgMicro) (Peldszus and Stede, 2015; Skeppstedt et al., 2018). In (Fishcheva et al., 2021) this corpus was expanded with machine translation of the Persuasive Essays Corpus (PersEssays) (Stab and Gurevych, 2014). XGBoost and BERT were applied to classify "for"/"against" premises.

Salomatina et al. (Salomatina et al., 2021) propose an approach to the partial extraction of the argumentative structure of a text by using patterns of argumentation indicators. They also try to recognize the relations between extracted arguments. (Ilina et al., 2021) develop a web resource for analysis of argumentation in popular science discourse. The annotation model is based on the ontology of argumentation and D. Walton's argumentation schemes (Walton et al., 2008). A scenario of argument annotation of texts allows constructing an argumentative graph based on the typical reasoning schemes.

To the best of our knowledge, there are no manually labelled publicly available datasets in Russian. In this work we present such dataset for the first time.

3   Dataset

The dataset is based on VKontakte users' comments discussing COVID-2019 news texts (Chkhartishvili et al., 2021). We choose the COVID-19 pandemic (and anti-epidemic measures in general) as the topic

of the dataset because we assume that the analysis of arguments on social measures against COVID-19 is still relevant for the modern society and especially for the understanding of the current public sentiment. From the gathered comment collection, sentences discussing masks, vaccines and quarantine were extracted using keywords (Nugamanov et al., 2021).

The annotation process included two stages: labelling by stance and labelling by premises. At both stages sentences were labelled in relation to the following claims:

1. "Vaccination is beneficial for society."
2. "The introduction and observance of quarantine is beneficial for society."
3. "Wearing masks is beneficial for society."

In the following subsections we describe the annotation process of the dataset for both sub-tasks: Stance Detection 3.1 and Premise Classification 3.2.

Stance	Premise	Numerical label
for	for	2
other (neutral/contradictory/unclear)	no argument	1
against	against	0
irrelevant	irrelevant	-1

Table 1: System of categories used to label the dataset.

3.1 Stance Annotation

The current dataset has been already annotated in (Nugamanov et al., 2021). In the current work the dataset was additionally checked and synchronized with premise annotation from the second step.

At the first stage of stance annotation each sentence was labelled by several experts (three on average). An annotator should indicate the stance it expresses towards each of the above-mentioned aspects (or indicate that the sentence is not relevant to the aspect). The annotators' group included professional linguists and psychologists. We consider four stance labels, namely:

- **for:** positive stance, which means that the speaker expresses his support for the topic;
- **against:** negative stance — the topic of discussion is not endorsed by the speaker;
- **other:** neutral stance (this label is used for factual sentences without any visible attitudes from the author); contradictory stance (for such a label, evident positive and negative attitudes should be seen in a message); unclear stance (the presence of a stance is seen, but the context of sentence does not give possibility to determine it);
- **irrelevant:** text does not contain stance on the topic.

The coding scheme for the stance annotation is presented in Table 1.

A sentence is considered to be relevant to an aspect, if at least two annotators considered it relevant. Sentences collected using keywords also can be irrelevant, for example a sentence mentioning Elon Musk ("Mask" in Russian spelling) is not relevant to the mask aspect.

3.2 Premise Annotation

At the second stage of annotation, the dataset was also annotated by premises for all three claims. The following four classes (labels) were used:.

- **for:** the stance is supported with argument in favor of the topic;
- **against:** the argument explains the author's negative outlook on the topic;
- **no argument:** no explanation is given for supporting/critisism of the topic;
- **irrelevant:** text does not contain stance and, consequently, premise on the topic.

The annotated sentences from the previous step were divided into three subsets: training, validation, and test (see Subsection 3.4). The labelling of each sentence by premises from the training and validation datasets was carried out by three annotators; the test sentences were labelled by four annotators. The final

labels for training and validation datasets were assigned with the agreement of at least two annotators, for test dataset – with the agreement of at least three annotators.

A sentence was considered as a premise if the annotator could use it to convince an opponent about the given claim, such as "*Masks help prevent the spread of disease.*" Detailed instructions for annotators are available in the competition repository[7].

The task of premise annotation should be separated from stance detection and sentiment analysis tasks. For example, the following statement does not contain a premise in relation to masks, although there is an author's stance "for": *It is high time to involve the city of "brides" in the production of protective masks*. It is also necessary to distinguish between sentiment polarity (positive and/or negative) and argumentation. In the following sentence there is a negative polarity towards quarantine, a positive polarity towards Trump, but no rational premises "for" or "against" quarantine are given: *And the fact that Trump did not introduce a suffocating quarantine is well done!*

The difference between the two tasks is illustrated in Table 2. For example, the first sentence possesses both stance and premise: the speaker expresses his negative attitude towards the vaccine by the reason of its short-term effectiveness. The second sentence, on the contrary, definitely supports vaccination without giving any specific arguments to support his/her opinion.

Text	Masks		Quarantine		Vaccines	
	Stance	Premise	Stance	Premise	Stance	Premise
И какой смысл в вакцине если антитела только 3 месяца? *(And what's the point of a vaccine if the antibodies work only for 3 months?)*	—	—	—	—	against	against
Должна быть вакцина которую, будут прививать с детства!!! *(There must be a vaccine that will be vaccinated from childhood!!!)*	—	—	—	—	for	no argument
Вот только там на момент, когда была 1000 выявленных, уже неделю карантин действовал. *(At the time when there were 1000 identified, quarantine had been in effect for a week.)*	—	—	other	against	—	—
Развитие ситуации: если соблюдать карантин месяц, то вирус будет остановлен. *(The development of the situation: if the quarantine is observed for a month, the virus will be stopped.)*	—	—	for	for	—	—
Вопрос к властям :почему из гос резерва не получили люди масок когда их не хватало или и резерва уже нет *(Question to the authorities : why didn't people get masks from the state reserve when there were not enough of them or there is no reserve anymore)*	for	no argument	—	—	—	—
Любители масок не ужели вы думаете что эта косметическая тряпочка поможет от вируса?! *(Mask lovers don't you really think that this cosmetic rag will help against the virus?!)*	against	no argument	—	—	—	—

Table 2: Examples for each topic — masks, vaccines, and quarantine (we keep the original spelling and punctuation). Note that for each topic, annotation of stances and premises was performed. Refer to Table 1 for the classification schema used to label the data. "Irrelevant" class is denoted as "—".

3.3 Dataset Verification

After completion of stance and premise annotation procedures, we verified the labels of the dataset. To this end, we looked through the contingency tables of stances and premises, and checked the following issues:

---

[7]https://github.com/dialogue-evaluation/RuArg/tree/main/annotation.

1. the sentence with an irrelevant label for one of the sub-tasks cannot be relevant for another sub-task;
2. the sentence with contradictory stance and premise (e.g., positive stance but premise "against") should be examined more carefully.

As a result, annotations for 289 sentences (3.0% from the whole dataset containing such issues) were revised and improved. Generally, if a sentence contains both a stance and a premise, then their polarity coincides (both "for" or both "against"). However, in 12 sentences the polarity is opposite. This is due to the fact that the sentence simultaneously contains the author's point of view and indicates the opponents' premises, for example: "*This is exactly why everyone should wear masks, but the main channels broadcast that masks are not needed and useless for healthy people.*"

3.4 Dataset Statistics

Each sentence has 6 labels: for each of the two sub-tasks (stance detection and premise classification) there is a label for each of the three aspects (masks, vaccines and quarantine).

The inter-annotator agreement was calculated by Krippendorff's alpha and it turned out quite high – 0.84. Dataset statistics are presented in Table 3. As one may observe the dataset is skewed (imbalanced). There are various schemes in the literature to perform evaluation of this kind of data (Rosenberg, 2012). We resort to a scheme by simply excluding the largest "irrelevant" class as it is done in Sentiment Analysis. For instance, at the SemEval-2016 Task 14 (Nakov et al., 2016) the organizers exclude the "NEUTRAL" class from the evaluation as the largest one.

The distribution of labels by class is shown in Figure 2

Dataset	Total	Stance			Premise			Irrelevant
		For	Other	Against	For	No argument	Against	
Masks								
train	6,717	704	1,832	594	339	2,451	340	3,587
val	1,431	148	388	126	62	542	58	769
test	1,402	147	401	123	63	523	85	731
all	9,550	999	2,621	843	464	3,516	483	5,087
Quarantine								
train	6,717	587	1 341	172	217	1,756	127	4,617
val	1,431	125	290	39	46	369	39	977
test	1,402	116	274	40	50	358	22	972
all	9,550	828	1,905	251	313	2,483	188	6,566
Vaccines								
train	6,717	374	866	418	149	1,238	271	5,059
val	1,431	78	183	92	24	282	47	1,078
test	1,402	75	181	81	21	262	54	1,065
all	9,550	527	1,230	591	194	1,782	372	7,202

Table 3: Statistics of the constructed dataset used in RuArg-2022 shared task.

4 Evaluation

The main performance metric in each of the two sub-tasks are $F1_{stance}$ and $F1_{premise}$ scores, which are calculated according to the following formula:

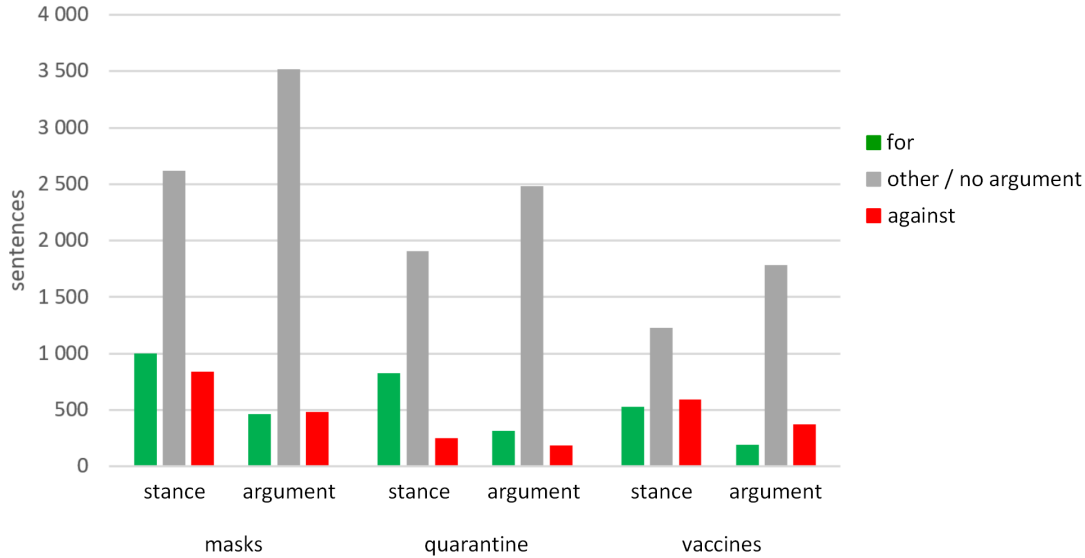$$F1 = \frac{1}{n} \sum_{c \in C} F1_{rel_c}, \qquad (1)$$

Figure 1: Distribution of the labels.

where $C = \{\text{"}masks\text{"}, \text{"}vaccines\text{"}, \text{"}quarantine\text{"}\}$, n is the size of C and $F1_{rel}$-score is macro F1-score averaged over first three relevance classes (the class "irrelevant" is excluded). Namely, the following procedure is used:

1. F1-scores are averaged over three out of four classes (the "irrelevant" class is excluded) – macro $F1_{rel}$-score is obtained for a given claim;
2. macro $F1_{rel}$-scores for all three claims are averaged – we get macro F1-score relative to the task (stance detection or premise classification);
3. For each of the three claims, F1-score is calculated for each class (label) separately.

As a result, two main macro $F1_{rel}$-scores are calculated – one for each sub-task. Participants' systems are ranked by these metrics (two separate lists). The $F1_{rel}$-score for claims and F1-score for individual classes (labels) will be also discussion in Section 7.

5   Baseline

We implement a simple baseline that finetunes the pre-trained ruBERT model (Devlin et al., 2019; Kuratov and Arkhipov, 2019) on the provided dataset. We chose "DeepPavlov/rubert-base-cased" model from Hugging Face [8]. We experiment with training a single model that predicts all the required labels. However, it did not performed well, so we finetune three pre-trained BERT models separately for three topics: "masks", "vaccines", and "quarantine". Each model comprises the following layers:

1. the pre-trained BERT layer with the unfrozen weights;
2. a dense layer for stance detection;
3. a dense layer for argument classification.

Then we applied categorical cross-entropy loss to train on both stance and argument labels simultaneously. The results are presented in Section 7.

6   Participating Systems

RuArg-2022 shared task attracted 16 participants, 13 of them participated in the final phase. We provide descriptions of the top 7 solutions which outperformed the baseline for at least one sub-task. We denote

---

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

each team either with its team name (if any) or with their CodaLab user names. In cases of multiple submissions from one team, we report only the best result. The scores of the teams are shown in Table 4.

camalibi (msu)   First, this team used RuBERT-classifier[9] to determine the relevance of the texts using NLI-method: to form an input example, a second sentence with the aspect ("masks", "quarantine", or "vaccination") was added to each original sentence from the dataset. The output 1 was for the "Relevant" result and 0 for "Irrelevant".

For the stance classification task the texts were pre-processed and then translated into English using pretrained seq2seq-model[10]. Then, each text was processed according to the rule: $keyword \rightarrow @*ASPECT*keyword@$, where $ASPECT$ is the aspect for which a given text is relevant and $keyword$ is the word from a list of words corresponding to each aspect.

Then for both RuArg sub-tasks the domain-specific BERT-classifier[11] was trained using NLI-method: for each text and each aspect for which a given text is relevant, six input examples were constructed (three for each stance label and three for each premise label). Final input examples looked like:

- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"Against quarantine"**,
- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"None-stance quarantine"**,
- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"In-favor quarantine"**,
- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"Negative to quarantine"**,
- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"Neutral to quarantine"**,
- "Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the @ * quarantine * quarantine @.", **"Positive to quarantine"**.

The stance or premise label was chosen as the one where the corresponding input example had the maximum softmax output.

sevastyanm (vyatsu)   This participant utilized pre-trained ruRoberta-large language model[12] which was trained on additional data obtained from "PersEssays_Russian" and "ArgMicro_Russian" datasets (Fishcheva et al., 2021) with similar annotation schemes. Both datasets were united by argumentative discourse units and used to train model to solve 4-class classification problem.

First, ruRoberta-large was fine-tuned on the united "PersEssays_Russian" and "ArgMicro_Russian" dataset with 8,780 units. Then, the model was fine-tuned separately on each of 6 tasks from the competition dataset *('masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument')*. For final class prediction the participant used token averaging and 2-layer linear neural network classifier. All models trained with $learning\_rate = 10^{-5}$ and $weight\_decay = 0.01$.

For the model trained on the additional dataset the participant used the following hyperparameters: $input\_size = 70, num\_epochs = 3, batch\_size = 16$. For model trained on the RuArg dataset hyperparameters were as follows: $input\_size = 100, batch\_size = 32, num\_epochs = \{2, 2, 4, 4, 4, 7\}$ for *'masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument'* respectively.

iamdenay (IICT)   This team used the pre-trained Crosslingual RoBERTa-large model and fine-tuned on the augmented data. They mostly augmented the data containing stances and arguments about *"quarantine"*. For the augmentation the participants used mT5 model to paraphrase sentences in order to increase the size of the text set. To increase accuracy of the proposed method they used six different

---

[9]https://huggingface.co/DeepPavlov/rubert-base-cased-conversational
[10]https://huggingface.co/Helsinki-NLP/opus-mt-ru-en
[11]https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2
[12]https://huggingface.co/sberbank-ai/ruRoberta-large

models, one per task {*'masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument'*}.

ursdth This team proposed a pipeline-based framework for the classification of texts with or without recognizable rhetorical structure. The first stage involved fine-tuning sequential model on the classification dataset including texts of different lengths and complexity. In the second stage, they frozen the base model and then trained a discourse-aware neural module on top of it for the classification of texts with discourse structure.

They used pre-trained Conversational RuBERT for the discourse unit classification. For texts with automatically recognizable discourse structure, they proposed a relation-aware Tree-LSTM over the discourse units' class predictions. Stance and premise labels were predicted jointly.

Both development and test datasets were treated as unseen, and the official development dataset was not used for the parameters adjustment. The predictions were obtained by averaging outputs from five models trained on cross-validation during experiments over labeled data. This is similar to an ensemble, where each model is trained using 80% of the train data.

sopilnyak (auteam) This team started with training a classifier to detect irrelevant sentences for each sub-task. They applied binary Logistic Regression classifier trained on TF-IDF features, calculated from BPE tokens.

Then they excluded irrelevant sentences and further trained the models (for each sub-task separately) as a blend of:
1. fine-tuned ruRoberta-large from Sber AI with a two-layer classification head on top. They unfrozen 30 top layers and used very low learning rate ($5 \cdot 10^{-6}$) to prevent model from over-fitting on a small dataset. Also they utilized weighted cross-entropy loss so that the results on unbalanced dataset would be more accurate.
2. Logistic Regression classifier on TF-IDF features calculated on BPE tokens.

kazzand This participant applied Transformer-based deep text feature extraction and hierarchical classification. Firstly, they trained simple TF-IDF + Logistic Regression pipeline for each text type (masks, quarantine, vaccines). Secondly, they trained 6 separate models for each task {*'masks_stance', 'masks_argument', 'quarantine_stance', 'quarantine_argument', 'vaccines_stance', 'vaccines_argument'*} using Sentence-BERT for embeddings computation served as input to the Logistic Regression or KNN model.

invincible The first step for the team was a preprocessing: they removed punctuation symbols, converted text to lowercase, and removed special symbols including the "[USER]" substring. They further used the DistilRuBERT model[13] to vectorize the text into a vector of numbers and saved as a row of the new matrix. This feature matrix was used as an input for the classification models.

Overall, there were nine models, three for each topic {"masks", "vaccines", "quarantine"}. The initial data were separated into three subsets corresponding to each topic. Then the following algorithm was applied: first, SVM model (with sigmoid kernel and balanced target) detected irrelevant sentences for each topic and classified them as "irrelevant" for both stance and argument types. Then for positive-classified sentences, two neural network models were applied. They consisted of Flatten layer, Dense layer with ReLU activation function, Dropout layer and final Dense layer with Sigmoid activation function. They used the following hyperparameters: $optimizer = $"adam", $loss = $ "sparse_categorical_crossentropy", $num\_epochs = 5$.

Importantly, since the classes are highly unbalanced (labels "for" and "against" are highly underrepresented), a random oversampling was applied to all dataframes before fitting the models. Accurate class-balancing allowed improving both scores significantly.

7 Results and Discussion

Table 4 presents respectively the results for "stance detecion" and "premise classificaiton" tracks.

---

[13]https://huggingface.co/DeepPavlov/distilrubert-tiny-cased-conversational

#	Participant	Base Transformer model	Additional data	Stance F1-score	#	Premise F1-score	#
1	camalibi	covid-twitter-bert-v2	Yes	0.6968	1	0.7404	1
2	sevastyanm	RuRoBERTa-large	Yes	0.6815	2	0.7235	2
3	iamdenay	RuRoBERTa-large	Yes	0.6676	3	0.6555	4
4	ursdth	RuBERT Conversational	No	0.6573	4	0.7064	3
5	sopilnyak	RuRoBERTa-large	No	0.5603	5	0.4438	10
6	kazzand	Sentence-BERT	No	0.5552	6	0.5603	6
7	morty	n/a	n/a	0.5353	7	0.5453	7
8	invincible	RuBERT Conversational	Yes	0.5286	8	0.5428	8
9	dr	n/a	n/a	0.4750	9	0.6036	5
10	baseline	ruBERT	No	0.4180	10	0.4355	9

Table 4: Competition results of the participant systems.
The places of participants for each sub-task are indicated in the brackets.

All the results are quite stable for both sub-tasks, only **sopilnyak** did not manage to overcome the premise baseline, demonstrating high results (top-5) at the stance detection sub-task. The range of the models used to solve the task is not wide: the participants choose between (ru)BERT, (crosslingual)RoBERTa(-large) and old good Logistic Regression model.

In comparison to the baseline, all the participants trained classification models separately for each sub-task. Evidently, multitask classification is more challenging than training classification models separately.

Interestingly, several best results were obtained with the help of the additional datasets or/and data augmentation (**camalibi** – top-1, **sevastyanm** – top-2, **iamdenay** – top-3 for stance detection and top-4 for premise classification, and also **invincible** – top-8). Top-1 **camalibi** used the special version of BERT model in which domain-oriented dataset was actually integrated; top-2 **sevastyanm** utilized additional dataset, top-3/top-4 **iamdenay** applied mT5 for paraphrase generation, top-8 **invincible** used random oversampling. From these observations we can assume that any kind of additional data is beneficial for these tasks, however, the more diverse the data is, the better.

The most different and outstanding approach in comparison to other participants was presented by the winner system of **camalibi**. This participant applied NLI method which performed best for both sub-tasks. Moreover, model trained on the English language was applied, therefore, **camalibi** did translate the whole dataset for the task.

We also compared the detailed results for the top 5 systems and the baseline. The scores are presented in Appendix A. From Figure 2 we can see that F1-scores for Premise Classification are slightly higher than for Stance Detection. The task of Stance Detection is equally hard for all three topics, whereas we can see that the scores for Masks Premise and Quarantine Premise are higher than Vaccines Premise results. It can be seen that the difference between the top 3 participants are not very much different from each other. As for the baseline results, we can see that the results for vaccine and quarantine are two times lower than the results of the (at least) top 4 participants. At the same time, Masks Stance and Premise results are higher than for vaccines and quarantine and not significantly different from the top results. To sum up, we can conclude that the algorithms for the top 3 results demonstrate similar results across different subsets.

8   Conclusion

We present the results of the first shared task on Argument Mining for Russian. For this shared task, we created a new dataset on the vital COVID-19 topic. We introduce and rely on the following claims:
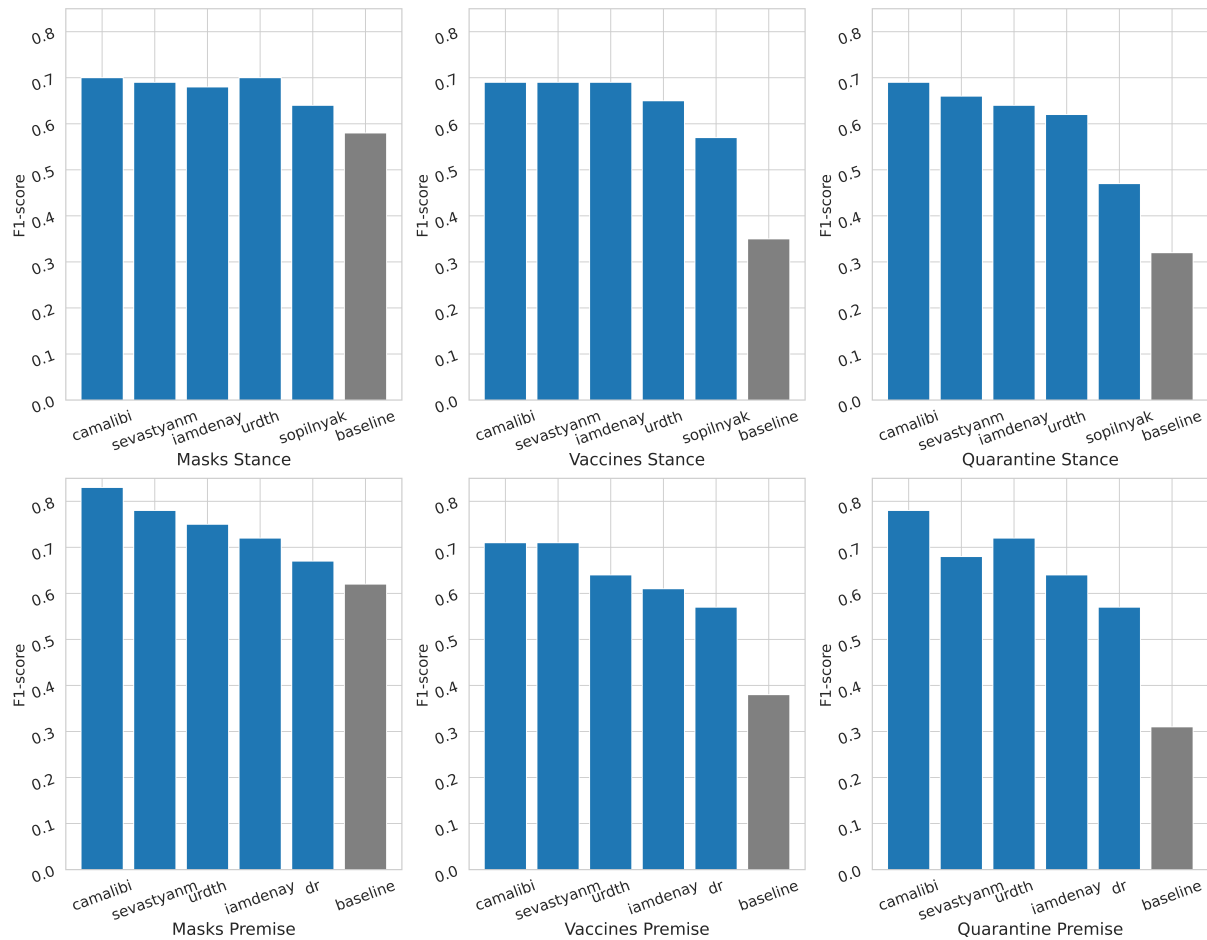
Figure 2: Results for the first top 5 participants for each sub-task.

"Vaccination is beneficial for society", "The introduction and observance of quarantine is beneficial for society", and "Wearing masks is beneficial for society".

Overall, 13 teams participated in the shared task, and more than half of them outperformed the baseline model. The winning system in both sub-tasks used the NLI (Natural Language Inference) variant of the BERT architecture, automatic translation into English to apply a specialized BERT model, pretrained on Twitter posts discussing COVID-19, and additional masking of target entities. This system showed for stance detection F1-score of 0.6968, for premise extraction F1-score of 0.7404 which considerably outperforms the proposed BERT-based baseline (F1-scores of 0.4180 and 0.4355, respectively).

According to the provided results, we see that the argument mining is a feasible task, especially on the COVID-19 dataset. All the data and codes are available online.[14] We hope that these materials will help to foster further research and developments in the area of argument mining for the Russian language.

As future work, we see it promising to explore more complex argument mining setups such as sequence tagging (Chernodub et al., 2019) or information retrieval (Bondarenko et al., 2020).

Acknowledgements

---

[14]https://github.com/dialogue-evaluation/RuArg
[15]https://rscf.ru/en/project/22-21-00885/

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8913–8931, Online, November. Association for Computational Linguistics.

Nejla Atabey. 2021. Science teachers' argument types and supporting reasons on socioscientific issues: Covid-19 pandemic. *International Journal of Psychology and Educational Studies*, 8(2):214–231.

Nina Bauwelinck and Els Lefever. 2020. Annotating topics, stance, argumentativeness and claims in Dutch social media comments: A pilot study. // *Proceedings of the 7th Workshop on Argument Mining*, P 8–18, Online, December. Association for Computational Linguistics.

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in german covid-19 social media. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P (to appear), virtual conference, August. Association for Computational Linguistics.

Alexander Bondarenko, Matthias Hagen, Martin Potthast, Henning Wachsmuth, Meriem Beloucif, Chris Biemann, Alexander Panchenko, and Benno Stein. 2020. Touché: First shared task on argument retrieval. // *Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, P 517—-523.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of Touché 2021: Argument Retrieval. // K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021)*, volume 12880 of *Lecture Notes in Computer Science*, P 450–467, Berlin Heidelberg New York, September. Springer.

Elena Cabrio and Serena Villata. 2020. *Proceedings of the 7th Workshop on Argument Mining*, Online, December. Association for Computational Linguistics.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, P 195–200, Florence, Italy, July. Association for Computational Linguistics.

Alexander Chkhartishvili, Dmitry Gubanov, and Ivan Kozitsin. 2021. Covid-19 information consumption and dissemination: A study of online social network vkontakte. // *2021 14th International Conference Management of large-scale system development (MLSD)*, P 1–5. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. // *Proceedings of the 8th Workshop on Argument Mining*, P 100–111, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-Lingual Argumentation Mining for Russian Texts. // *Proceedings of the 8th International Conference "Analysis of Images, Social networks and Texts" (AIST 2019), Lecture Notes in Computer Science*, P 134–144.

Irina Fishcheva, Valeriya Goloviznina, and Evgeny Kotelnikov. 2021. Traditional machine learning and deep learning models for argumentation mining in russian texts. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2021"*, P 246–258.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. // *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, P 493–503, Hong Kong, China, November. Association for Computational Linguistics.

Daria Ilina, Irina Kononenko, and Elena Sidorova. 2021. On Developing a Web Resource to Study Argumentation in Popular Science Discourse. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2021"*, P 318–327.

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. 2015. A shared task on argumentation mining in newspaper editorials. // *Proceedings of the 2nd Workshop on Argumentation Mining*, P 35–38, Denver, CO, June. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Survey*, 53(1):1–37.

Dilek Küçük and Fazli Can, 2021. *Stance Detection: Concepts, Approaches, Resources, and Outstanding Issues*, P 2673–2676. Association for Computing Machinery, New York, NY, USA.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. // *Proceedings of the 5th Workshop on Argument Mining*, P 22–28, Brussels, Belgium, November. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 7871–7880, Online, July. Association for Computational Linguistics.

Manling Li, Revanth Gangi Reddy, Ziqi Wang, Yi-Shyuan Chiang, Tuan M Lai, Pengfei Yu, Zixuan Zhang, and Heng Ji. 2022. Covid-19 claim radar: A structured claim extraction and tracking system. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) Demo Track*.

Zhenghao Liu, Chenyan Xiong, Zhuyun Dai, Si Sun, Maosong Sun, and Zhiyuan Liu. 2020. Adapting open domain fact extraction and verification to COVID-FACT through in-domain language modeling. // *Findings of the Association for Computational Linguistics: EMNLP 2020*, P 2395–2400, Online, November. Association for Computational Linguistics.

Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, and Marco Winckler. 2022. Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, 2(4):1301–1323, 02.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. // *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2016)*, P 31–41.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. // *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, P 1–18, San Diego, California, June. Association for Computational Linguistics.

Eduard Nugamanov, Natalia Loukachevitch, and Boris Dobrov. 2021. Extracting sentiments towards COVID-19 aspects. // *Supplementary 23rd International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2021*, P 299–312.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. // *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, P 98–107, New York, NY, USA. Association for Computing Machinery.

Andreas Peldszus and Manfred Stede. 2015. An Annotated Corpus of Argumentative Microtexts. // *Proceedings of the 1st European Conference on Argumentation*, P 801–815.

Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. 2021. Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance. // *Proceedings of the 8th Workshop on Argument Mining*, P 12–23, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. 2018. Cross-lingual argumentative relation identification: from English to Portuguese. // *Proceedings of the 5th Workshop on Argument Mining*, P 144–154, Brussels, Belgium, November. Association for Computational Linguistics.

Andrew Rosenberg. 2012. Classifying skewed data: Importance weighting to optimize average recall. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 3:2239–2242, 01.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. // *ACL/IJCNLP*.

Natalia Salomatina, Irina Kononenko, Elena Sidorova, and Ivan Pimenov. 2021. Identification of connected arguments based on reasoning schemes "from expert opinion". *Journal of Physics: Conference Series*, 1715.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Matthias Schildwächter, Alexander Bondarenko, Julian Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. Answering comparative questions: Better than ten-blue-links? // *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, P 361–365, New York, NY, USA. Association for Computing Machinery.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. // *Proceedings of the 5th Workshop on Argument Mining*, P 155–163.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. // *Proceedings of the International Conference on Computational Linguistics*, P 1501–1510.

Manfred Stede and Schneider Jodi. 2018. *Argumentation Mining*. Morgan & Claypool.

Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.

Nhat Tran and Diane Litman. 2021. Multi-task learning in argument mining for persuasive online discussions. // *Proceedings of the 8th Workshop on Argument Mining*, P 148–153, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, P 1338—1352.

Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace. 2020. *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. // *Proceedings of the 20th Workshop on Biomedical Language Processing*, P 131–142, Online, June. Association for Computational Linguistics.

A All results

From the detailed results for each label we can see that the top 1 result is not always the best approach: many models from the top list perform equally well or even outperform the winner in many cases. For instance, from Tables 5, 7 and 8 we can see that **sevastyanm** and **urdth** are very competitive approaches. The top 1 ranking is achieved by demonstrating stable results across different subsets and good (normally best) precision scores. Interestingly, the best recall scores on quarantine subsets and the vaccines premise subset is achieved by the baseline (more than 0.95 points).

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	**1.00**	**1.00**	**1.00**	**0.68**	**0.66**	**0.67**	0.80	0.83	**0.81**	**0.65**	0.61	0.63	**0.71**	**0.70**	**0.70**
sevastyanm	**1.00**	**1.00**	**1.00**	0.65	0.63	0.64	0.80	0.82	**0.81**	0.62	0.62	0.62	0.69	0.69	0.69
iamdenay	**1.00**	0.99	**1.00**	0.67	0.65	0.66	0.77	0.81	0.79	0.61	0.56	0.58	0.68	0.67	0.68
urdth	**1.00**	**1.00**	**1.00**	0.62	0.65	0.63	**0.82**	0.80	**0.81**	0.64	**0.65**	**0.65**	0.69	**0.70**	**0.70**
sopilnyak	0.98	**1.00**	0.99	0.64	0.50	0.56	0.77	**0.86**	**0.81**	0.63	0.50	0.56	0.68	0.62	0.64
baseline	0.99	0.99	0.99	0.46	0.49	0.47	0.77	0.77	0.77	0.50	0.48	0.49	0.58	0.58	0.58

Table 5: Results for masks stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	**1.00**	**1.00**	**1.00**	0.78	**0.73**	**0.76**	**0.94**	0.94	**0.94**	**0.73**	**0.84**	**0.78**	**0.82**	**0.84**	**0.83**
sevastyanm	**1.00**	**1.00**	**1.00**	**0.85**	0.61	0.71	0.93	**0.96**	**0.94**	0.65	0.70	0.67	0.81	0.76	0.78
urdth	**1.00**	**1.00**	**1.00**	0.67	**0.73**	0.70	0.93	0.93	0.93	0.68	0.60	0.64	0.76	0.75	0.75
iamdenay	**1.00**	**1.00**	**1.00**	**0.85**	0.47	0.61	0.89	**0.96**	0.93	0.64	0.62	0.63	0.79	0.68	0.72
dr	**1.00**	0.99	**1.00**	0.77	0.47	0.58	0.88	**0.96**	0.92	0.59	0.46	0.52	0.75	0.63	0.67
baseline	**1.00**	0.99	0.99	0.56	0.41	0.48	0.89	0.92	0.91	0.44	0.51	0.47	0.63	0.61	0.62

Table 6: Results for masks premise for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	**1.00**	**1.00**	**1.00**	0.78	0.58	0.67	0.72	**0.86**	**0.79**	**0.71**	0.56	**0.63**	**0.74**	0.67	**0.69**
sevastyanm	**1.00**	**1.00**	**1.00**	0.71	**0.74**	**0.73**	**0.75**	0.75	0.75	0.63	**0.59**	0.61	0.70	**0.69**	**0.69**
iamdenay	**1.00**	**1.00**	**1.00**	**0.79**	0.60	0.69	0.71	**0.86**	0.78	0.70	0.52	0.60	0.73	0.66	**0.69**
urdth	**1.00**	**1.00**	**1.00**	0.67	0.57	0.61	0.70	0.82	0.76	0.68	0.52	0.59	0.68	0.64	0.65
sopilnyak	0.99	**1.00**	**1.00**	0.61	0.43	0.51	0.67	0.79	0.72	0.54	0.41	0.47	0.61	0.55	0.57
baseline	0.99	**1.00**	0.99	0.43	0.15	0.22	0.56	0.85	0.67	0.38	0.11	0.17	0.46	0.37	0.35

Table 7: Results for vaccines stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	**1.00**	**1.00**	**1.00**	**0.79**	0.57	**0.67**	**0.89**	0.94	**0.92**	0.55	0.52	0.54	**0.75**	0.68	**0.71**
sevastyanm	**1.00**	**1.00**	**1.00**	0.63	**0.59**	0.61	0.90	0.92	0.91	**0.67**	**0.57**	**0.62**	0.73	**0.69**	**0.71**
urdth	**1.00**	**1.00**	**1.00**	0.56	**0.59**	0.58	**0.89**	0.88	0.88	0.45	0.48	0.47	0.64	0.65	0.64
iamdenay	**1.00**	**1.00**	**1.00**	0.60	0.50	0.55	0.86	0.85	0.86	0.34	0.52	0.42	0.60	0.63	0.61
dr	**1.00**	**1.00**	**1.00**	0.55	0.39	0.46	0.85	0.91	0.88	0.39	0.33	0.36	0.60	0.54	0.57
baseline	**1.00**	0.99	0.99	0.43	0.11	0.18	0.80	**0.95**	0.87	0.33	0.05	0.08	0.52	0.37	0.38

Table 8: Results for vaccines premise for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	0.99	**1.00**	**1.00**	**0.88**	0.35	**0.50**	0.84	0.83	0.83	**0.70**	0.81	**0.75**	**0.80**	**0.66**	**0.69**
sevastyanm	0.99	**1.00**	**1.00**	0.57	**0.40**	0.47	**0.85**	0.78	0.82	0.62	0.79	0.70	0.68	**0.66**	0.66
iamdenay	0.99	**1.00**	0.99	0.67	0.25	0.36	0.84	0.85	**0.84**	**0.70**	0.75	0.72	0.73	0.62	0.64
urdth	0.99	**1.00**	**1.00**	0.56	0.35	0.43	0.84	0.71	0.77	0.56	**0.82**	0.67	0.65	0.63	0.62
sopilnyak	0.98	**1.00**	0.99	0.00	0.00	0.00	0.77	0.80	0.78	0.59	0.67	0.63	0.45	0.49	0.47
baseline	**1.00**	0.99	**1.00**	0.00	0.00	0.00	0.65	**0.97**	0.77	0.62	0.11	0.19	0.42	0.36	0.32

Table 9: Results for quarantine stance for top-5 participants.

Method	Irrelevant			Against			Other			For			Macro Average		
	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1	Pr	R	F1
camalibi	0.99	**1.00**	**1.00**	**0.60**	0.27	0.37	0.91	0.96	**0.93**	**0.85**	0.68	0.76	**0.84**	0.73	**0.77**
sevastyanm	0.99	0.99	0.99	0.44	**0.50**	**0.47**	0.91	0.92	0.92	0.72	0.62	0.67	0.69	0.68	0.68
urdth	0.99	**1.00**	**1.00**	0.47	0.41	0.44	**0.95**	0.92	**0.93**	0.73	**0.88**	**0.80**	0.72	**0.74**	0.72
iamdenay	0.99	**1.00**	**1.00**	0.33	0.41	0.37	0.91	0.90	0.91	0.70	0.60	0.65	0.65	0.64	0.64
dr	0.99	**1.00**	**1.00**	0.42	0.23	0.29	0.88	0.94	0.91	0.66	0.42	0.51	0.65	0.53	0.57
baseline	**1.00**	0.99	**1.00**	0.00	0.00	0.00	0.82	**0.99**	0.90	0.50	0.02	0.04	0.44	0.34	0.31

Table 10: Results for quarantine premise for top-5 participants.