ZenPropaganda: A Comprehensive Study on Identifying Propaganda Techniques in Russian Coronavirus-Related Media

Anton Chernyavskiy[1], **Svetlana Shomova**[2], **Irina Dushakova**[2], **Ilya Kiriya**[2], **and Dmitry Ilvovsky**[1]
[1]Faculty of Computer Science, National Research University Higher School of Economics, Russia
[2]Institute of Media, National Research University Higher School of Economics, Russia
{acherniavskii, sshomova, idushakova, ikiria, dilvovsky}@hse.ru

Abstract

The topic of automatic detection of manipulation and propaganda in the media is not a novel issue; however, it remains an urgent concern that necessitates continuous research focus. The topic is studied within the framework of various papers, competitions, and shared tasks, which provide different techniques definitions and include the analysis of text data, images, as well as multi-lingual sources. In this study, we propose a novel multi-level classification scheme for identifying propaganda techniques. We introduce a new Russian dataset `ZenPropaganda` consisting of coronavirus-related texts collected from Vkontakte and Yandex.Zen platforms, which have been expertly annotated with fine-grained labeling of manipulative spans. We further conduct a comprehensive analysis by comparing our dataset with existing related ones and evaluate the performance of state-of-the-art approaches that have been proposed for them. Furthermore, we provide a detailed discussion of our findings, which can serve as a valuable resource for future research in this field.

Keywords: propaganda detection, manipulative techniques, fake news, Transformer-based models, RoBERTa

1. Introduction

Although the issue of manipulation and propaganda in the media is not new, it does not lose its urgency and requires ongoing research attention. In recent years, this problem has gained even more relevance due to the emergence of new digital media practices (Caled and Silva, 2022), the transmedia environment (Martin-Neira et al., 2023), the development of generative artificial intelligence (Wach et al., 2023), and the advent of technologies for manipulating textual and visual content (Arnaudo et al., 2021). At the same time, we face a controversial situation: the capabilities of algorithms increase the risks of propaganda impact on the audience, and on the other hand, these algorithms are actively involved in the fight against inaccurate information (Bagdasaryan and Shmatikov, 2022). Therefore, understanding the capabilities of modern artificial intelligence methods in shaping information processes has become a key objective of this research.

Our contributions can be summarized as follows:

- We introduce a revised and extended multi-level classification of propaganda techniques.
- We present a new Russian dataset `ZenPropaganda` comprised of COVID-19-related texts extracted from Vkontakte and Yandex.Zen with fine-grained labeling of manipulative spans.
- We perform comparative analysis of `ZenPropaganda` with existing related datasets and evaluate the performance of state-of-the-art approaches proposed for them.
- We provide a detailed discussion of our findings and release our dataset and code at https://github.com/aschern/ru_zen_prop.

2. Related Work

The proliferation of "fake news" led to the active research in the field of detecting propaganda. Propaganda is typically characterized by a set of various manipulative techniques, which may include emotional appeals or logical fallacies. However, there exists variation among researchers regarding the precise set of propaganda techniques to be considered, as well as their respective definitions (Torok, 2015).

Jewett (1940) identifies seven techniques, Weston (2000) lists at least 24 and the Wikipedia[1] provides the list of 72 techniques. These techniques were partially explored in previous research tasks like hate speech detection (Gao et al., 2017) and computational argumentation (Habernal et al., 2018). SemEval-2020 Task 11 (Da San Martino et al., 2020) requested for a fine-grained propaganda detection that complemented existing approaches. In 2021, a new SemEval task introduced a multimodal dataset of memes annotated with an extended set of techniques (Dimitrov et al., 2021).

[1]https://en.wikipedia.org/wiki/Propaganda_techniques

LREC-COLING 2024, pages 17795–17807
20-25 May, 2024.

Other studies include WANLP-2022 shared task on detecting 20 propaganda techniques in Arabic tweets (Alam et al., 2022), as well as in Bulgarian COVID-19 related social media (Nakov et al., 2021). Additionally, a new SemEval-2023 task titled "Persuasion Techniques in Online News in a Multilingual Setup" (including Russian) (Piskorski et al., 2023), which employed a taxonomy of 23 persuasion techniques. In this paper, we present novel two-level classification of propaganda techniques, which is described below, and mark up a new Russian-language popular domain.

Automatic detection of propaganda fragments has been the subject of extensive research in shared tasks. Among these, the most successful methods have been those based on Transformer-based models like BERT, RoBERTa, XLNet, ALBERT, DeBERTa, mBERT, XLM-RoBERTa and their ensembles (Jurkiewicz et al., 2020; Chernyavskiy et al., 2020; Morio et al., 2020; Tian et al., 2021; Gupta et al., 2021; Sadeghi et al., 2023; Liao et al., 2023; Purificato and Navigli, 2023).

3. Our Dataset: `Zen-Propaganda`

3.1. Manipulative Content in the Media

Although various aspects of propagandistic or manipulative presentation of information in the media have been studied, there is still a gap in the research field regarding generalization and the creation of a comprehensive scheme for annotation.

Thus, the first key theoretical decision for the research design was the choice of manipulative techniques for annotation and building a hierarchy to develop an overarching scheme. To achieve this, we took as a basis the existing schemes used by the developers of similar algorithms, and we supplemented them with findings from other studies (Fedorov and Levitskaya, 2020; Anisimova et al., 2021; Da San Martino et al., 2020; Arnaudo et al., 2021), while ensuring compatibility in analytical procedures.

However, we encountered several challenges during the process. Firstly, there were interpretative and terminological issues. In Russian research, there is a lack of unified approaches to interpreting the concepts of propaganda and manipulation and to describing their tools. Secondly, there were analytical challenges in determining the inclusion of manipulative and other techniques that affect the reliability and relevance of the material.

These issues are tightly connected to the complexity of automating the analysis of media messages for manipulation. In general, both researchers and fact-checkers agree on the limited ability of algorithms to accurately interpret manipulative/propaganda content. Fact-checking work is often done manually, algorithms, rather, suggest only paying attention to a "doubtful" piece of material (as one of the examples for the analysis of COVID-19 misinformation, see (Brennen et al., 2020)).

Not being able to completely solve this problem, we would like to emphasize that our analysis of texts focused solely on techniques that could be identified within the messages themselves, without relying on external sources. As a result, we were unable to address common manipulative techniques such as misrepresentation of someone's position (when the interlocutor's thesis is replaced) or distortion of opponents' opinions by attributing statements or actions that they did not actually make, or to verify the accuracy of quotes and their translations, as this would require access to additional information. This significantly distances our work from fact-checking and allows us to focus on intra-text manipulation. One more problem is that the identified manipulative techniques are operationalized with varying degrees of accuracy.

To systematize the variety of techniques, three types were identified:

- narrative techniques or manipulative techniques that refer to the whole text, they are rather poorly systematized and are not easily amenable to automated analysis;
- manipulative techniques that refer to specific fragments of text;
- indicators of manipulation within the text (not necessarily related to separate techniques and can be automatically highlighted—for example, many exclamation marks, words written in caps, etc.).

This distinction shows different approaches to operationalizing manipulative techniques. When considering narrative techniques, there may not be one specific fragment of the text to which they can be applied; instead, they can be simply "smeared" throughout the text. Furthermore, these techniques can only be detected while switching from one text to another in the same time period. While it is possible to annotate the techniques that characterize the entire text, this research does not focus on the techniques that can be noticed in the corpus as a whole. This serves as both a limitation and a perspective of the research.

The markers of manipulation in the text demonstrate a high level of adaptability for automation of the annotation. These markers can be easily marked without being tied to individual techniques, since they can both serve as parts of multiple techniques at once and be independent signs. Typically, these markers are used as the basis for highlighting automatically generated manipulative content and are well recognized by the algorithms.

1) Storytelling prevails over facts (this will probably largely coincide with the division of texts into opinion and reporting)
2) Mocking, trolling (at a topic, phenomenon, organization, personality, state/country, nation, idea, symbol): used to defame something/someone, to show something insignificant and frivolous, not worthy of attention (this will probably largely coincide with annotating texts as satire)
3) Irrelevant data / uncheckable data (e.g., no exact names or sources)
4) Emotional "load" of the text – strong/negative emotions (appeal to fear or panic cause)
5) Discouraging critical thought and rational thinking: "shock content", appeal to basic emotions, references to traumatic and paranormal phenomena
6) Repetition: constant, obsessive repetition of certain statements, regardless of their truth
7) Obfuscation, Intentional vagueness, Confusion
8) Unbalanced representation of extremism (showing only the most radical positions as the only existing ones): unbalanced emphasis on only positive or only negative facts and arguments while ignoring the opposite;
9) Oversimplification: the plot "ordinary people, maximum simplicity, rubbing into trust", a bet on trusting relationships with a wide audience, its support under the pretext that ideas are maximally simplified, the communicator's suggestions have a positive value, since they are supposedly close to ordinary people ("I'm like you")
10) Slogans, myths, stereotypes
11) Personal promotion and self-promotion (a person, a group, a company)
12) Excessive usage of allegories and metaphors (each individual metaphor or allegory may not be a technique by itself, but their overuse indicates an imbalance in the text)

Table 1: Characteristics of the entire text.

3.2. Manipulative Techniques

3.2.1. LEVEL ONE: Characteristics of the Text as a Whole

All text are marked by category: *Opinion*, *Reporting* (objective informing on relevant issues), *Satire*, *Ignore* (if the language is non-understandable or the annotation is irrelevant), *Other*. In addition, at the level of the entire text, we suggest making notes from Table 1 if such characteristics are present in it. Examples of techniques are given in Appendix A.

Among the techniques that relate to the entire text, one can indicate *Irrelevant or uncheckable data*: such text either presents information that is not related to the topic under discussion and leads away from the core argument(s) or is constructed on the basis of unverifiable or unconfirmed information (for example, there are no indications of sources or the sources are vague). It can also be based on the author's conjectures and/or assumptions. This type is particularly prevalent in the discourses of the Russian Internet (Runet), especially on blogging platforms and social networks.

Obfuscation, or excessive obscurity, is another technique that characterizes the whole text and can be traced in the coronavirus discourses of the Runet. By obfuscation or excessive obscurity, we understand the deliberate complexity of the text, for example, deliberate scientificity or the use of incomprehensible words that different audiences can interpret differently.

Shock content is an appeal to human fears, to emphatically shocking or traumatic stories, to the information (either true or not necessarily) that blocks critical perception. This is a manipulative structure often used as the "core" of the coronavirus text.

3.2.2. LEVEL TWO: Characteristics of Text Fragments (Persuasion Techniques)

There are six big categories, as was proposed by Piskorski et al. (2023): *Justification*, *Simplification*, *Distraction*, *Call*, *Manipulative Wording*, *Attack on Reputation*. Each of them has subcategories presented in Table 2 (see Appendix A for examples).

As for manipulative characteristics that relate to separate fragments of texts, *Exaggeration or minimization* – an attempt to depict something in overly "dark" colors or, on the contrary, to pretend that what is being described has little meaning – can be mentioned. For example, "Coronavirus: is it a cover-up operation? With a regular flu, people are not prohibited from going to work or communicating... they are not".

Another "popular" manipulation technique found in many texts is *Appeal to authority*. We define it as a reference to authority, including appeals to an authority that is irrelevant to the topic being commented on or an expert without a confirmed current status.

The concept of *Appeal to values* refers to the inclusion of references to virtues and values, often accompanied by pretentious and loud statements, as well as mentions of various types of values that are not directly related to the topic under discussion.

3.3. Corpus Formation and Data Collection

The research is focused on coronavirus-related materials, which were selected based on the current agenda. Only texts with at least 1000 characters were selected to ensure comprehensive coverage of the studied techniques. The selection itself was based on the current development of the media

Meta-Label	Label
Justification	1) Appeal to authority (including references to an authority that is irrelevant to the topic being commented on, an expert without a confirmed current status)
	2) Appeal to popularity
	3) Appeal to values, including references to virtue
	4) Appeal to fear / prejudice
	5) Greenwashing: justification through appeal to green politics
	6) Bluewashing: justification through participation in international humanitarian initiatives (often as UN projects, but not only)
	7) Flag waving (appeal to national interests to justify ideas)
	8) Rumours
Manipulative Wording	1) Loaded language
	2) Sensational and/or provocative headings
	3) Repetition
	4) Exaggeration / Minimization
	5) Obfuscation, vagueness, obscurantism
	6) Statistical deception (shift of emphasis on one indicator, its exaggeration, vague wording according to the data obtained, confusion of significance in numbers with practical significance, etc.)
Distraction	1) Strawman
	2) Red Herring
	3) Whataboutism
	4) Appeal to Hypocrisy ('Tu quoque')
	5) Distraction by scapegoat (a combination of "strawman" and adhominem - to assign someone to blame in order to remove criticism from another person)
	6) Substitution of an idea / topic / issue
Call	1) Slogan
	2) Conversation Killer
	3) Appeal to Time
	4) "you should", "never do...", "you must..."
Simplification	1) Causal Simplification
	2) False Dilemma or no Choice
	3) Consequential Simplification
	4) Simplified Interpretation
	5) Stereotypes (an attempt to evoke negativity towards an alternative scenario, often based on prejudice)
	6) "I am like you" or "like everyone else" (an attempt to convince the target audience to join in and take a course of action because "everyone else is doing the same thing"): this is what all or "positive" nations / parties / groups do
Attack on Reputation	1) Labelling
	2) Hate speech, slang, name calling: demonization, offensive epithets, metaphors, names related to a particular phenomenon / organization / country / nation / person / idea, etc., are used to discredit something / someone
	3) such negative concepts as "authoritarianism", "aggression", "enemy", "imperialism", "terrorism", "militarism", "nationalism", "occupation", "racism", "totalitarianism", "junta" are exploited. And, on the contrary, such positive concepts as "brotherhood", "democracy", "friendship", "health", "quality", "love", "peace", "patriotism", "victory", "superiority", "prosperity", "equality", "freedom", "commonwealth", "happiness", "success", etc.
	4) Casting Doubt
	5) Guilt by Association (Reductio ad Hitlerum)

Table 2: Characteristics of text fragments.

space (Entman and Usher, 2018; Peeters and Maeseele, 2023): studying the processes of manipulation, today it is necessary to consider the participation of not only the so-called "rogue actors", who create manipulative content, but also the platforms that distribute it, as well as the audience that is influenced by such content when forming their own picture of the world or when interacts with similar information (Culloty and Suiter, 2020).

To address the suggestions that (*i*) not only the producers of inaccurate or manipulative content are important, but also the platforms on which they are distributed, (*ii*) not only content produced within the framework of the institution of professional journalism is important, but also by users, a corpus of texts was collected from two platforms:

- Vkontakte—one of the most popular social networks among Russian users, which was used as a source of user-generated content;
- Yandex.Zen—a blogging platform that was used as a source of content created by professional media, but which principles of content demonstration are based on user preferences and smart feed (as of 2022, almost 70 million people used Zen monthly, and the number of active bloggers was 100,000).

A total of 125 texts published on these two platforms between 2020 and 2022 about COVID-19 were selected and manually annotated.

Label	Count	Count (texts)
(WT) Opinion	107	107
(WT) Reporting	9	9
(WT) Satire	4	4
(WT) Other	5	5
(WTO) Irrelevant/uncheckable data	132.0	73.0
(WTO) Slogans, myths, stereotypes	62.0	48.0
(WTO) Emotional "load"	53.0	43.0
(WTO) Storytelling	51.0	48.0
(WTO) Promotion	46.0	32.0
(WTO) Shock content	38.0	26.0
(WTO) Mocking, trolling	36.0	18.0
(WTO) Obfuscation	23.0	18.0
(WTO) Allegories and metaphors	22.0	19.0
(WTO) Extremums	14.0	13.0
(WTO) Repetition	14.0	14.0
(WTO) Oversimplification	11.0	11.0

Table 3: The distribution of labels related to whole text (WT) and mainly to text but can also characterize specific phrases (WTO).

3.4. Data Labeling

The corpus, composed of 125 texts, was manually annotated using a special annotation tool, Doccano, which allowed the annotators (*i*) to select text with an accuracy of 1 character, which was important for training the algorithms to accurately identify manipulative fragments with a precision of one sign, (*ii*) to assign one or more values to the selected fragments, which was used in annotating the fragments that were parts of more than one manipulative technique without ranging the techniques and marking all of them. In the case of manipulative techniques that characterized the whole text, the title of the text was annotated with the corresponding label.

The task of annotating the texts was carried out by two annotators—the same researchers who developed the classification system. The Inter-Annotator Agreement calculated using Cohen's κ was greater than 0.8, which indicates a high level of agreement. It is also important to emphasize that after the initial annotation, both coders cross-checked the labels, removing all controversial issues and reaching consensus.

The texts were annotated in their original form without any modifications. This approach was taken as even small details, such as the number of spaces or punctuation marks, as they could potentially serve as indicators of inaccurate, manipulative, or propaganda content.

3.5. Dataset Statistics

A total of 125 documents were annotated, which contained 2396 manipulative segments. Table 3 contains statistics for each class related to the entire text, including the number of examples and the number of examples per text. In some cases, the WTO (*Whole Text Other*) tags refer to specific phrases and do not characterize the entire document. Therefore, their number is greater than the number of texts. At the same time, each document is assigned exactly one WT *Whole Text*) label.

Table 4 shows statistics for the labeled manipulative segments. Additionally, we indicate the average length in characters and words, as well as the number of spans that end with a period or punctuation mark (normalized by the number of examples in the class).

We also conducted an analysis to determine the number of spans of other classes that are present in the document when a specific class is present. In this calculation, we did not include the span for statistics on itself, so the values on the diagonal are not maximal. In Table 5 this data is organized by rows. Additionally, we normalized rows by class weights. To enhance clarity, the 12 largest classes were selected. Not the most popular classes have the most weight in each row, which indicates that there are connections between classes.

3.5.1. Comparison to Existing Datasets

We compared our dataset `Zen-Propaganda` to a closely related dataset from the SemEval 2020 Task 11 (Da San Martino et al., 2020). One key difference we observed is the length of the labeled manipulative spans. In our dataset, the maximum average length of these spans is 330 characters, while in the SemEval competition the maximum length was only 130 characters. Moreover, this disparity holds true not only for classes with large lengths, but also on average. For example, the minimum average length for classes in the top-20 is 70 characters in our dataset, compared to 25 characters in the SemEval dataset.

Additionally, we found differences in the size and distribution of the classes. The dataset in SemEval-2020 included 536 articles with around 9000 labeled spans. In `Zen-Propaganda` we have 125 articles with around 1900 labeled segments. Consequently, the average number of spans per article is roughly similar between the two datasets. However, the distribution of classes is different. In the SemEval dataset, about half of all labeled spans had the label *Loaded Language* or *Name calling or labeling*. In our dataset, these classes are also popular (in the top 5), but their proportion is much smaller. There are also two compatible classes: *Appeal to authority* and *Casting Doubt*. Note that the complexity of different classes varies, making it challenging to directly compare the results of models trained and evaluated on these datasets.

We also conducted a comparative analysis of our dataset with the multilingual SemEval-2023 dataset (Piskorski et al., 2023), which includes texts in the Russian language. SemEval-2023 dataset consists of a total of 221 articles in Russian labeled with 4684 propaganda spans. So, it is comparable to our dataset in terms of the number of propa-

Label	Count	Count (texts)	Length	Length (words)	Punct. ending	Dot ending
Hate speech, slang, name calling	193.0	56.0	103.35	14.05	0.41	0.32
Appeal to authority	186.0	57.0	160.77	22.05	0.39	0.28
Casting Doubt	167.0	65.0	215.27	31.31	0.78	0.77
Labelling	156.0	57.0	97.29	13.41	0.43	0.31
Loaded language	155.0	74.0	112.75	16.32	0.55	0.49
Causal Simplification	88.0	46.0	230.72	33.39	0.83	0.77
Appeal to Hypocrisy	87.0	42.0	212.46	29.99	0.71	0.63
Negative / Positive concepts	79.0	35.0	98.19	13.05	0.51	0.44
Appeal to fear / prejudice	71.0	43.0	238.04	34.2	0.63	0.58
Statistical deception	65.0	43.0	236.18	36.08	0.74	0.74
Appeal to values	53.0	23.0	206.02	29.55	0.74	0.74
Simplified Interpretation	49.0	34.0	321.02	45.84	0.67	0.61
Substitution of an idea	49.0	31.0	273.24	39.76	0.69	0.67
Exaggeration / Minimization	48.0	30.0	173.9	25.42	0.71	0.65
Distraction by scapegoat	47.0	29.0	252.81	34.96	0.57	0.51
Sensational and/or provocative headings	44.0	37.0	72.05	9.23	0.45	0.45
“you should”	44.0	32.0	103.25	15.07	0.73	0.66
Rumours	40.0	23.0	187.3	27.02	0.75	0.68
Strawman	33.0	18.0	204.82	30.21	0.82	0.79
False Dilemma	31.0	18.0	150.94	22.58	0.77	0.65
Flag waving	30.0	15.0	198.03	27.73	0.77	0.77
Obfuscation, vagueness, obscurantism	27.0	15.0	330.3	43.89	0.63	0.59
Repetition	20.0	13.0	230.35	33.25	0.85	0.85
Stereotypes	19.0	13.0	133.58	18.95	0.79	0.79
Appeal to popularity	17.0	12.0	100.82	13.82	0.41	0.18
Guilt by Association	15.0	12.0	163.33	22.93	0.47	0.47
Consequential Simplification	14.0	11.0	139.14	20.57	0.64	0.64
“I am like you”	13.0	11.0	131.38	21.46	0.77	0.77
Slogan	13.0	12.0	129.38	18.08	0.54	0.46
Appeal to Time	12.0	9.0	115.25	18.08	0.75	0.67
Red Herring	8.0	6.0	209.75	30.88	0.88	0.75
Conversation Killer	7.0	5.0	32.0	4.71	0.71	0.57
Whataboutism	7.0	6.0	214.29	31.57	0.71	0.57
Greenwashing	4.0	2.0	191.5	29.25	1.0	1.0
Bluewashing	2.0	2.0	64.5	7.5	0.5	0.0
Call	1.0	1.0	56.0	9.0	0.0	0.0

Table 4: The distribution of labels related to text fragments (persuation techniques). In addition to the length and number, the number of examples ending in punctuation marks is indicated.

ganda techniques per document. Similar to our dataset, the most popular *WT* class in SemEval-2023 was *Opinion*, but the second most prevalent class, *Reporting*, took up almost 30% of the dataset, whereas in our `Zen-Propaganda` dataset it has less than 10%. However, it is important to note that our dataset is sourced from completely different domain, specifically posts from Vkontakte and Yandex.Zen (social media), whereas SemEval dataset is sourced from news outlets. We believe that our classification of propaganda characteristics is applicable across various topics in social media. Although the distribution of classes may vary, the underlying classification patterns are expected to remain consistent.

4. Experimental Analysis

The exploratory analysis of the dataset was validated through the evaluation of the NLP models' performance. To this end, we divided the task into two subtasks, as was suggested in SemEval-2020 Task 11 (Da San Martino et al., 2020):

- span identification: binary classification of tokens to highlight segments of propaganda;
- technique classification: multi-class classification of selected segments.

To solve these subtasks, we employed the top-performing non-ensemble approach proposed in the SemEval-2020 competition for identifying English-language propaganda (Chernyavskiy et al., 2020). This approach is based on the Transformer-style model (RoBERTa). Additionally, we incorporated task-specific modifications to enhance the performance of the models for both subtasks.

4.1. Span Identification (SI)

To train the model, we converted span markup to token labeling with the BIO encoding scheme (*Begin*, *Inside*, *Outside*). So, we represented the spans as labeled tokens. Similarly, when applying the model and obtaining the final result, we performed an inverse transformation, which was implemented using the Spacy library.

Following recommendations for the English task (Jurkiewicz et al., 2020; Chernyavskiy et al., 2020), we additionally modified RoBERTa by incorporating a Conditional Random Field (CRF) layer (Lafferty et al., 2001). This modification allowed the model to learn class-level interactions in addition to its default token-level interactions. Furthermore, we specified manual restrictions on predicted labels based on the impossibility of constructing some label sequences. For example, the *Inside* label cannot be located between two *Outside* labels.

Class	Simpl. Int.	Causal Simpl.	Hate speech	Casting Doubt	Loaded lang.	Labelling	Neg./ Pos. con-cepts	Stat. dec.	Appeal to values	Appeal to fear / preju-dice	Appeal to Hypocrisy	Appeal to auth.
Simplified Interpretation	0.306	0.449	0.612	0.633	**0.735**	0.612	0.327	0.449	0.204	0.429	0.347	0.551
Causal Simplification	0.318	0.477	0.398	**0.636**	**0.636**	0.500	0.341	0.636	0.216	0.500	0.568	0.557
Hate speech, slang, name calling	0.363	0.368	0.710	0.580	0.699	**0.772**	0.539	0.404	0.368	0.430	0.482	0.585
Casting Doubt	0.299	0.503	0.503	0.611	**0.784**	0.689	0.389	0.539	0.240	0.539	0.521	0.551
Loaded language	0.394	0.374	0.555	**0.684**	0.523	0.581	0.394	0.439	0.265	0.432	0.381	0.503
Labelling	0.353	0.538	0.763	0.699	**0.776**	0.635	0.513	0.564	0.314	0.506	0.596	0.571
Negative / Positive concepts	0.215	0.405	0.658	0.620	0.608	**0.722**	0.557	0.392	0.380	0.494	0.418	0.544
Statistical deception	0.354	0.523	0.554	**0.846**	0.800	0.600	0.400	0.338	0.246	0.569	0.538	0.662
Appeal to values	0.264	0.415	0.679	0.528	**0.755**	0.642	0.623	0.415	0.566	0.453	0.340	0.642
Appeal to fear / prejudice	0.296	0.507	0.592	0.563	**0.662**	0.620	0.394	0.535	0.183	0.394	0.479	0.577
Appeal to Hypocrisy	0.253	0.552	0.609	**0.828**	0.759	0.713	0.402	0.621	0.218	0.563	0.517	0.529
Appeal to authority	0.452	0.441	0.634	0.661	0.618	0.586	0.403	0.500	0.253	0.473	0.425	**0.694**

Table 5: Co-occurrence of labeled segments in the corpus.

4.2. Technique Classification (TC)

We employed the RoBERTa model for the technique classification subtask. As it was proposed for English-language data, as an additional feature description of the span we used its length and the aggregated (averaged) representation of tokens from the last layer of the neural network before the classifying head.

In addition, we provided the relevant context for classification. This was achieved by connecting each span with its corresponding text fragment using a separator. This fragment was obtained by extending each span to the boundaries of the preceding and succeeding sentences.

4.3. Experimental Setup

As for the baseline, we adapted the approach proposed for English tasks and utilized the Russian-language BERT-style model. In our preliminary experiments we chose the strongest model RoBERTa [2] compared to its other versions and RuBERT [3] (we observed +8.5% of the f1 score compared RuBERT in the TC task for the best checkpoint). This model was pre-trained on TAIGA corpus, which includes literary texts, news, texts from social networks, scientific articles. We randomly splitted the training dataset into training and validation parts by articles, using 20% of them for validation.

Subtask SI The RoBERTa-CRF model was trained end-to-end for 25 epochs with early stopping on a validation sample (we saved the model every epoch and chose the best at the end of training), with a batch size of 4 and a training step of 1e-5. To achieve statistical stability, We chose the best of three runs.

Subtask TC The RoBERTa model was trained for 14 epochs with early stopping on a validation, with a training step of 1e-5, a batch size of 24, and 300 warm-up steps to improve initialization. We chose the best of three runs. Additionally, we tried to use model initialization from the previous subtask (SI), but this did not improved results of the final model.

4.4. Evaluation Measures

For the span identification subtask, we measured the F1-score by symbols. Additionally, we calculated the F1-score by sentences (as, generally, spans have a large length). A sentence was labeled as manipulative if it contained manipulative fragments.

For evaluation of the technique classification subtask, we used standard machine learning metrics to assess the quality of classification: accuracy score (proportion of correctly labeled examples) and micro-averaged F1-score (which also takes into account class imbalance).

[2] https://huggingface.co/blinoff/roberta-base-russian-v0

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

4.5. Experimental Results

4.5.1. Whole Text Classification

It is challenging to learn whole text labels using only the current markup, since there is not enough data even for the simplest methods based on word counts as the classes are strongly imbalanced (the majority of texts labeled as *Opinion*). In our experiments, we achieved an accuracy of 0.9 with a macro-averaged F1-score of 0.4 on the test sample for a Random Forest model with 10 trees on top of TF-IDF embeddings.

4.5.2. Whole-Text-Other Classification

To solve the subtask of Whole Text multi-label classification with WTO labels, we utilized a fully-connected network on top of TF-IDF embeddings and trained it for 10 epochs in a multilabel setup. The model primarily focused on learning the most popular classes, achieving a binary accuracy of 0.85 on the validation set. Attempts to incorporate weights inversely proportional to the absolute/root/logarithm of the class frequency did not yield improvement due to the limited number of examples for some classes (less than 10). Notably, the predicted classes consistently included *Irrelevant data / uncheckable data* and *Opinion* with high accuracy. Additionally, when considering the top 3 classes or setting a classification threshold of 0.3, the model also predicted *Emotional "load"* and *Storytelling* classes with a precision of 60%. A potentially more precise solution may involve the implementation of few-shot learning techniques employing the Russian GPT. Our initial attempt at zero-shot learning utilizing GPT-3.5-Turbo yielded inconsistent and frequently deviated results with the specified target classes.

4.5.3. Span Identification

On the test sample, we obtained an F1-score of 0.30 for the intersection of spans. CRF layer improved the score of the standard RoBERTa by 2%. The achieved results are lower than the metrics reported for a similar task in the English language using the best single-model solution (0.45). However, it can be explained by the peculiarities of the `ZenPropaganda`: we have another classes (and class definitions) and completely different types of texts, moreover spans in our dataset are longer. We further calculated scores at the sentence level to gain a more detailed understanding of the model's performance. It achieved F1-score of 0.43.

4.5.4. Technique Classification

Our approach achieves accuracy of 0.385 and F1-score of 0.173, which is much lower than values in the English SemEval-2020 dataset with the F1-score of 0.63 for the best single-model solution. The main reason for the low value of F1 metric is a class imbalance. In addition, we have a large number of classes in the dataset compared to its size (36). The F1-score for each class is presented in Table 6. We observed that the model tried to learn only popular classes. So we have zero F1-score for low-frequency classes. A possible solution here is to combine classes into meta-classes to remove these low-frequency ones. We are not comparing our scores with SemEval-2023 (Piskorski et al., 2023) because it operated with another subtasks.

We combined labels in meta-labels that are associated with our six big classes: *Justification*, *Simplification*, *Distraction*, *Call*, *Manipulative Wording*, *Attack on Reputation*. The resulting dataset is also unbalanced but does not contain extremely small classes of 1-2 instances. Table 7 shows the frequencies of each of the meta-classes in the corpus with the corresponding F1-score of models trained on the source and this new corpus. The source model achieves accuracy of 0.45 and F1-score of 0.42, which is closer to the result of a similar model

Label	F1
Hate speech,slang,name calling	0.441
Appeal to authority	0.764
Casting Doubt	0.442
Labelling	0.291
Loaded language	0.413
Causal Simplification	0.293
Appeal to Hypocrisy	0.356
Negative/Positive concepts	0.421
Statistical deception	0.417
Appeal to values	0.133
Substitution of an idea	0.200
Sensational and/or provocative headings	0.545
"you should"	0.667
Appeal to fear/prejudice	0.077
Total	0.173

Table 6: Technique classification. Model was trained for all 36 classes. Score for other classes are 0.

Label	Count	F1	F1 (meta)
Justification	403	0.436	0.509
Simplification	214	0.361	0.300
Distraction	231	0.274	0.328
Call	77	0.500	0.538
Manipulative Wording	359	0.390	0.340
Attack on Reputation	610	0.550	0.605
Total		0.42	0.44

Table 7: Meta-class technique classification. Results for models trained on original labels and metalabels (F1 meta).

for the English SemEval-2020 corpus. The model trained on meta-labels surpasses this result, reaching accuracy score of 0.48 and F1-score of 0.44.

5. Discussion

Despite the fact that the list of manipulative techniques was carefully compiled based on previous research, some of them were found to constantly overlap, while others were rarely detected or merely not detected in the corpus. This poses challenges for effectively training the models. In addition, the labeled domain has its own specifics, which is expressed not only in terms of distribution, but also in the length of the spans of individual classes.

Another observation was that authors often used the same two or three techniques throughout the text. A high number of the texts that used direct references to well-known conspiracy theories showed this omission both in our annotating scheme and in the theoretical sources we used. The introduction of conspiracy theories to the list of manipulative techniques could be done in further research on the topic. It should also be noted that a number of techniques that we initially attributed to the entire material, were actually well identified in specific phrases or individual fragments of the text.

All this indicates that, at the theoretical level, work on understanding current manipulative media tools should continue. The list of techniques and media effects that contribute to this form of manipulation is still open and deserves further consideration.

6. Conclusion and Future Work

We presented `ZenPropaganda`, a dataset for detecting manipulative techniques in Russian coronavirus-related texts. The dataset consists of 125 texts and almost 2400 expertly annotated manipulative segments. We conducted a comprehensive analysis by comparing our dataset with existing related datasets and assessed the performance of state-of-the-art approaches that have been proposed for them. Furthermore, we provided a discussion of our findings and results.

In future work, we intend to expand our investigation by experimenting with an extended list of manipulative techniques and exploring different grouping schemes for them. Additionally, we plan to label texts from other popular Russian domains.

Acknowledgments

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

Limitations

The dataset presented in this study comprises a collection of Russian-language texts that are limited to a specific topic and obtained from specific sources. Also, the existing state-of-the-art methods suggested for closely related datasets and tasks, which we have employed as baselines, demonstrate low performance. So, these methods cannot be effectively utilized for the complete automated detection of manipulative techniques.

Ethics and Broader Impact

We would like to point out that work on computational propaganda detection could potentially be misused by malicious actors, e.g., with the intention to restrict freedom of speech. However, it is important to emphasize that the primary objective of such research is to enhance media literacy by raising awareness about the intricate mechanisms of manipulation employed through diverse propaganda techniques. We believe that the benefit of this kind of research outweighs the potential drawbacks.

We would also like to warn that the use of large-scale Transformers requires a lot of computations and the use of GPUs/TPUs for training, which contributes to global warming. This is a bit less of an issue in our case, as we do not train such models from scratch; rather, we fine-tune them on relatively small datasets. Moreover, running on a CPU for inference, once the model is fine-tuned, is perfectly feasible, and CPUs contribute much less to global warming.

7. Bibliographical References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tatiana Anisimova, Svetlana Chubai, and Elena Gimpelson. 2021. Forms of manipulation in the discourse of social advertising. *SHS Web of Conferences*, 108:05001.

Daniel Arnaudo, Samantha Bradshaw, Hui Hui Ooi, Kaleigh Schwalbe, Vera Zakem, and Amanda Zink. 2021. Combating information manipulation: A playbook for elections and beyond.

Eugene Bagdasaryan and Vitaly Shmatikov. 2022. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 769–786. IEEE.

J Brennen, Felix Simon, Philip Howard, and Rasmus Nielsen. 2020. Types, sources, and claims of covid-19 misinformation.

Danielle Caled and Mário J Silva. 2022. Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. In *Journal of computational social science, volume 5*, pages 123–159.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

Eileen Culloty and Jane Suiter. 2020. *Disinformation and Manipulation in Digital Media: Information Pathologies*.

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

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Robert M Entman and Nikki Usher. 2018. Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation. *Journal of Communication*, 68(2):298–308.

Alexander Fedorov and Anastasia Levitskaya. 2020. Typology and mechanisms of media manipulation. *International Journal of Media and Information Literacy*, 5.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 763–772, New Orleans, Louisiana. Association for Computational Linguistics.

Arno Jewett. 1940. Detecting and analyzing propaganda. *The English Journal*, 29(2):105–115.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA.

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

Juan Ignacio Martin-Neira, Magdalena Trillo-Domínguez, and María Dolores Olvera-Lobo. 2023. Science journalism against disinformation: decalogue of good practices in the digital and transmedia environment. *Scientific Journal of Communication and Emerging Technologies*, 21(1).

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009, Held Online. INCOMA Ltd.

Maud Peeters and Pieter Maeseele. 2023. Ideological crystallization: rethinking the alternative-mainstream binary in times of populist politics. *Journalism*, page 146488492311611.

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

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

Aryan Sadeghi, Reza Alipour, Kamyar Taeb, Parimehr Morassafar, Nima Salemahim, and Ehsaneddin Asgari. 2023. SinaAI at SemEval-2023 task 3: A multilingual transformer language model-based approach for the detection of news genre, framing and persuasion techniques. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2168–2173, Toronto, Canada. Association for Computational Linguistics.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1082–1087, Online. Association for Computational Linguistics.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming.

Krzysztof Wach, Cong Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt. *Entrepreneurial Business and Economics Review*, 11(2):7–30.

Anthony Weston. 2000. *A Rulebook for Arguments*. Hackett Student Handbooks.

A. Examples of Manipulative Techniques

In this section, we show typical examples of text fragments that contain manipulative techniques.

Irrelevant or uncheckable data On April 15, 2020, one of the VKontakte users published the following text (with a link leading to another resource): “When did the flock decide anything? thought Soros to himself. That’s what the flock is for – to be driven in the direction the shepherd needs. Is it in vain that he sponsors the media of most countries, non-profit charitable organizations, the IOC, and the WHO? It’s time for them to serve the “global profiteer.” It was decided to start Operation “Coronavirus” with the countries that were most ‘fed up’ with the current politics of the Republican Party, namely China and Iran, that readily agreed, not wanting to endure Trump for another four years, and so the pandemic of 2020 began.”. Here and further in the text, data that does not have relevant confirmation and has not undergone the fact-checking procedure is presented as facts.

Obfuscation, or excessive obscurity A typical example (from the same VKontakte network): “If the nano-“vaccinated” survive the first doses of graphene oxide introduced into their blood, it breaks down in the body due to neutralizing antibodies responsible for its breakdown. Once the graphene and its toxicity disappear, so do our neutralizing antibodies to the substance, which also trigger our immune globulins. That’s why they force you to get a booster shot every 3 months, claiming that you no longer have immunity – to maintain the level of this toxic substance in the body.”

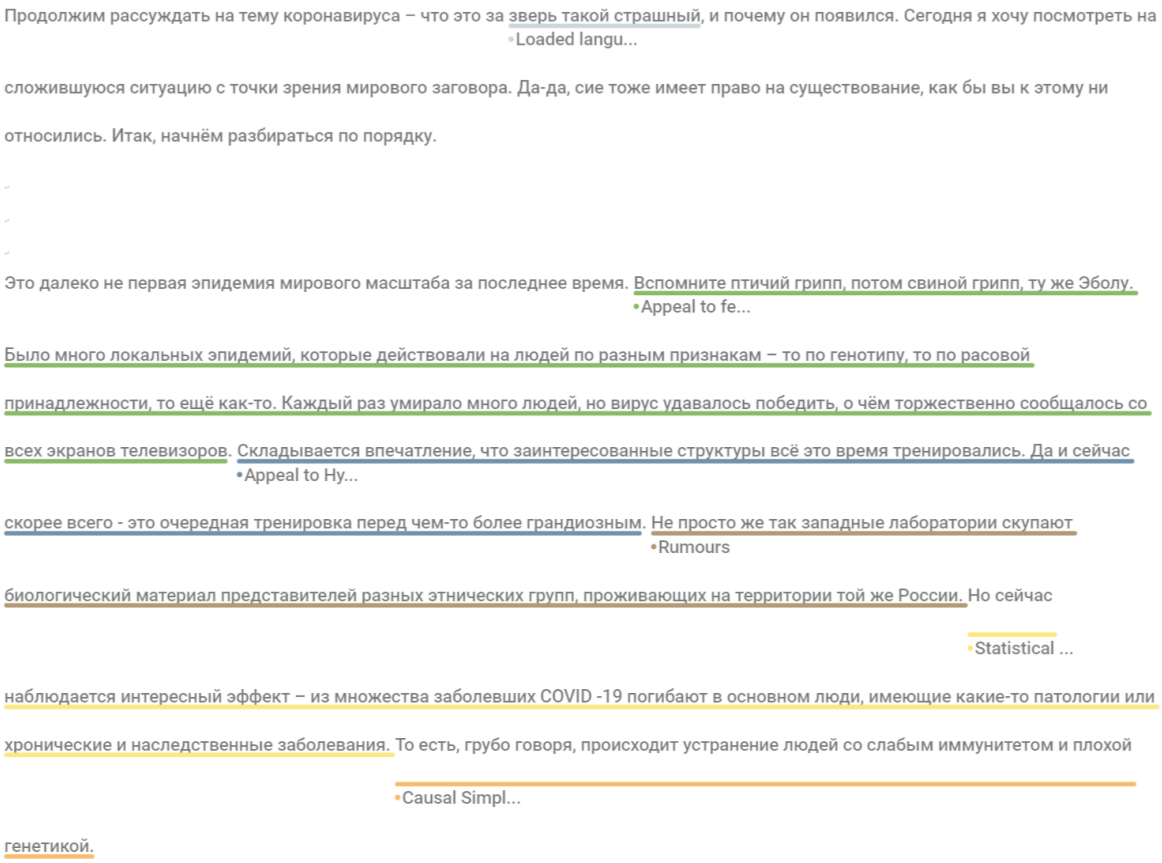

Figure 1: A screenshot of the labeled text from the annotation tool (Doccano).

Shock content Here is a fragment of one of the texts from the Yandex.Zen platform: “In China in January 2019, as many as seven engineers disappeared under strange circumstances, working with 5G. Their colleagues said that they were preparing a letter to the UN about the dangers of these technologies. And after the launch of 5G, the coronavirus suddenly appeared throughout China... It turns out that at 60 Hz, oxygen turns into a microwave oven. Everyone who gets into the area of a 5G tower will die from oxygen starvation. It is almost like how coronavirus patients are now suffering from oxygen deficiency. In the range of 40–50 Hz, high pressure immediately rises under the influence of an electromagnetic pulse on the neurons of the brain. It will feel like your skull is about to explode...”

Exaggeration or minimization An example of an exaggeration we can see in the following fragment from the text from Yandex.Zen “Coronavirus: is it a cover-up operation?”: “With a regular flu, people are not prohibited from going to work or communicating... they are not establishing (in fact) a curfew and a paramilitary situation, and they are not punished for evading admission pills and vaccines up to execution.”

Appeal to authority V. Zhirinovsky was a famous Russian politician, but he is hardly an expert in the field of virology. However, in one of the Yandex.Zen texts, we can see the following statement: “Vladimir Volfovich Zhirinovsky is an informed politician who knows a lot about the power behind the scenes. Behind extravagant behavior, a thinking person has long been hidden, as if in a case. Much of what he wrote about big-world politics took place at the dawn of fresh Russian capitalism. Zhirinovsky is not your primitive television demagogue-political scientist. Read what he tells us all about “virus policy.””

Appeal to values This technique can be seen in the following example from a text on Yandex.Zen: “The virus encourages people to pay attention to spiritual values and to understand that life can end at any moment. What will we take with us? Nothing but our conscience and moral and spiritual development. An excessive race for material values has led our planet to an ecological crisis... This disease is a reaction to our behavior, and it is not surprising that it began in China, the country that pollutes the atmosphere the most and where almost every living creature of the Earth is eaten.”

Screenshot from the annotation tool Figure 1 shows a screenshot of the labeled text from the annotation tool. It contains a lot of different manipulative techniques. Below, we provide a translation of this text (manipulative techniques are highlighted in brackets and in bold).

"Let's continue to talk about the coronavirus—what kind of [terrible beast][**Loaded language**] this is, and why it appeared. Today I want to look at the current situation from the point of view of a global conspiracy. Yes, yes, this also has a right to exist, no matter how you feel about it. So, let's start to figure it out in order. This is not the first global epidemic in recent times. [Remember bird flu, then swine flu, the same Ebola. There were many local epidemics that affected people according to different characteristics—sometimes by genotype, sometimes by race, or something else. Each time many people died, but the virus was defeated, which was solemnly reported from all TV screens.][**Appeal to fear/prejudice**] [It seems that the interested structures have been training all this time. And even now, most likely, this is just another training session before something more grandiose.][**Appeal to Hypocrisy**] [It's not just that Western laboratories buy biological material from representatives of different ethnic groups living on the territory of Russia.][**Rumours**] [But now an interesting effect is being observed—out of the many people who become ill with COVID-19, mostly people with some pathologies or chronic and hereditary diseases die.][**Statistical deception**] [That is, roughly speaking, there is an elimination of people with weak immunity and poor genetics.][**Causal simplification**]"