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

Moscow, June 15–18, 2022

The dataset for presuicidal signals detection in text and its analysis

Igor Buyanov
FRC CSC RAS
Moscow, Russia
buyanov.igor.o@yandex.ru

Ilya Sochenkov
FRC CSC RAS
Moscow, Russia
sochenkov@isa.ru

Abstract

The paper says about dataset for presuicidal signal detection in Russian posts from social media. To the best of our knowledge, it is a first dataset of a such type for this language. We develop a collection methodology and conduct linguistic analysis of completed dataset. We also build a classification baseline with machine learning models to solve the detection task.

Keywords: suicide; dataset; methodology; natural language processing
DOI: 10.28995/2075-7182-2022-21-81-98

Датасет для задачи распознавания пресуицидальных сигналов в тексте и его анализ

Буянов Игорь
ФИЦ ИУ РАН
Москва, Россия
buyanov.igor.o@yandex.ru

Соченков Илья
ФИЦ ИУ РАН
Москва, Россия
sochenkov@isa.ru

Аннотация

В статье представлен датасет для распознавания пресуицидальных сигналов в постах на русском языке в социальных сетях. Насколько нам известно, это первый датасет такого типа на русском языке. Мы разработали методологию сбора и провели лингвистический анализ полученного датасета. Мы также провели эксперименты по распознаванию пресуицидальных сигналов методами машинного обучения.

Ключевые слова: суицид, датасет, методология, обработка естественного языка

1 Introduction

Despite the rapid development of technologies that have significantly improved the living conditions of many people, the WHO reports 800,000 thousand annual suicides worldwide [13]. Suicide, in addition to tangible economic losses for the state, is hard to experience by the surrounding people and the action that can no longer be corrected. Judging by the fact that a recent study showed an increase in the level of depression [14], the problem of suicide will become more serious, as depression is considered to be one the suicidal factors [16] [17].

There are various non-profit organizations both in Russia and abroad that prevent such a terrible outcome by searching for potential suicides and carrying out preventive actions with them. The main source of search for such people is social networks such as Twitter, VK, Instagram, Telegram, etc., where people, mostly young ones, along with memes, sometimes post their experiences, even very frank.

Often, for people on the verge of suicide, expressing their feelings on a social network is a kind of valve that allows them to relieve the tension a little. In addition to direct expressions, a person who has decided to commit suicide sometimes leaves notes about his decision with information about the place and the chosen method. If such information could be detected immediately, it would be possible to save these people. In a less extreme case, it would be possible to track the individual problems like early depression or emerged physical self-harm before these problems severely damage individual mental health.

In recent years, a large number of papers have been published where the authors study the problems of detecting depressive behaviour based on data from social networks. Unfortunately, most of these works concentrate on English and concern the prediction of certain outcome like whether the person will commit a suicide in a predetermined time period. In this work to the best of our knowledge, we present a first Russian language dataset built from Twitter that is dedicated to a study of signals that people shows on their road to a possible decision of suicide.

As a result of the study, the following results were obtained:

- We collect a dataset, containing texts of messages in Russian from personal pages showing suicidal intentions or close to this condition. The dataset contains markup on the presence of features by which volunteers assess the condition of people.
- We discovered some language characteristics that are specific for people with a risk to commit a suicide at least on Twitter.
- We proposed the baseline implementations solving the aforementioned tasks of presuicide signal detection. The code and dataset are available.[1]

2 Related work

Applying the NLP techniques in the mental health domain is vastly possible with access to social media data. As a common source of data including post texts, the researchers utilize Reddit and Twitter. The former has a subcategory that is dedicated to a mental health problem so sometimes users directly report their diagnosis there which can be used to build a quality dataset. Almost the same happens on Twitter where users may post their diagnoses to find emotional support [8]. However, these posts had to be verified in order to be sure that the post contains no jokes, sarcasm and other unrelated phenomenon [18] [19].

This approach allows researchers to build a dataset for the identification of users having depression or PTSD [4] [18], a dataset with signs of depression [5] based on which the task of Early depression detection (eRisk) was organized, and a unique suicidal dataset [6] created from died and survived from committing suicide person's Twitter account. A list of currently available datasets for the mental health domain can be found in chapter 3.1 of a survey [7].

In this work [1] authors show that there is a statistical value between mental health and using Offensive Language. Again, the source of the data was Reddit.

Another dataset building method is to create a questionnaire application based on popular social network like Facebook. The users, who want to take a participation, give agreement under Terms of Services to collect their publically available data such status text, gender, age, etc. This approach was applied to study linguistic difference in user's personality [20].

Speaking of Russian language based works, it's worth to mention the paper, in which authors collect the depression posts from Vkontakte by utilizing a list of depression-related keywords and provide analysis of collected data [21]. In other work [22], authors managed to collect essay that was written on neutral topic by persons with a diagnosed depression. They provide analysis of dataset by showing the difference in a set of depression markers between depressive essays and control ones.

3 Task definition

The task set in the study is as follows. Having submitted the text to the input, the machine learning model should assign the text to one of five categories. During the paper we will refer to categories as next indices.

1. **Texts describing negative events that occurred with the subject in the past or in the present** — messages that are factual, describing negative moments that can happen to a person, such as attempts and facts of rape, problems with parents, the fact of being in a psychiatric hospital, facts of self-harm, etc.

[1] https://github.com/Astromis/research/tree/master/presuicidal_detection_dataset

2. **Current negative emotional state** — messages containing a display of subjective negative attitude towards oneself and others, including a desire to die, a feeling of pressure from the past, self-hatred, aggressiveness, rage directed at oneself or others.
3. **Messages about the intention of suicide** — messages containing an explicit declaration of suicidal actions. Messages that contain questions about suicide methods also fall into the same category.
4. **Messages with a suicidal theme** — the text of messages that are not directly related to the user but have a suicidal topic.
5. **Neutral** is the category in which messages that are not included in the above list fall.

Here we explain how we form these categories. In the course of the work of the non-profit organization, volunteers process accounts in social networks, in the post of which a third-party search engine found matches with keywords that carry a suicidal meaning. Processing consists of searching account posts containing signals about the possible presence of suicidal behaviour. Such signals can be indirect, such as, for example, stories about constant problems in the family or a university, and direct — the clear expression of a suicide intension. After evaluating the founded signals, the volunteer assigns to a particular user his suicidal status having three levels: low, medium, and critical (the highest level). The formulation of these categories based on volunteer's needs when they try to classify the user status.

The first category was formed from the considerations that negative events can leave an emotional trigger that can destabilize a person's psyche, increasing the likelihood of suicide if such thoughts arise. The more such triggers, the more vulnerable a person is. The second category is an indirect indicator of a person's mental state, which is also cumulative – if the density of messages with similar content increases, then the person becomes mentally unstable. The third category is self-explainable in a view of finding people with suicidal behaviour. Sometimes people don't expose direct emotions but uses death-related poetry or expressions. We can't include it in previous categories so we allocate a fourth one.

Notice that the second category is similar for the more general task of sentiment prediction where the task is to identify whether the text is either negative, neutral or positive, but in this work, we narrow the definition of text negativity. In our dataset, some texts also can be assessed as negative. It may be, for example, statements that a character from a game or TV series is annoying, but such negative texts do not carry meaningful information for our task.

4 The methodology of dataset creation

Using the collected database of annotated users, we download the texts of Twitter users' posts that had a medium and critical status. Further, all texts were annotated manually by several trained and guided non-psychologists annotators. At the time of writing, there were a critically small number of volunteers engaged directly in detecting users, and there was also no unified data collection software where volunteers could immediately mark messages with the necessary features. For this reason, outside people were hired and trained to annotate downloaded texts.

One of the benefits of the hired annotators is their personal responsibility that increase the quality compared to the crowdsourcing, and direct communication, that allows to give them a feedback on their work. Moreover, based on a feedback from the annotators, we constantly improve an instruction. The major drawback is the high cost, that didn't allow us to annotate the dataset with an overlap so we don't report inter-annotator agreement. We will remove this drawback in a future version of dataset as it is constantly improved.

We compile an instruction, which describes the categories, phenomena falling under the certain categories, as well as some general recommendations. The annotation was divided in several rounds. In each round the annotator receives data block consisting of 3-5 thousand texts, annotates it and sends it back. We manually verify 5% of each block and if the number of errors was no more than three cases per thousand examples, we accept the block and send the annotator a feedback.

Among the problems faced by the annotators, we can highlight attempts to interpret texts based on their own beliefs and personal experience, ambiguous meaning of some texts, texts containing complex phrasal expressions and sarcasm, and texts representing two classes.

To compensate an absence of inter-annotator agreement and ensure the quality of the dataset, after the annotation was finished, we apply a cleaning procedure using the TracIn [12] algorithm. Originally developed and tested in the field of computer vision, the algorithm can be adapted to any type of data, including text.

5 Analysis of the collected dataset

In this section, we provide some remarkable findings that we discovered during a dataset analysing process. First, Table 1 summarizes class distribution in a resulting dataset. We see that the neutral text is a majority class in a dataset despite the source of texts being persons with medium and high suicidal risk. From the perspective of our task, we, unfortunately, couldn't gather a comparable amount of texts that represent classes three and four so these categories will not be considered. However, the remaining categories also have a rather small number of examples compared to neutral texts, creating a strong imbalance of classes. This can be explained by the fact that the social network as a whole is not a "book of complaints" — people write there on various topics, including to distract themselves.

Class name	**Amount of examples**
Neutral text	27619
Current negative emotional state	2809
Texts describing negative events	2131
Messages with a suicidal theme	205
Messages about the intention of suicide	21

Table 1 – Class distribution of a collected dataset

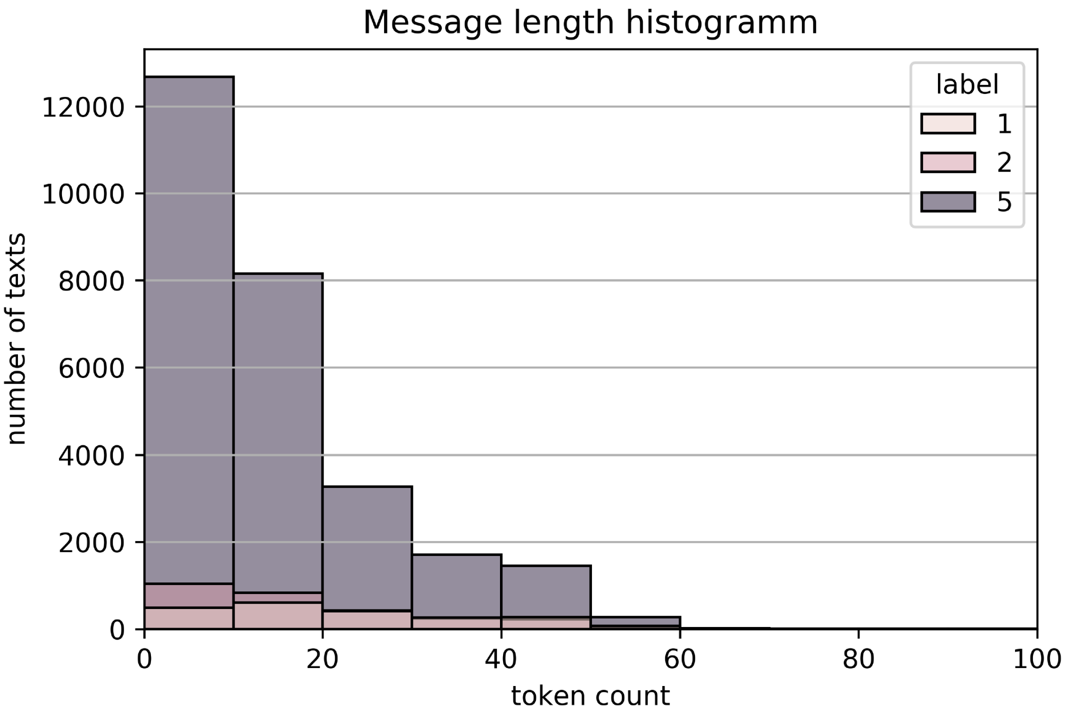

Figure 1 – Count distribution of texts by classes

Later, when we will build the baseline, we unite first and second classes in order to increase representativeness. To visualize how we transform classes for different purposes, we provide the Table 2.

Source set of categories	Set for dataset analysis	Set for baseline
{1,2,3,4,5}	{1,2,5}	1∪2,5

Table 2 – Dataset label transformation

In Figure 1 we can see the distribution of token length of text. As we can see, the distribution is convinced with our expectations that Twitter is a microbloging platform.

As part of texting, emojies have become an essential component in text communications. The main goal of emojis is to help better express person's feels, intentions sometimes even art. During the collection of the corpus, we preserve emojis in text and do a basic analysis.

In our dataset we got, 12551 emojis with over 483 unique set. We chose the top 10 emojis by frequency and build a count to class ratio distribution that is depicted in Figure 2. We excluded the Triangular red flag because it appears that about 1200 times of usage is distributed through 23 posts. We see that the Loudly crying face has the highest value for the second class which is consistent with our expectation of class semantics. On the other hand, we see that the Pleading face which we might expect to be also important for the second class has the highest value for the five class. We also see that heart-related emojis also lies in five class that also looks coherent.

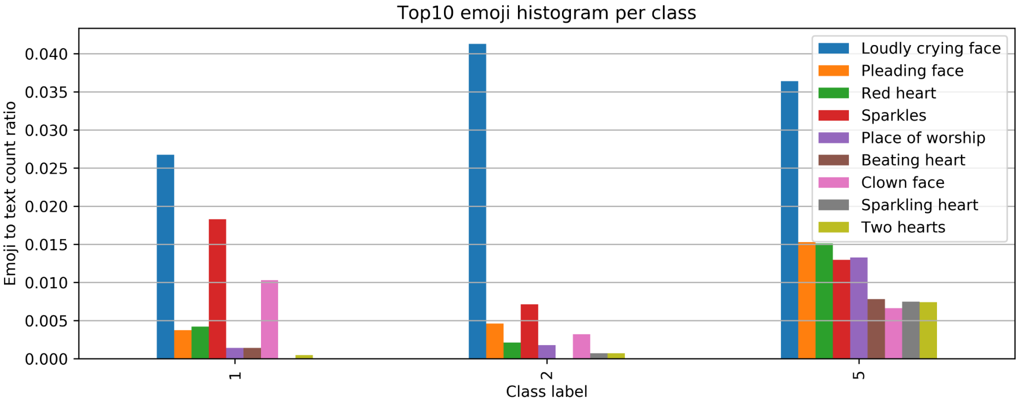

Figure 2 – Emoji to text count ratio distribution

We examine the lexicon of the dataset by a comparing method. To do this we took a more general twitter dataset that is used for sentiment analysis [9]. The first thing we investigated is unique words that characterize the language of the dataset. We acquire these words by substituting the set of general dataset words from a set of ours. We highlight the top 20 such words by frequency in Table 3.

Word	Count	Word	Count
мем	102	бсд	52
дазай	97	секси	50
геншин	82	ментальный	50
мью	81	пж	49
тикток	76	атсума	48
краш	74	рпп	48
рп	65	фд	46
хорни	63	осама	45
дилюк	57	эстетика	43
вайб	55	косплей	40

Table 3 – Specific words for our dataset

From this table, we can see special Twitter language like «мью» - transliterated short version of word "mutual" that means the person with which user has a mutual subscription with another one. We also can see some meme-words like «хорни» (transliteration from «horny» that means sexual arousal), shortcuts like "рп, пж, фд". We also see a name «геншин» which is a name of online videogame Genshin Impact and a Bungou stray dogs shortcut "бсд" which is a name for manga and anime TV show. There are also names of persons from these two universals.

Another method of lexicon analysis we applying is recently proposed allotaxonometry[10]. The goal behind this method is a comparison of any two systems, entity of which has a rank and this rank is distributed according to Zipf law. As part of that comparison, rank-divergence metric was proposed to understand the most important entities from two systems. Given two rank list R_1, R_2 of two systems with entity τ and hyperparameter α the rank-divergence metric can be computed as follow

$$D_\alpha^R(R_1||R_2) = \frac{1}{N_{1,2;\alpha}} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{\left[r_{\tau,1}\right]^\alpha} - \frac{1}{\left[r_{\tau,2}\right]^\alpha} \right|^{\frac{1}{\alpha+1}}$$

where $N_{1,2;\alpha}$ is a normalization factor (see the formula 7 from [10]). We again use the general sentiment analysis Twitter dataset as the opposite system. We construct an intersectional vocabulary from two corpora removing stop words and normalizing tokens. Then we compute the rank-divergence metric with $\alpha = 1/3$ as it was reported to deliver a good balance between entities with high and low ranks.

Table 4 shows the top 50 words sorted by Rank-Turbulence divergence and also shows the word rank in both corpuses that help two understand a direction of a rank change. From this table, we can clearly see that obsessive lexicon is a vital component of texts from our corpus. We might assume that this phenomenon relates to findings in work [1] in which authors show that there is a statistically significant relationship between mental health and offensive language usage.

index	word	rtd	rank in our corpus	rank in common one	index	word	rtd	rank in our corpus	rank in common one
0	бл**ь	3.070	7	175	25	обидно	1.241	868	131
1	завтра	2.429	70	8	26	день	1.240	8	4
2	блин	2.403	82	9	27	зачет	1.239	5573	334
3	на**й	1.968	44	745	28	заболевать	1.239	763	122
4	болеть	1.941	136	19	29	киев	1.227	12436	473
5	скучать	1.834	517	44	30	аниме	1.226	249	2786
6	сегодня	1.688	15	5	31	приходиться	1.218	387	83
7	человек	1.632	4	11	32	комп	1.217	1995	214
8	жаль	1.553	429	57	33	просто	1.215	5	10
9	жалко	1.553	1075	92	34	дома	1.209	232	60
10	пробка	1.516	3516	165	35	винд	1.208	25752	640
11	нг	1.463	2243	149	36	друг	1.208	27	76
12	тип	1.456	75	558	37	по**й	1.207	183	1435
13	школа	1.382	123	32	38	печально	1.194	5553	365
14	свой	1.375	6	14	39	заканчиваться	1.194	328	77
15	жизнь	1.360	16	46	40	снег	1.186	971	152
16	серия	1.346	556	87	41	ппц	1.181	9842	478
17	буквально	1.338	178	2093	42	выздоравливать	1.171	8477	460
18	печаль	1.338	2509	194	43	блиин	1.169	22574	672
19	нету	1.332	2371	191	44	выходной	1.163	505	106
20	чел	1.309	219	2915	45	обновлять	1.153	4337	354
21	пятница	1.305	1683	169	46	скоро	1.138	125	42
22	е**ть	1.291	111	740	47	ау	1.138	820	32308
23	пи***ц	1.254	41	144	48	личность	1.138	457	7089
24	суицид	1.251	615	31353	49	порез	1.121	800	26578

Table 4 – Rank-Turbulence divergence values for the top 50 words

In addition, we compute a thematic value characteristic (TVC) [11]. TVC represents the value of a word w for some particular topic σ compared to all other topics in a given corpus c. A TVC value ΔI^+can be computed as next

$$\Delta I(w, c, \sigma) = IDF(w, c \backslash \sigma) - IDF(w, \sigma)$$

$$\Delta I^+(w, c, \sigma) = \Delta I(w, c, \sigma) * X\big(\Delta I(w, c, \sigma)\big)$$

where IDF is inverse document frequency, X is a Heaviside function. Table 5 contains words with the highest TVC value for three classes.

Finally, we examine the common (pseudo-)syntactic patterns of the sentences by mining POS trigrams associated with a certain label. We use the Russian POS tagger from the NLTK package. Having got the POS tags we create trigrams and then compute the PMI score between each trigram and text label where a certain trigram occurs. In Table 6 a list of the top 10 trigrams[2] for three labels is presented.

[2] https://yandex.ru/dev/mystem/doc/grammemes-values.html

To explore the statistical significance of these findings we compute the Mann–Whitney U test statistics for these tag sets, with values being presented in Table 7. As we can see the difference between the tagset for 1 and 2 classes against 5 class and vice versa has statistical significance, although the difference between class 1 and 2 has not.

	Class 1	**Class 2**	**Class 5**		**Class 1**	**Class 2**	**Class 5**
0	прл	сато	мью	**10**	диагноз	прорыдать	дазай
1	антидепрессант	чудовище	чуять	**11**	селфхармить	упорно	петь
2	рвота	усталый	сезон	**12**	препарат	рас***рить	солнышко
3	больничка	медосмотр	вайб	**13**	желчь	здохнуть	фанфик
4	галлюцинация	поплакать	хорни	**14**	выстраивать	ничтожный	спи**ить
5	побочка	подпускать	добавлять	**15**	тревожка	пусто	геншин
6	бессонница	забиваться	картинка	**16**	трезвый	шататься	рт
7	частичка	унижение	читатель	**17**	биполярка	уе**сь	косплей
8	кп	подавлять	вкус	**18**	до**ывать	кулак	мило
9	порез	комок	ау	**19**	перечить	будовать	тикток

Table 5 – Top 20 words by TVC

Class 1	**Class 2**	**Class 5**
V|S-PRO|S	V|S-PRO|ADV	PR|V|S
CONJ|S-PRO|S-PRO	CONJ|V|S-PRO	CONJ|ADV|PR
CONJ|V|S-PRO	PR|A-PRO=m|S	A=pl|S|NONLEX
S-PRO|ADV-PRO|V	ADV|V|S	S-PRO|PR|A-PRO=pl
S|CONJ|PART	S|NONLEX|NONLEX	PRAEDIC|V|<none>
V|V|S-PRO	ADV|V|PR	CONJ|S-PRO|A=n
ADV-PRO|S-PRO|V	V|CONJ|S	A=n|CONJ|S-PRO
PART|PART|V	PRAEDIC|<none>|<none>	CONJ|CONJ|S
V|CONJ|PR	ADV|V|V	S|S-PRO|A=m
S|CONJ|PR	ADV|V|CONJ	S-PRO|S-PRO|CONJ

Table 6 – Postag trigrams for classes

	Class 1	**Class 2**	**Class 5**
1 class tagset	1	8.5e-1	1.8e-4
2 class tagset	7.0e-1	1	3.1e-3
5 class tagset	2.4e-3	2.4e-3	1

Table 7 – The Mann–Whitney U test results

6 The baseline classifier

In this work, we also provide a baseline for solving the established problem. In this section, we describe of a whole pipeline.

At first, we preprocess the dataset by removing all punctuation, set the case to lower, filter emojis and non-alphabetic characters. We also remove all text that contains only one token. As a vectorization procedure, we employ count vectorization which is essentially a vector with dimensionality equal to power of vocabulary and entities representing the number of times a certain word occurs in the given text. We also used BERT distilled model for the Russian language named rubert-tiny2[3]. As embeddings, a CLS token from the last hidden state were used. As we mentioned before, due to the lack of third and fourth categories, we exclude them from considerations. Moreover, as we mentioned before, we combine first two classes into one that makes it more representative. We assume that this new class will carry certain negatively gained emotions anyway so it might be distinguished from the neutral class.

			precision		**recall**		**f1**		**f1 macro**	
			mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
method	**vec**	**class**								
Isolation Forest	**BERT**	**0**	0.558	0.351	0.000	0.000	0.001	0.000	0.334	0.000
		1	0.500	0.000	1.000	0.000	0.667	0.000		
	Count	**0**	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.000
		1	0.500	0.000	1.000	0.000	0.667	0.000		
Local Outlier Factor	**BERT**	**0**	0.301	0.022	0.009	0.001	0.017	0.001	0.338	0.001
		1	0.497	0.001	0.979	0.003	0.659	0.001		
	Count	**0**	0.502	0.000	0.997	0.000	0.668	0.000	0.344	0.002
		1	0.768	0.033	0.011	0.002	0.021	0.004		
Logistic Regression	**BERT**	**0**	0.948	0.002	0.765	0.005	0.847	0.003	0.680	0.003
		1	0.382	0.004	**0.777**	0.011	0.512	0.004		
	Count	**0**	0.909	0.007	0.754	0.018	0.824	0.013	0.617	0.018
		1	0.313	0.021	0.598	0.030	0.410	0.024		
LogReg Stack	**BERT**	**0**	0.947	0.002	0.774	0.004	0.852	0.002	0.680	0.005
		1	0.391	0.010	0.770	0.006	0.518	0.009		
	Count	**0**	0.923	0.035	0.632	0.299	0.693	0.302	0.617	0.018
		1	0.292	0.078	0.656	0.190	0.384	0.065		
OneClassSVM	**BERT**	**0**	0.486	0.015	0.577	0.220	0.512	0.108	0.685	0.005
		1	0.491	0.013	0.401	0.202	0.414	0.130		
	Count	**0**	0.621	0.006	0.374	0.001	0.467	0.001	0.538	0.180
		1	0.552	0.002	0.772	0.007	**0.644**	0.003		
Random Forest	**BERT**	**0**	0.856	0.004	0.994	0.001	0.920	0.002	0.558	0.006
		1	**0.770**	0.032	0.112	0.007	0.196	0.010		
	Count	**0**	0.856	0.005	0.991	0.001	0.918	0.003	0.545	0.010
		1	0.671	0.028	0.098	0.012	0.171	0.019		
XGBoost	**BERT**	**0**	0.899	0.004	0.931	0.003	0.915	0.001	**0.703**	0.007
		1	0.548	0.018	0.445	0.016	0.491	0.013		
	Count	**0**	0.899	0.004	0.923	0.003	0.911	0.002	0.693	0.005
		1	0.516	0.016	0.442	0.009	0.476	0.009		

Table 8 – Experiment results

We experiment with several models including traditional classification methods like Random Forest, Logistic Regression, and XGBoost and models for outlier detection like Isolation Forest, Local outlier factor and, One class SVM. The motivation for the latter is a significant class imbalance, so we can view

[3] https://huggingface.co/cointegrated/rubert-tiny2

a neutral class as a class representing text that person types being mentally stable. We assume that a number of mentally stable people largely outnumber the number of unstable ones. Moreover, we assume that even high suicidal risk person's messages are not always exposed informative signals [15]. On the other hand, texts from combined classes can be treated as a non-common case. Finally, we use a composition of Logistic regressions called stacking. We split the dataset into equal blocks, where each block consists of a full number of outlier class and an equally sized normal class. On all blocks, we separately train the logistic regression model. After that, we train a final logistic regression model on a whole dataset using predicted probabilities from early train models as features.

We also estimate class weights and set them as hyperparameter in classification models. Other hyperparameters we left as default.

As a metric we use precision, recall and F1-measure. We evaluate all models ten times each time mixing random state of models and train/test split. Table 8 summarize the results.

From this table, we can see that Isolation Forest and Local Outlier Factor doesn't work in this setting as perfect recall with a half precision says that classificator assigns one class for all examples. Another observation is that in almost all settings the BERT embeddings as expected outperform the simple count vectorization method except OneClassSVM for detecting both classes. The best result by precision shows RandomForest based on BERT, although Random forest shows the worst result by recall. The best recall showed the Logistic Regression. The OneClassSVM with count vectorizer shows the best F1 score for the first class. It's interesting, that tree based ensemble methods show high recall for zero class. Unlike Random Forest, BERT based XGBoost classifier shows a much better result by recall for the first class that leads to the best macro F1 across all settings and, thus, setting our baseline for the task.

7 Conclusion and future work

In this work, we introduce a new task of detecting messages that express some clues about possible person mental instability. We develop a methodology for collecting a Russian dataset from open data from social media. We also analyse the dataset and found various language features that characterize such texts. Finally, we investigate various settings to build a baseline classifier. Overall we see that this task is quite challenging as the highest precision we can archive is only 0.75 on the classification task. On the other hand, we see that outlier detection method One Class SVM shows the best performance by the F1 score for a class of interest so it might be a promising way to continue to work with this task in outlier detection setting. Nevertheless, the best macro F1 shows the XGBosst classifier. Probably, with accurate hyper parameter search it is possible to archive better results.

In the future, we plan to collect more data which will include not only the text but also images, audio and social interactions. We believe that multimodality brings new findings and ideas to better understand the behaviour of people with high suicidal risk and thus give us more accurate methods to find and help them.

Acknowledgements

We would like to thank to our data annotators Ermakova Darina, Ulyanova Irina, Kalinovskaya Tatiana, our volunteer Brodskaya Alexandra for user seeking and psychologist Elizaveta Kluchikova for meaningful consultation. The reported study was funded by RFBR according to the research project № 21-011-44242.

Reference

[1] Ana-Maria Bucur, Marcos Zampieri, Liviu P. Dinu. An Exploratory Analysis of the Relation Between Offensive Language and Mental Health // Computing Research Repository. — 2021. — Vol. arXiv: 2105.14888. — version 2. Access mode: https://arxiv.org/abs/2105.14888

[2] Siyang Liu et al. Towards Emotional Support Dialog Systems // Computing Research Repository. — 2021. — Vol. arXiv: 2106.01144. — version 1. Access mode: https://arxiv.org/abs/2106.01144

[3] Ning Wang et al. Learning Models for Suicide Prediction from Social Media Posts // Computing Research Repository. — 2021. — Vol. arXiv: 2105.03315. — version 1. Access mode: https://arxiv.org/abs/2105.03315

[4] Glen Coppersmith et al. CLPsych 2015 Shared Task: Depression and PTSD on Twitter // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Denver, Colorado, 2015. — P. 31–39.

[5] Losada D.E., Crestani F., A Test Collection for Research on Depression and Language Use. — Springer, Cham, 2016. — Vol. 9822

[6] Sean MacAvaney et al. Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task // Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology. Online, 2021. — P. 70–80.

[7] Muskan Garg. Quantifying the Suicidal Tendency on Social Media: A Survey // Computing Research Repository. — 2021. — Vol. arXiv: 2110.03663. — version 1. Access mode: https://arxiv.org/abs/2110.03663

[8] Moreno, Megan A. et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depression and anxiety, 2011. — Vol. 28,6

[9] Rubtsova U. (2012), An automatic construction and analysis of short text corpus (microblog posts) for the task of developing and training of sentiment classifier. [Avtomaticheskoye postroenie i analiz korpusa korotkikh tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora], Knowledge engineering and semantic web technologies [Injeneria znanii i tekhnologii semanticheskogo weba], Saint Petersburg, pp. 109-116

[10] P. S. Dodds et al. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems // Computing Research Repository. — 2021. — Vol. arXiv: 2002.09770. — version 1. Access mode: https://arxiv.org/abs/2002.09770

[11] D. A. Devyatkin et al. (2013) Method of thematic clustering of large-scale collections of scientific and technical documents. [Metod tematicheskoy klasterizatsii mashtabnikh kollektsiy nauchno-tekhnicheskikh dokumentov], ITCS [ITiVS], Moscow, pp. 68-78

[12] Garima Pruthi et al. Estimating Training Data Influence by Tracing Gradient Descent // Computing Research Repository. — 2020. — Vol. arXiv: 2002.08484. — version 3. Access mode: https://arxiv.org/abs/2002.08484

[13] Dévora Kestel and Mark van Ommeren et al. Suicide in the world — World Health Organization, 2019. — Vol. 1

[14] Bollen J. et al. Historical language records reveal a surge of cognitive distortions in recent decades. — Proc Natl Acad Sci USA, 2021 — Vol. 1

[15] Cavazos-Rehg PA et al. A content analysis of depression-related Tweets. — Comput Human Behav, 2016 — Vol. 1

[16] Craig J. Bryan and M. David Rudd, Brief Cognitive-Behavioral Therapy for Suicide Prevention. — Guilford Press, 2018 — Vol. 1

[17] Popov U. V., A.A. Pichikov, Suicidal behavior in adolescents. [Suicidalnoe povedenie u podrostkov] — SpecLit, 2017 — Vol. 1

[18] Glen Coppersmith et al. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015 — P. 1–10

[19] De Choudhury M. et al. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media // Proceedings of the SIGCHI conference on human factors in computing systems, 2016 — P. 2098-2110

[20] H. Andrew Schwartz et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach — PloS one, 2013 — vol. 8

[21] Narynov S. et al. Dataset of depressive posts in Russian language collected from social media // Data in Brief, 2020 — vol. 29

[22] Stankevich M., Smirnov I. et al. Predicting Depression from Essays in Russian // Proceedings of "Computational Linguistics and Intellectual Technologies" DIALOGUE, 2019 — P. 637-647

Appendix A Инструкция для разметчиков

Общая информация

Вам будет представлен ряд сообщений, собранный с просторов русскоязычного Твиттера в формате Excel таблицы. Необходимо их распределить их в соответствии с ниже заданной классификацией. Разметка нужна для создания классификатора, который упростит волонтерам поиск людей, находящимся на грани самоубийства, для их последующего консультирования и оказания помощи. Без преувеличения можно сказать, что, выполняя эту работу, вы вносите вклад в спасение чьей-то жизни.
В случае если у вас возникнет вопрос как именно следует классифицировать сообщение, то необходимо такие сообщения сгруппировать отдельно, поставить ту метку, к которой склоняетесь больше всего, и описать ваши сомнения. Полученный файл прислать заказчику на супервизию.

Классификация

Жирным шрифтом выделены сами классы, а маркированным список указаны **не исчерпывающие** подобласти класса, которые призваны помочь составить представление о контенте. Критичность указывает приоритет класса при совместном появлении классов в одном сообщении.

1) *[Критичность: средняя]* **Исторические или текущие негативные события** – сообщения, носящие фактический характер, описывающие негативные моменты, которые могут произойти с человеком, такие как:

- попытки и факты изнасилования,
- проблемы с родителями (ненависть к ним, непонимание, алкоголики, насилие с их стороны),
- проблемы с друзьями/отношениями (отсутствие друзей, разрыв любовных и дружественных отношений, конфликты)
- издевательства в школе и травля,
- факты применения физической силы,
- факт нахождения в психиатрической больнице,
- психиатрический диагноз (депрессия("депра"), шизофрения, биполярное расстройство(биполярка), тревожность, СДВГ, ПТСР),
- факт употребления медикаментов (антидепрессантов, успокаивающих и т.д.)
- попытки в прошлом или фантазии о суициде,
- факт употребления наркотиков, алкоголя,
- проблемы со сном,
- проблемы со здоровьем
- факты самоповрежедний - текст, в котором говорится о том, что человек причиняет себе физическую боль. Чаще всего это выражается в порезах.
- проблемы с питанием - анорексия, заявления о том, что стошнит после приема пиши, невозможность нормально есть
- бедность (личная или семейная)
- проживание с больным родственником
- выраженная низкая самооценка
- пережитое недавно эмоциональное потрясение
- факты криминального характера

2) [*Критичность: низкая*] **Текущее негативное эмоциональное состояние –** сообщения, содержащие отображение субъективного негативного отношения к себе и окружающим:

- заявления о том, что нет сил, терпения,
- желание умереть,
- ощущение одиночества,

- ощущение, что "все плохо",
- рассуждения о тщетности жизни,
- ощущение давления прошлого
- фантазии и высказывания желаний о порезах
- ненависть к себе
- агрессивность, ярость, направленная на себя или на других

3) *[Критичность: высокая]* **Сообщения о намерении суицида**, отличается от "желания умереть" именно декларацией действий, например, "завтра в 7 иду на железную дорогу, всем спасибо за внимание", «завтра набираю ванну и беру нож» или поиск способов типа "какую веревку выбрать" или "смертельная доза таблеток", «насколько глубоко надо резать вены»

4) **Суицидальная тематика -** все то, что как-то связано с суицидом, но трудно в определении или не попадает в другие категории. Например,

- "Зачем нужны парные луки, татухи и всякая лабудень, если можно просто совершить парный суицид"
- "нежелаете совершить со мной суицид?"
- "соверши чистое, весёлое и энергичное самоубийство"

5) Сообщения, не имеющие отношения к суицидальной тематике

Антипаттерны

- Выражение эмоций, связанных с сохранением жизни, следует отнести к положительным примерам

Замечания

1. Сообщения на иностранном языке необходимо помечать, как не имеющие отношения к суициду
2. Тег <emoji></emoji> обозначает эмоджи, использованные автором. Хотя они довольно сильно могут мешать, они необходимы для изучения взаимосвязи использования эмоджи с состоянием людей. Просьба проявить терпение
3. В случае, если встречается текст, который попадает под две категории, то следует выбирать ту, которая более критична. Критичность категории указана в ее описании. Например, фраза: «Я БОЛЬШЕ ТАК НЕ МОГУ!!! Опять эти еб**ые флешбеки про изнасилование» несет в себе категории 1 и 2, но гораздо важнее то, что у человека есть факт переживания насильственных действий. Или же пример «Все, на**й все, вы все меня за**али! пошла на мост!», где пересекаются 2 и 3, но нам важно, что человек собрался совершить суицид.
4. При оценке сообщения недопустимо пытаться их интерпретировать, опираясь на личный опыт или знания – нельзя только лишь по твиту сказать действительно ли сообщение является попыткой привлечения к себе внимания, обманом или реальным положением дел. Посему, если текст попадает под одну из категорий, то необходимо его пометить этой категорией, вне зависимости от контекста.