Automatic Detection of Implicit Aggression in Russian Social Media Comments

Shulginov V.A.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
shulginov.val@yandex.ru

Mustafin R. Zh.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
rmustafin.art@gmail.com

Tillabaeva A.A.
Higher School of Economics /
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices
alinka99-t@mail.ru

Abstract

This article studies the characteristics of implicit and explicit types of aggression in the comments of a Russian social network with the means of machine learning. As it is hypothesized that expression of aggression depends on local norms, the dataset contains the comments collected from a single social media community. These comments were divided into three classes: polite communication, implicit aggression, and explicit aggression. Trying different combinations of data preprocessing, we discovered that lemmatization and replacement emojis with placeholders contribute to better results. We tested several models (Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, XGBoost, RuBERT) and compared their results. The study describes the misclassifications and compares the keywords of each class of comments. The results can be helpful while enhancing the algorithm of detection of implicit aggression

Keywords: politeness; verbal aggression; implicit aggression; machine learning; social media
DOI: 10.28995/2075-7182-2021-20-636-645

Автоматическое определение скрытой речевой агрессии в русскоязычных социальных сетях

Аннотация

В данной статье представлен принцип автоматического определения характеристик имплицитной и эксплицитной речевой агрессии в русскоязычных социальных сетях. Поскольку предполагается, что проявление агрессии зависит от локальных коммуникативных норм, датасет содержит комментарии, опубликованные в одном интернет-сообществе. Эти комментарии были разделены на три класса: кооперативная коммуникация, скрытая речевая агрессия и явная речевая агрессия. Используя различные комбинации признаков предобработки данных, мы обнаружили, что лемматизация и замена эмодзи на плейсхолдеры способствуют получению лучших результатов. Кроме того, мы протестировали несколько моделей машинного обучения (Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, XGBoost, RuBERT) и сравнили их результаты. В исследовании описываются ошибки классификации и сравниваются набор лексических маркеров для каждого класса комментариев. Полученные результаты могут быть полезны при усовершенствовании алгоритма обнаружения скрытой агрессии.

Ключевые слова: лингвистическая вежливость; речевая агрессия; скрытая агрессия; машинное обучение; социальные сети

1 Introduction

Aggressive behavior has a negative impact on participants of online communication [Warner & Hirschberg, 2012]. Social media such as Facebook and Twitter state in the usage policy [5], [12] that they are concerned about hateful user-generated content (abusive language, hate speech, cyberbullying, and trolling). Automation tools detecting aggression may be used to help moderators. For example, the Russian social network VKontakte introduced a new function available for the administrators and moderators of online communities: they will be able to filter the comments containing threats or hate speech.

It is important to note that the task of automatic detection of verbal aggression differs from the task of sentiment analysis [Cambria et al., 2017]; [Lukashevich, 2017]. It has not been studied properly because aggression is a complex sociocultural phenomenon. Verbal behaviour cannot be described by the dichotomy of aggression and politeness. It is more likely to be a continuum between two poles: completely rude interaction and polite respectful interaction [Locher, 2006]. In certain contexts, the words that are usually attributed to impolite behavior can be used either in cooperative or confrontational interaction. While detecting verbal aggression, ideally, we should consider both the intention of the speaker to conduct a face-threatening act and the hearer's perception of that. It might be traced if we consider the broad context of the message. However, even for human beings, it is difficult to determine the initial intention of the speaker and the internal state of the recipient, especially when the aggression is implicit and expressed with sarcastic, insincere politeness. There have already been attempts of automatic aggression detection in social media. The International Workshop on Semantic Evaluation SemEval-2019 motivated many researchers to study this topic. One of its tasks required identifying and categorizing offensive language in social media. Although our study is quite similar to that subtask, there are considerable differences. In addition to the basic differences in the language of texts and the source of data, a substantial difference in taxonomy must be mentioned. The works presented on SemEval-2019 included only two classes: not offensive and offensive. Offensive texts were characterized by the presence of obscene words. In contrast, we distinguish three classes of comments (polite, explicitly aggressive and implicitly aggressive) paying attention to the type of intention. Verbal aggression often contains vulgar lexis but it is not a definitive factor. This approach allows to detect the implicit aggression including sarcasm and irony which may not have vulgar lexis as their distinctive lexical features. Another workshop that should be considered is the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1). The taxonomy of TRAC included three classes just as the taxonomy of our study: overtly aggressive, covertly aggressive and non-aggressive texts [Aroyehun & Gelbukh, 2018]. However, the criteria of labelling were not described properly, and it was not clear what kind of texts were defined as covertly aggressive.

The aim of this paper is to study the features of implicit and explicit aggression using the means of machine learning. There are six tasks to be completed: to create a model detecting both explicit and implicit verbal aggression in Russian social media comments; to select and collect a corpus of comments; to preprocess the data obtained; to select the methods of data processing; to train the model; to test the model and to draw conclusions.

2 Methods

2.1 Data Collection

As stated above, we focus on social media comments. The data were collected in the social network Vkontakte, more specifically in the BORSCH (БОРЩ) community. The decision to select a particular community for data collection was motivated by the hypothesis that the way how aggression is expressed varies depending on the local norms and community standards.

Local norms are formed in communities of practice, which define the social engagements. Penelope Eckert states that "a community of practice is an aggregate of people who come together around mutual engagement in an endeavor" [Eckert, 2006]. The conventionalization of meaning is the result of the collective activity in which they share their experience. The three main criteria for identifying a community of practice, according to Wenger [1998], are (1) mutual engagement; (2) a joint enterprise; and (3) a shared repertoire. Thus, each community of practice has its own standards of aggressive communication.

The presumed consistency of the norm within a single community was supposed to contribute to better results. The total number of the comments collected was 28,272. Then a subset of the comments that were written in reply to another comment and containing more than three words was chosen. Such restrictions were meant to provide more detailed context helping the annotators label data more accurately, that was especially challenging in case of implicit aggression. The decision to select comments that were written in reply to other comments can be explained by the fact that annotators need the context to label them. The imbalanced dataset contains 7,225 comments in total: 5,058 comments in the training and 2,167 comments in the test dataset. The balanced dataset includes 5,687 comments in total: 3,984 comments in the training dataset (1,328 comments in each class) and 1,703 comments in the test dataset [https://github.com/alinatl/Implicit-Aggression].

2.2 Taxonomy and Labeling

The comments were divided into three classes of comments: polite (cooperative) comments, explicitly aggressive comments and the implicitly aggressive ones. Annotators marked up each comment using one of the three labels and considering the context of its use (previous and subsequent comment in the thread). Lexical markers and the strategy of each participant were taken into account:

0 - polite (cooperative) comments. It is the class of comments which correspond to the Grice's Cooperative Principle. According to Leech, politeness is "a constraint observed in human communicative behaviour, influencing us to avoid communicative discord or offence, and maintain communicative concord" [Grice, 2007]. These comments do not contain face-threatening acts towards another participant or any social groups.

1 - implicitly aggressive comments. The speaker performs a face-threatening act using politeness strategies which are clearly insincere. The speaker's intent is exhibited only in the context. There are the following key markers:

- vocatives that do not bear negative connotation by themselves (*старик — old man, сынок — boy, дружочек — little buddy, девушка — girl, оно — it*);
- markers of politeness and impoliteness intertwined (Х*орошая история. Жаль, что враньё. It's a good story. Too bad it's a lie*);
- question containing implicit aggression *(Ты глупый? Are you silly?*);
- offensive expressions exhibiting emotional state of the speaker (*Бля, как можно этого не видеть. Shit, how can you not see that?*);

2 - explicitly aggressive comments. In these comments, face-threatening acts are performed in a direct, clear, unambiguous and concise way in circumstances where face is not irrelevant or minimized. The comments contain different types of insults (personalized negative vocatives, personalised negative assertions, personalised negative references, personalised third-person references that are negative from the point of view of the target), name-calling, casting aspersions and pejorative speech.

Each comment was manually labelled by two annotators in order to minimize inaccuracy. When the labels did not match, the cases were discussed collectively and the disagreement was mitigated. Besides, the dataset was balanced, i.e. the number of comments in each class was the same.

2.3 Data Preprocessing

On purpose or unintentionally, people tend to change the graphic form of words in online communication. That is why one of the crucial tasks of preprocessing is to unify the data keeping the potential markers of verbal aggression.

The first stage of preprocessing includes tokenization [https://www.nltk.org/], casting all the words to lowercase and spelling correction with YandexSpeller [https://github.com/oriontvv/pyaspeller]. These operations were done in all variants of preprocessing because spelling mistakes can worsen the results. In the second stage, we tried all combinations of the methods related to six independent token-based features (lemmatization, emoji, punctuation, named entities, vulgar words, stopwords). All of them might be meaningful for aggression detection. During the actual experiments all possible combinations of preprocessing methods are tested in terms of the model score. There were 192 possible combinations in total. All possible characteristics can be seen in Table 1.

Lemmatization, punctuation removal, vulgar words removal and stopwords removal could be executed or not. On the one hand, participants of the community which was chosen for this study do not usually follow the rules of punctuation. In order to mitigate the inconsistency of punctuation in the corpus, punctuation marks can be removed. Stopwords, including prepositions, conjunctions and particles, can be also removed, because they do not contain any meaningful information, which is a common NLP practice. On the other hand, such fields of study as stylometry do not exclude stopwords nor punctuation marks when the method based on N-grams is applied. That is why all possible variants should be tried.

High variability of word forms in the Russian language makes it reasonable to apply lemmatization. Due to the typological features of Russian as a fusional language, one word can have many forms. In order to allow an algorithm to consider all the forms as one word, they are lemmatized.

Replacement of named entities and emojis with placeholders is applied when the dataset size is small and the number of distinct named entities, vulgar words and emojis is insufficient for statistical analysis. When these tokens are replaced, the information about particular named entities and emojis is lost. Nevertheless, this preprocessing method makes it possible to examine whether their presence constitutes a significant feature or not. Another possible variant is to remove all named-entities and emojis. Besides, emojis can be also replaced with specific classifying labels instead of placeholders: positive, negative or neutral.

All the parameters of preprocessing and the methods can be seen in Table 1.

Lemmatization		**Emoji**				**Punctuation**		**Named entities**			**Vulgar words**		**Stopwords**	
Yes	No	Keep	Remove	Replace with placeholders	Replace with labels	Keep	Remove	Keep	Remove	Replace with placeholders	Keep	Remove	Keep	Remove

Table 1. Parameters of data preprocessing

We used TF-IDF vectorizer because it is characterized by the adequate balance between high quality of vectorization and computational complexity.

2.4 Training Aggression Detection Models

Five models were trained to determine the baseline: Naive Bayes, Logistic Regression, Linear Classifiers with SGD Training, Random Forest, and XGBoost. The baseline model using simple algorithms should be surpassed by the final model. This baseline is established because there were no examples of studies with absolutely identical research design. We include the class of implicit aggression and focus on intention rather than on lexis while labelling.

The performance metric used was weighted average f1. Thus, 960 models were trained: 192 combinations of preprocessing techniques for each of 5 classifiers. The detailed explanation of how the best preprocessing type was selected is provided below.

We selected the top 20 models with the highest f1-score for each classifier. It allowed us to select the most stable preprocessing pipeline and model kind (i.e. classifier) showing the highest results.

No.	**logreg**	**xgboost**	**bayes**	**forest**	**sgd**
83	0.596	0.558	-	0.566	0.586
19	0.594	0.562	-	0.572	0.587

131	0.591	0.556	-		0.588
51	0.595	0.559	-	0.562	-
80	-	-	0.578	0.558	0.587
87	0.590	0.557	-	0562	-
81	0.592	0.559	-	0.561	-
21	0.590	0.556	-	0.565	-

Table 2. Types of preprocessing and the top-ranked results

Table 2 exhibits the types of preprocessing that were included in the top 20 results with at least 3 model types showing top-ranked results. According to it, the most successful variants of preprocessing pipeline for the majority of the model types were the variants 83 and 19. The methods applied in the five best types of preprocessing are provided in Table 3.

No.	**emojis**	**lemmatization**	**NER**	**punctuation**	**stopwords**	**vulgar**
21	replace	yes	no	keep	keep	del
51	del	yes	del	keep	keep	del
19	del	yes	no	keep	keep	del
83	del	yes	replace	keep	keep	del
87	label	yes	replace	keep	keep	del
81	no	yes	replace	keep	keep	del
131	del	no	del	del	keep	del
80	no	yes	replace	keep	keep	keep

Table 3. A comparison of preprocessing methods

7 out of 8 preprocessing variants included lemmatization, vulgar words deletion also appeared to be successful (7/8), in half of the cases the named entities were replaced by placeholders and in half of the cases emojis were deleted.

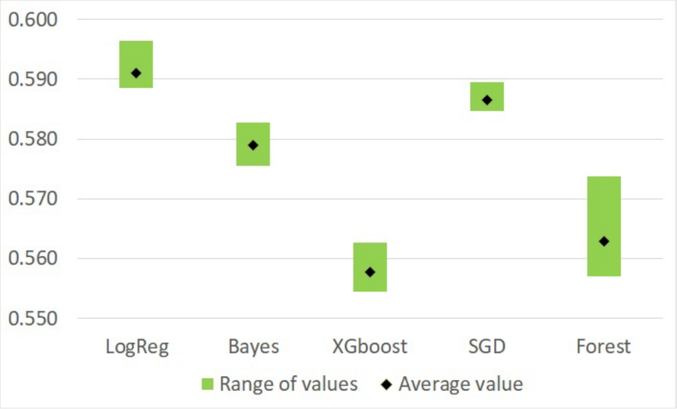

Figure 1. Value of f1-score metrics for all model types based on the results of 10 iterations

We calculated average values of the f1 score for models in top 20. From the table obtained we can state that Bayes classifier (0.579), Logistic Regression (0.591), and Linear Classifiers with SGD Training (0.587) demonstrated the best performance. Thus, Logistic Regression classifier combined with relevant preprocessing achieved the highest score and was defined as the baseline.

In order to outperform the baseline, we fine-tuned a transformer neural network based on the RuBERT model. This is a BERT (Bidirectional Encoder Representations From Transformers) trained on the set of Russian texts from the corresponding Wikipedia branch. That transformer language model achieves state-of-the-art results in a broad range of NLP tasks [Devlin J. and al., 2019]. For RuBERT model training we use both balanced and imbalanced datasets.

2.5 Keywords

When the models are ready, the following task is to define the lists of keywords for each class among the comments that the model labelled correctly, to compare the lists and to examine if there are regular patterns in terms of lexis.

We used three algorithms for selecting the keywords of each class: RAKE, Text Rank and Summa. RAKE (Rapid Automatic Keyword Extraction) calculates the weight of keywords (word scores) for words and phrases split by stopwords and punctuation marks. Tokens are presented as arrays and then are split into sequences of contiguous words at phrase delimiters and stop word positions. That is why this algorithm often selects collocations. We used the Python RAKE module [https://github.com/fabianvf/python-rake] with the following parameters: maxWords = 3, minFrequency = 2, whereas maxWords is the maximum number of the keywords, and minFrequency is the minimum keywords occurrence. TextRank and Summa are graph-based ranking algorithms that function as a voting and recommendation system that takes into account the relationships between words (vertices). As a result, we detected the keywords for all the classes. 10 keywords for each class are presented in Table 3 (0 - politeness, 1 - implicit impoliteness, 2 - explicit impoliteness).

RAKE			**TextRank**			**Summa**		
Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
крайняя мера last resort	общая генная цепочка common gene sequence	emoji	это it	это it	что what	всё everything	свой own	emoji
создать to create	emoji	своя мамаша own mom	что what	что если what if	emoji	emoji	твой yours	твой yours
случайность accident	достаточно самомнениее go enough	твой интернет your internet	все быть all be	emoji	твои your	весь entire	emoji	свой own
иметь to have	покупательская способность buying power	падаль старая old carrion	how	твои yours	own	year film	человек human	идти to go

Время time	религиозное чувство religious feeling	весь похуй all fuck	так so	свои own	идти to go	мочь понять человека to be able to understand a human	который which	который which
человек human	интеллект медузы jellyfish intelligence	мерзость filth	если они if they	как how	все all	просто just	мочь can	всё everything
хороший good	мочь повлиять to be able to affect	тупой Вася dumb Vasya	там there	они they	этот this	наш our	весь entire	тупой dumb
сильный strong	право твоё your right	свой own	можно can	так so	быть to be	хороший good	просто just	человек human
цена price	просто видеть just see	делать to do	только only	такой this	тупой dumb	большой large	хороший good	жопа ass
маска mask	признавать начало to recognize the beginning	жить to live	emoji	человек human	клоун clown	новый new	начать to start	ебать to fuck

Table 4. Keywords lists

Having analyzed the intersection of the words of three classes, we explored the main topics of the community “BORCH”. We also detected the keywords of each class. The words in the lists were categorized into several semantic groups. The results obtained are discussed in the next section.

3 Results

The baseline model was the model using Logistic Regression. The best variants of preprocessing were the variants number 19 and 83. They both included vulgar words and emojis deletion and lemmatization. In the 19th variant, named entities were not removed, while in the 83rd they were replaced with placeholders. Both in 19th and 83rd punctuation marks and stopwords were kept. This might indicate that users omit punctuation marks when using obscene vocabulary. The conclusion to be drawn is that the fact of the presence of named entities helps the model detect verbal aggression in social media comments.

№	Model	F1	Precision	Recall
19	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.59	0.59
	Random Forest	0.57	0.58	0.58

	XG Boost	0.56	0.57	0.57
83	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.58	0.59
	Random Forest	0.57	0.57	0.57
	XG Boost	0.56	0.58	0.57
80	Naive Bayes	0.58	0.58	0.58
	Log Reg	0.59	0.59	0.59
	SGD	0.58	0.58	0.58
	Random Forest	0.56	0.56	0.57
	XG Boost	0.54	0.56	0.56
RuBERT balanced		0.65	0.65	0.66
RuBERT imbalanced		0.66	0.66	0.67

Table 5. F1-score, precision and recall

The model based on RuBert demonstrated the highest score (0.66) on imbalanced dataset in automatic detection of aggression and overcame the baseline model Logistic Regression (0.60). At the same time, attention should be paid to the high accuracy of the explicit aggression detection (0.82).

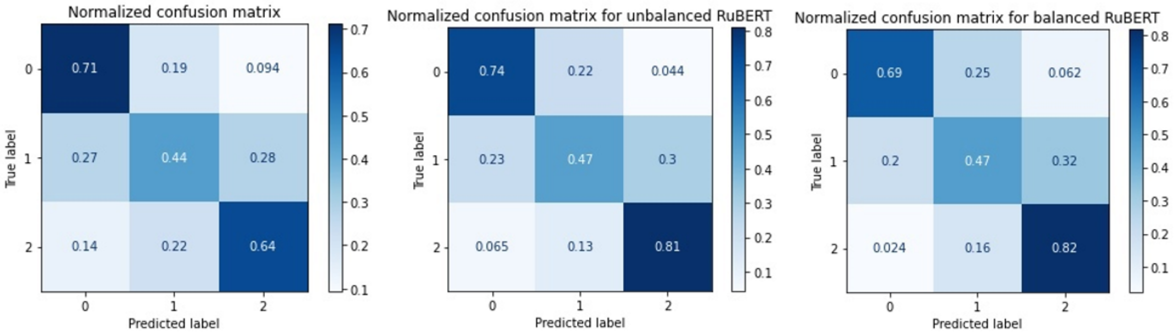

Figure 2.Confusion matrix of the Logistic Regression model with the 19th preprocessing variant/ the balanced RuBERT model/ the unbalanced RuBERT model

Having analysed the misclassifications of both models, we discovered that the model based on RuBERT attributed the comments of the class 1 correctly in 47% of the cases. It made a mistake and attributed the comments of that class to the class 0 in 36% of cases and to the class 2 in 17% of cases. The misclassifications between the classes 0 and 1 can be explained by the similarity in terms of their lexical features. It can be proved by the keywords representing each class. The keywords extracted for each class by all three algorithms (RAKE, Text Rank and Summa) are shown in Table 5.

Class 0	вода (water), значит (to mean), пользоваться (to use) место (place), кстати (by the way), говорить (to talk), время (time), маска (mask), бумажка (paper), классика (classics), вообще (at all), менять (to change), начать (to start), статистика (statistics), вариант (option), думать (to think), Россия (Russia), ситуация (situation), рубль (ruble), государство (state), просто (simply), посмотреть (to look), пример (example), остаться (stay), шина (tire), офигеть (be shocked), построить (to build), emoji, право (right), никто (nobody), курс (course), жить (to live),строить (to build)
Class 1	решить (to decide), месяц (month), ответ (answer), дело (case), жизнь (life), платить (to pay), значит (to mean), начало (beginning), заслужить (to deserve), никто (nobody), говорить (to speak), жопа (ass), точно (accurately), считать (to consider), хотеть (to want),

	видео (video), уровень (level), мочь (to be able to), мнение (opinion), перечитать (to reread), ждать (to wait), зарплата (salary), работать (to work), верить (to believe), человек (human), вообще (at all), сидеть (to sit), мама (mom), Россия (Russia), благодаря (due to), развивать (to develop), пора (it's time), невозможно (impossible), доказать (to prove), рубль (ruble), работа (job), весь (entire), видеть (to see), государство (state), молодец, давать (to give), Путин (Putin), пост (post), сказать (to say), посмотреть (to watch), Вася (Vasya). ясно (clear), жить (to live), показать (to show), таракан (cockroach), норма (norm), emoji, слово (word), покупать (to buy), понять (to understand), делать (to do), проблема (problem), понимать (to understand), продолжать (to continue), глаз (eye)
Class 2	параша (slop-pail), нормально (normal), дебил (moron), знать (to know), вместо (instead), шлюха (slut), говорить (to talk), жопа (ass), понятно (clear), хотеть (to want), читать (to read), ебать (to fuck), мразь (scum), мамкин (mom's), мочь (to be able), написать (to write), сосать (to suck), kremlebot, высер, говно (shit), ватник, лахта, пиздец, власть (authorities), почему (why), сказать (to say), смотреть (to watch), дурачок (fool), жить (to live), друг (friend), emoji, идти (to go), слово (word), делать (to do), работать (to work), понимать (to understand), сука (bitch)

Table 6. Keywords for each class

As demonstrated in Table 5, the class 2 contains the largest number of expressive lexis. There are several pejorative words marking political and ideological views (*(1) Kremlebot, (2) vatnik (3) lahta*), insults referred to promiscuity (*(4) slut, (5) to fuck, (6) to suck*), insults of family members (*(7) Mom's* (mamkin)) and scatological terms (*(8) shit*). The keywords of the comments with implicit aggression do not have offensive meaning by themselves and in many cases coincide with the words of class 0 which typify the main theme of the community (*(9) state, (10) right, (11) situation, (12) ruble, (13) mask*). It demonstrates that the lexical-based approach [Njagi et al, 2015] of aggression detection is not effective. Marked words in the class 1 can be used without addressee (*(14) ass*) or imply aggression only in particular contexts (*(15)Vasya* (this name is associated with a simpleton, a foolish person)*, (16) cockroach (*the Belarusian president's derogatory nickname)).

The keywords that are unique for the class of implicit aggression can contain substandard words (*(17) big head (bashka), (18) to get drunk, (19) to shit up*) or just name verbal aggression (*(20) boorish, (21) rudeness*) but they are not invective.

4 Conclusions and Future Work

The purpose of this study was to explore approaches to the automatic detection of implicit aggression in comparative perspective with the detection of explicitly aggressive and polite speech. The article discussed data collection, preprocessing and train modelling.

Several conclusions were drawn. First, we discovered that on the stage of data preparation lemmatization and keeping stopwords and punctuation marks contribute to better results. We also suppose that the comments with vulgar lexis do not contain punctuation marks more often. Second, certain similarities between polite communication and implicit aggression in terms of keywords make lexical features insufficient for accurate detection of implicit aggression. Third, the winning model was the model based on RuBERT. The f1 of this algorithm is 0.66, which is higher than the baseline and the best result presented for the similar task in TRAC-1 (f1 0.64). This result is still lower than the best result achieved by the participants of SemEval-2019 (f1 0.82), but it can be explained by a substantial difference in taxonomy. Our taxonomy includes not only polite and explicitly aggressive comments but also implicitly aggressive. This class is interjacent: it is at the same time closer to the explicitly aggressive class in terms of intention and to the polite class in terms of vocabulary. It allows to detect sarcasm and irony which do not bear any specific lexical markers. However, the absence of such markers worsen the results.

As for possible solutions to the problem of low accuracy, several solutions might be proposed. For instance, we could specify the taxonomy, analyse other linguistic features of implicit aggression (syntax, POS), consider pragmatics and identify interjacent classes between polite and impolite communication. The size of the dataset is, probably, not ample to ensure the stable work of the model.

The topic of automatic detection of implicit aggression in social media has many paths for further research. This study can be used as a base for the future research of implicit aggression as a linguistic phenomenon and automatic detection of aggression in communication. For instance, it is possible to use crowd-sourcing and to create a larger dataset, to collect a corpus of other languages or use other social media as a source. The methods also can vary: further studies might classify comments differently, consider other linguistic features or choose alternative ways of data processing.

References

[1] Aroyehun Segun T., Gelbukh Alexander. Aggression Detection in Social Media: Using Deep Neural Networks, Data Augmentation, and Pseudo Labeling // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC) 2018. — Santa Fe, USA, 2018. — P. 90–97.

[2] Cambria Erik, Poria Soujanya, Gelbukh Alexander, and Thelwall M. Sentiment analysis is a big suitcase. — IEEE Intelligent Systems, Vol. 32(6), pp. 74–80.

[3] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). — Minneapolis, USA, 2019.— P. 4171–4186.

[4] Eckert Penelope. Communities of Practice // Encyclopedia of Language and Linguistics / ed. by K. Brown. 2nd edition. Amsterdam: Elsevier, 2006, pp. 683–685

[5] Facebook's Policy & Usage Guidelines. Access mode: https://developers.facebook.com/docs/messenger-platform/policy/

[6] Grice Paul. Logic and conversation // Syntax and Semantics. Vol. 3: Speech Acts / ed. by P. Cole, J. Morgan. New York: Academic Press, 1975, pp. 41–58.

[7] Kecskes Istvan. Intercultural Pragmatics. — Oxford: Oxford University Press, 2014.

[8] Kumar Ritesh, Ojha Atul Kr, Malmasi Shervin, Zampieri Marcos. Benchmarking aggression identification in social media // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC) 2018. — Santa Fe, USA, 2018. — P. 1–11.

[9] Leech G. N. Politeness: Is there an East-West divide? // Journal of Politeness Research. 2007. Vol. 3 (2), pp.167–206.

[10] Levin Y.I. About obscene expressions of the Russian language [Ob obscennih virajeniyah russkogo yazika]. Selected Works. Poetics. Semiotics. [Izbrannie trudi. Poetika. Cemiotika], Moscow, 1998, pp. 809-819.

[11] Locher M. A. Polite behavior within relational work: The discursive approach to politeness // Multilingua. Journal of Cross-Cultural and Interlanguage Communication. 2006. Vol. 25 (3), pp. 249–267.

[12] Lukashevich N.V. Automatic sentiment analysis methods [Avtomaticheskie metody analiza tonal'nosti], Automatic natural language text processing and data analysis [Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh], NIU HSE, Moscow, pp. 127-179

[13] Twitter Rules and policies. Access mode: https://help.twitter.com/en/rules-and-policies

[14] Warner W., and Hirschberg Julia. Detecting hate speech on the world wide web // Proceedings of the Second Workshop on Language in Social Media 2012. — Stroudsburg, USA,2012. — P. 19–26.

[15] Wenger, E. Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press, 1998.

[16] Zampieri Marcos, Malmasi Shervin, Nakov Preslav, Rosenthal Sara, Farra Noura, Kumar Ritesh. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) // Proceedings of the 13th International Workshop on Semantic Evaluation. — Minneapolis, USA, 2019. — P. 75–86.