NATIONAL RESEARCH UNIVERSITY

Valery Shulginov
Alina Tillabaeva
Rashid Mustafin

Automatic Detection of Implicit Aggression in Russian Social Media Comments

Moscou, 2021

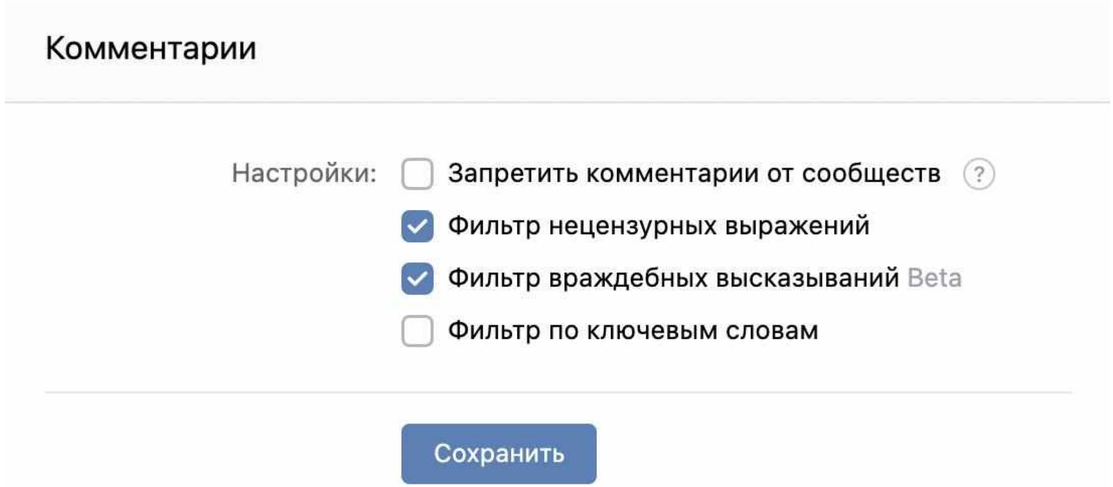
Plan

1. Introduction
2. Data collection
3. Taxonomy and labelling
4. Data preprocessing
5. Training aggression detection models
6. Keywords
7. Conclusions and future work

Introduction

1) Verbal behaviour is more likely to be a continuum between two poles: completely rude interaction and polite respectful interaction [Locher, 2006].

1) In certain contexts, the words that are usually attributed to impolite behavior can be used either in cooperative or confrontational in-teraction. 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.



Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

Tasks

- 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.

Taxonomy

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin

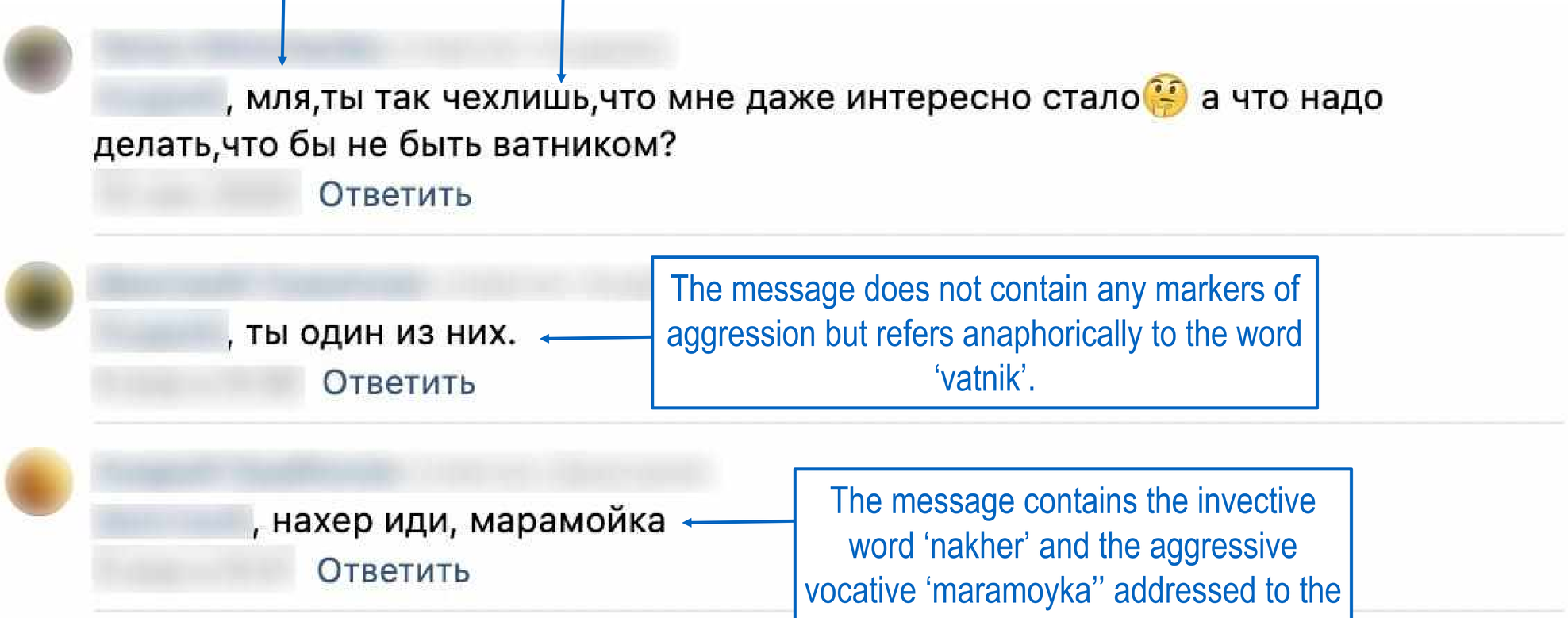
0 - Polite (cooperative) comments	1 - Implicitly aggressive comments	2 - Explicitly aggressive comments
These comments do not contain face-threatening acts towards another participant or any social groups.	• vocatives that do not bear negative connotation by themselves (*старик — old man, сынок — boy, дружочек — little buddy, девушка — girl, оно — it*); • markers of politeness and impoliteness intertwined (*Хорошая 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?*).	• 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, • pejorative speech.

Data Collection

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin

The obscene word is used as an interjection; aggression is not directed towards the interlocutor.

The word ‘chekhlish’ does not express aggression out of the context.



The message does not contain any markers of aggression but refers anaphorically to the word ‘vatnik’.

The message contains the invective word ‘nakher’ and the aggressive vocative ‘maramoyka’’ addressed to the interlocutor.

Data Collection

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin

Criteria of community selection:

— the total number of subscribers;
— speech aggression is a typical discourse;
— long threads in the comments;
— regularity of users' participation in the communication.

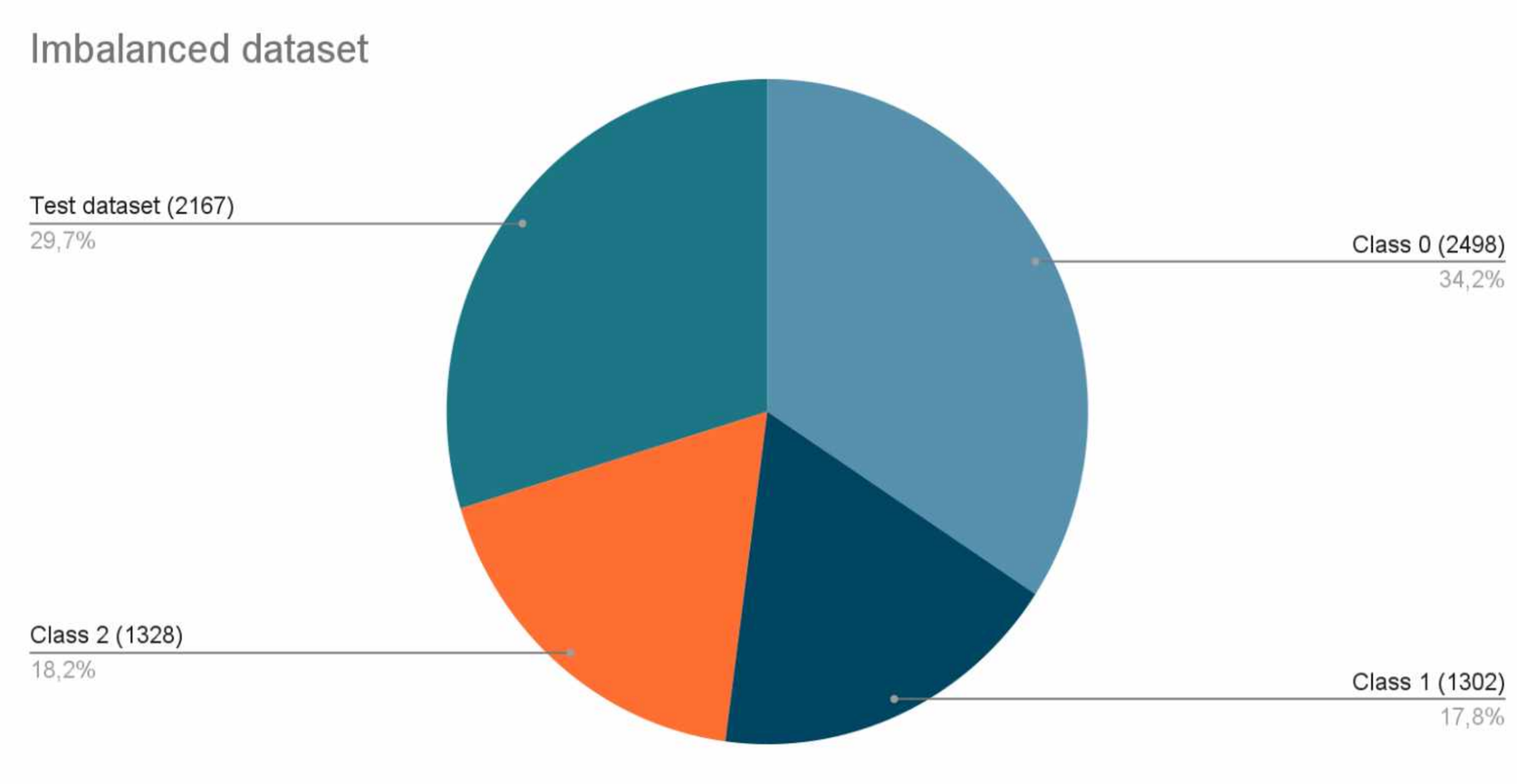
Criteria of comments selection:

— the comment written in response to another comment;
— the comment contains more than three words;
— the comment contains text (not only an image or an audio file).

Data Collection

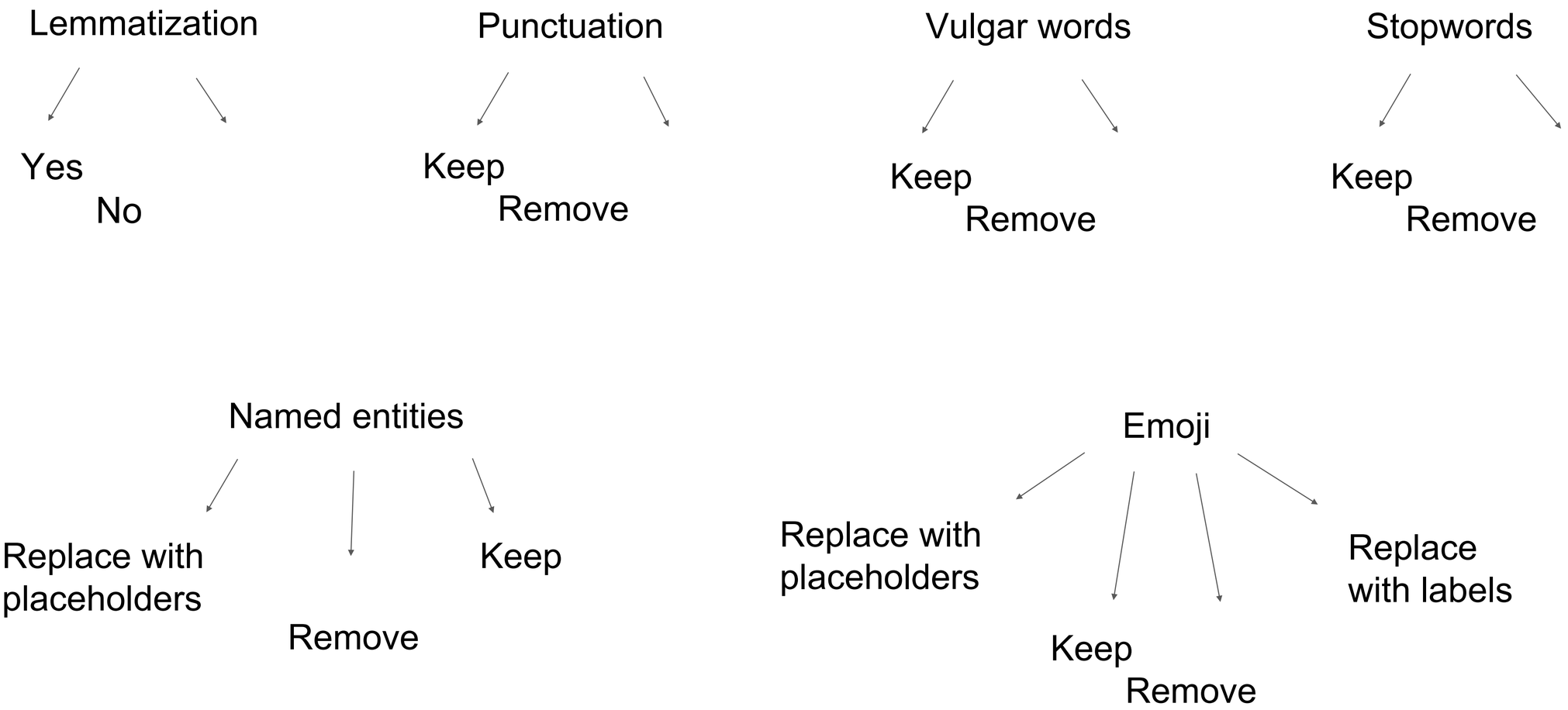
V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin

The total number of the data collected:
28,272 comments,
The imbalanced dataset: 5,486 comments,
The balanced dataset: 3,982 comments,
The test dataset: 1 ,703 comments.



Data Preprocessing

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin



Results

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin

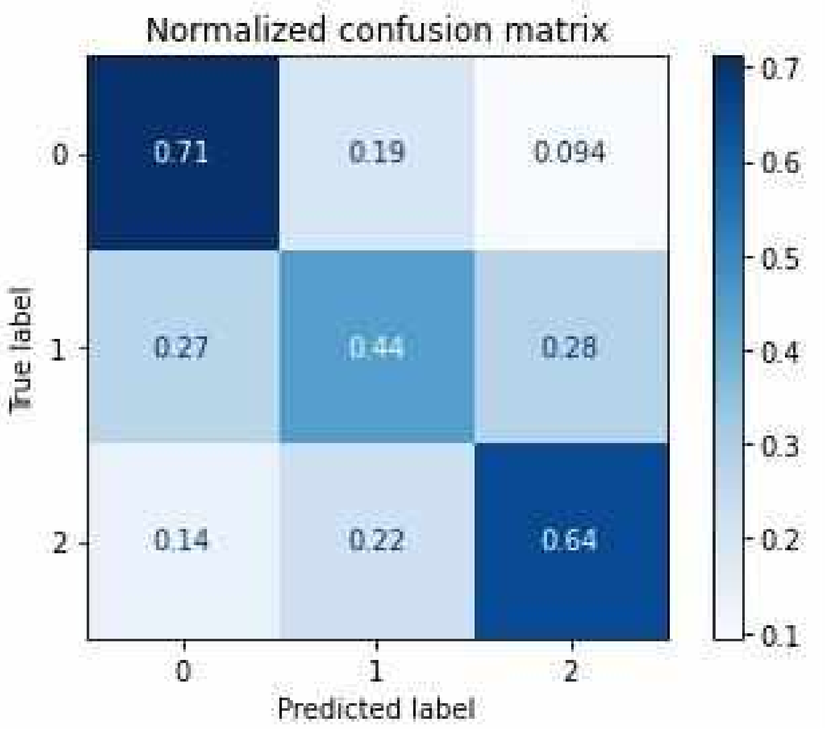
№	Model	F1	Precision	Recall
19	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.59	0.59
	Random Forest	0.57	0.58	0.58
	XG Boost	0.56	0.57	0.57
83	Naive Bayes	0.56	0.57	0.56
	Log Reg	0.60	0.60	0.60
	SGD	0.59	0.58	0.59
	Random Forest	0.57	0.57	0.57
	XG Boost	0.56	0.58	0.57
80	Naive Bayes	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	0.66	0.67
RuBERT imbalanced		0.66	0.66	0.67

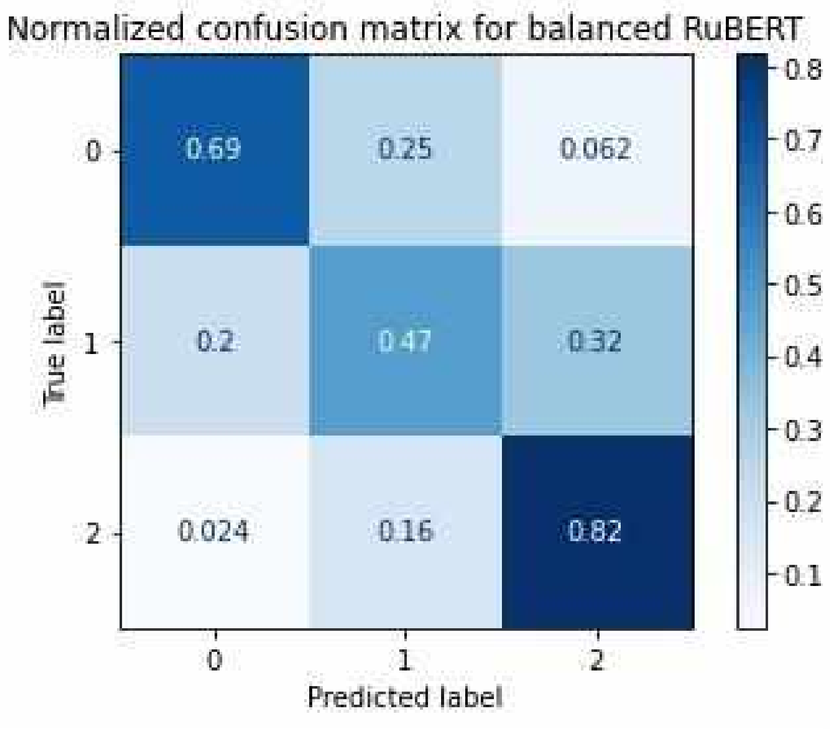
Preprocessing methods

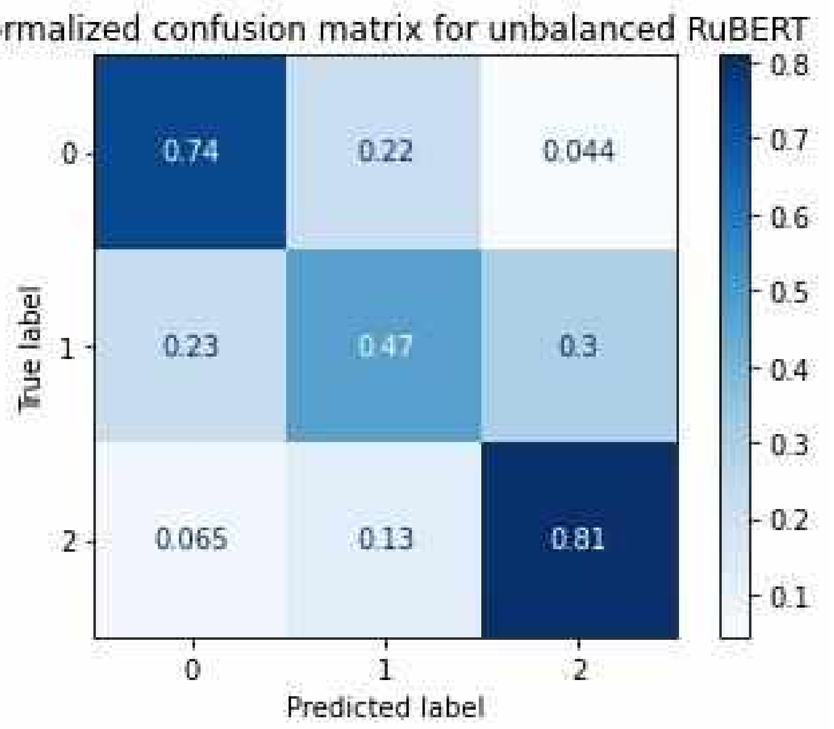
No.	emojis	lemmatization	NER	punctuation	stopwords	vulgar
19	del	yes	no	keep	keep	del
83	del	yes	replace	keep	keep	del
80	no	yes	replace	keep	keep	keep

Results

V.A. Shulginov, A.A.Tillabaeva, R.Zh.Mustafin







Keywords

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)

Conclusions and Future Work

- Lemmatization and keeping stopwords and punctuation marks contribute to better results.
- Similarities between polite communication and implicit aggression in terms of keywords make lexical features insufficient for accurate detection of implicit aggression.
- The f1 of the RuBERT model is 0.66.
- The principles of labelling are to be enhanced with a context-based approach.

Q&A

Contacts: Valery Shulginov shulginov.val@yandex.ru
Alina Tillabaeva alinka99-t@mail.ru
Rashid Mustafin rmustafin.art@gmail.com