HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

MASTER'S DISSERTATION

Classification of toxic and inappropriate texts on sensitive topics

Faculty of Computer Science: 01.04.02 «Applied Mathematics and Informatics»

Student

Kundyz Onlabek
Mathematics of Machine Learning
June 1, 2022

Research Advisor

Dmitry Ilvovsky
Associate Professor, PhD

Moscow 2022

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Классификация токсичных и неуместных текстов на деликатные темы

Факультет Компьютерных Наук: 01.04.02 «Прикладная Математика и Информатика»

Студент_____

Кундыз Онлабек

Математика Машинного Обучения

Июнь 1, 2022

Научный Руководитель_____

Дмитрий Ильвовский

Доцент, Кандидат Наук

Москва 2022

Annotation

In this work, the notion of toxicity is discussed, which has lately become a hot topic in Natural Language Processing. There have been numerous studies on identifying toxicity in textual communication as well as preventing it by understanding its premises. Toxicity can take many forms, ranging from derogatory remarks to activities that spread from the internet realm to the actual world, endangering people's well-being and even lives. Therefore, it is critical to prevent such abuse in textual form. To begin with, non-toxic online communication is more enjoyable and productive. Second, acceptance of online toxicity indicates that it is acceptable offline as well, so failing to prevent it early can have major dramatic effects.

While toxicity has attracted the attention of many scientists, it is not the only sort of unwanted content that can damage human-to-human talks. In this work, it is stated that toxic messages are a subset of a larger class of *inappropriate* communications. They contain:

- Messages that can offend an individual or a group of people by inflicting insult, threat, generalization, incorrect information, or disrespectfully raising sensitive concerns.
- Messages that can endanger a person's health if they follow the message's advice or logic.

Hence, inappropriate messages are those that can frustrate the reader directly or indirectly by offering incorrect or malicious information. Toxicity is one type of messages in this category, but inappropriateness includes more than just toxicity. A message that encourages drug use or suicide, for example, can be written in a non-toxic (and even supportive) manner, but its content is unquestionably harmful, implying that social media administrators would like to detect such messages.

Overall, the main aim of the research is to work primarily with inappropriate texts for the Russian language: investigate which machine learning models work better in the classification of inappropriate data and its detoxification. The research gap that is being closed by this work is that the existing models are created and trained to solve English language problems, while there is not much work done to deal with Russian language toxicity.

As a result, the inappropriate dataset was classified by two baseline models (logistic regression and BERT), and the main method, the F-1 score of the latter showed good results that gave a significant improvement over baselines. In addition, data was detoxified with the usage of the recently released RuGPT3 model, and the results were pretty average, as 40 % of the text was paraphrased clumsily or not changed at all. Further improvements to the model are necessary.

Аннотация

В этой работе обсуждается понятие токсичности, которое в последнее время стало популярной темой в области обработки естественного языка. Было проведено множество исследований по выявлению токсичности в текстовой коммуникации, а также по предотвращению ее путем понимания предпосылок. Токсичность может принимать разные формы, начиная от унизительных замечаний и заканчивая действиями, которые распространяются из интернета в реальный мир, подвергая опасности благополучие и даже жизни людей. Поэтому важно предотвращать подобные злоупотребления уже в текстовой форме. Начнем с того, что нетоксичное онлайн-общение более приятное и продуктивное. Во-вторых, принятие онлайн-токсичности указывает на то, что она приемлема и в реальной жизни, поэтому неспособность предотвратить ее на раннем этапе может иметь серьезные драматические последствия.

Хотя токсичность привлекла внимание многих ученых, это не единственный вид контента, который может повредить общению между людьми. В этой работе утверждается, что токсичные сообщения являются подмножеством более широкого класса *неуместных* сообщений. В них содержатся:

- Сообщения, которые могут оскорбить человека или группу людей путем нанесения оскорбления, угрозы, обобщения, неверной информации или неуважительного обсуждения деликатных вопросов.

- Сообщения, которые могут поставить под угрозу здоровье человека, если он следует совету или логике сообщения.

Следовательно, неуместными сообщениями являются те, которые могут прямо или косвенно навредить читателю, предлагая неверную или вредоносную информацию. Токсичность — один из типов сообщений в этой категории, но неуместность включает в себя больше, чем просто токсичность. Например, сообщение, поощряющее употребление наркотиков или самоубийство, может быть написано в нетоксичной (и даже поддерживающей) манере, но его содержание, несомненно, вредно, а это означает, что администраторы социальных сетей хотели бы обнаруживать такие сообщения.

В целом, основная цель исследования — работать с неуместным текстом на русском языке: выяснить, какие модели машинного обучения лучше работают при классификации неуместных текстов и их детоксикации. Пробел в исследованиях, который устраняется этой работой, заключается в том, что существующие модели создаются и обучаются для решения проблем английского языка, в то время как для борьбы с токсичностью русского языка проводится не так много работы.

В результате датасет неуместного текста был классифицирован двумя базовыми моделями (логистическая регрессия и BERT) и основным методом. Оценка F-1 основного метода показала хорошие результаты, которые дали значительное улучшение по сравнению с базовыми. Кроме того, данные были детоксицированы с использованием недавно выпущенной модели RuGPT3, но результаты были довольно средними, так как 40 % текста было перефразировано неуклюже или не изменено. Необходимы дальнейшие доработки модели.

Contents

Chapter 1

Introduction

Nowadays, toxicity is a big issue in online communication. It can take many forms, ranging from derogatory remarks to activities that spread from the internet realm to the actual world, endangering people's well-being and even lives. Therefore, it is critical to prevent such abuse in textual form. To begin with, non-toxic online communication is more enjoyable and productive. Second, acceptance of online toxicity indicates that it is acceptable offline as well, so failing to prevent it early can have major dramatic effects.

As a result, toxicity detection has lately become a hot topic in Natural Language Processing (NLP) [1, 2, 3, 4, 5]. There have been numerous studies on identifying toxicity in textual communication as well as preventing it by understanding its premises. Some of these methods have already been used in commercial companies. Instagram, for example, offers to reformulate toxic comments before posting them.

The fact that big language models, such as GPT-2 [6], are trained on large-scale user-generated data from the Internet is another less evident but severe hazard of toxicity spreading. If there is toxicity in the data, generation models can learn to make poisonous texts. The Microsoft Tay chatbot, for example, was shut down after it began sending unacceptable tweets after being fine-tuned with user data [7]. There have been many similar failures by dialogue assistants in various languages. Luda Lee, a Korean chatbot, made rude comments against sexual minorities and individuals with disabilities [8]. Yandex's chatbot Alice espoused radical ideas, endorsed violence, and encouraged suicide [9]. A customer was advised by Russian bank Tinkoff's Oleg chatbot to have her fingers amputated [10]. The training process of all these chatbots, like Tay, was done based on user interactions, tweets, and other user-generated texts. Toxicity should, hence, be identified not only to make human-to-human contact safer but also to prevent dangerous human prejudices from being transmitted to robots. When it comes to human-to-machine communication, the toxicity produced by a machine might harm a company's reputation. This provides yet another reason for corporate research groups to focus on this issue.

While toxicity has attracted the attention of many scientists, it is not the only sort of unwanted content that can damage human-to-human talks and affect language model training data.

Babakov et al. [11] propose that toxic messages, which have been studied extensively in NLP, are a subset of a larger class of *inappropriate* communications.

Some prior works [12] have already looked into the concept of inappropriateness. Yet, Babakov et al. [11] define inappropriate messages as those that could be damaging to any of the conversation's participants. They contain:

- Messages that can offend an individual or a group of people by inflicting insult, threat, generalization, incorrect information, or disrespectfully raising sensitive concerns.
- Messages that can endanger a person's health if they follow the message's advice or logic.

Hence, inappropriate messages are those that can frustrate the reader directly or indirectly by offering incorrect or malicious information. Toxicity is one type of messages in this category, but inappropriateness includes more than just toxicity. A message that encourages drug use or suicide, for example, can be written in a non-toxic (and even supportive) manner, but its content is unquestionably harmful, implying that social media administrators would like to detect such messages.

When applied to a firm or organization rather than an individual, the concept of appropriateness becomes considerably stricter. This requires the necessity to manage the messages generated by corporate chatbots. Any doubtful claim made by such a chatbot can increase users' displeasure and result in significant reputational damage.

Overall, the main aim of the research is to work primarily with inappropriate texts for the Russian language: investigate which machine learning models work better in the classification of inappropriate data and its detoxification. The research gap that is being closed by this work is that the existing models are created and trained to solve English language problems, while there is not much work done to deal with Russian language toxicity.

Chapter 2

Literature Review

2.1 Toxicity

Regarding the definition of toxicity, there are many English text corpora that have been annotated for toxicity's appearance or absence. Some works even identify the degree of toxicity and the topic of the toxicity. Yet, the scientific community could not agree on a definition of the term "toxicity", therefore, each study works with a variety of interpretations. Some resources refer to any unwelcome behavior as toxicity and don't differentiate between the two [13]. The preponderance of researchers, on the other hand, prefers finer-grained annotation. Jigsaw's Wikipedia Toxic comment datasets [1, 2, 3] are the most extensive English toxicity datasets accessible, allowing to work with a variety of toxicity categories (*toxic, insult, obscene, identity hate, threat*, etc). *Personal assault, aggression*, and *toxicity* are examples of granularity used in English Wikipedia talk pages based on WAC corpora [14]. In the Evalita 2018 Task on Automatic Misogyny Identification [15], a different method to toxic behavior classification was used: *misogynous, discredit, sexual harassment, stereotype, dominance*, and *derailing*. Some datasets focus primarily on a single category of toxicity, such as *offensive language* [16], *hate speech* [17, 18, 19, 20], and *microaggression* [21].

Toxicity varies along several axes. Some studies focus only on severe offenses such as *hate speech* [4], while others look into more subtle assaults [5]. Offenses might be directed at a single person, a group or the general public [22], and they can be explicit or implicit [23, 24].

The toxicity issue has also been researched in the Russian language. There are publicly accessible datasets of toxic comments in Russian: the Russian language toxic comments dataset (binary) [25] and Toxic Russian comments dataset (*normal, insult, threat, obscenity*) [26]. Also, the VK (Russian social network) dataset of 100,000 comments, that were analyzed to determine if they have different types of *hate speech/abuse* [27].

Considering setups for toxicity annotation, toxicity is commonly expressed as a binary (*toxic/safe*) or fine-grained (*obscenity / insult / threat /...*) sentence-level label in annotated data.

However, the concept of toxicity is ambiguous, and people frequently debate whether or not toxicity exists in a specific example. As a result, toxicity annotation is vulnerable to biases, as demonstrated by Waseem and Hovy [28], who found that crowd workers regard misogyny to be less harmful than other types of toxicity.

There are several methods for obtaining more objective annotation. A simple way is to collect numerous judgments on each sample and take the majority label as the ground truth, or to calculate the toxicity score as the percentage of ”toxic” labels, which was implemented in Jigsaw datasets [2, 3]. Since all individuals engaging in annotation may have a similar background and reflect the same biases in the data, this method is still prone to biases, yet, it provides pretty good judgements [5].

This bias can be reduced if the annotation is carried out by a wide number of people from various backgrounds. This can be accomplished through crowdsourcing, which involves enlisting the help of hundreds of people to complete a task. Crowdsourcing, on the other hand, suffers from another problem in manual annotation: the presumption that all users are equally trustworthy. This is not the case in a crowdsourcing scenario, when the number of workers is large and the ability to verify their dependability and comprehension of the task is restricted. In such scenarios, more advanced label aggregation methods exist, such as the Dawid-Skene approach [29]. This is an iterative algorithm that dynamically defines more trustworthy annotators as those whose annotations match the annotations of other users the most often. These aggregation algorithms, on the other hand, cannot combat user bias; they can only improve the distillation of the ground-truth answers. Aggregation cannot minimize ambiguity when the ground truth is intrinsically ambiguous, as it is with toxicity.

If binary labels are replaced with a fine-grained scale, such as 1-to-5 or 1-to-100, the debate over the presence of toxicity can be partially resolved. However, this scale still does not eliminate user prejudice and may cause annotators to become confused [30].

Dinan et al. [31] propose more sophisticated labeling setup to address the discrepancy and to expand the data with non-trivial cases of toxicity. They create a pipeline where crowd workers are told to ”fool” a pre-trained model. They attempt to create offensive adversarial messages that are misclassified as non-offensive. In a subsequent paper [32], the authors extend this pipeline by proposing a human-in-the-loop framework. Crowd workers there use an adversarial approach to detect harmful messages from the dialogue model. The utterances that pass the adversarial challenge are incorporated into the final datasets in each of these pipelines, making the systems trained

on them more robust to different kinds of toxicity.

Aside from direct labeling, there is an indirect method of generating a ranking of objects that is similar to the one obtained by fine-grained annotation. Instead of assessing the toxicity of sentences directly, an annotator could be asked to evaluate two sentences and identify which is more toxic. These comparisons can be combined to get a ranking. This method is less biased and easier for annotators than direct annotation on a binary or fine-grained scale.

Regarding the dependence of toxicity on a particular topic, insults may not always have a topic, although there are undoubtedly topics that might be toxic (sexism, racism, and xenophobia). Some publications do not discuss toxicity in general but instead focus on a specific subject. Waseem and Hovy [28] deal with racism and sexism, while sexist and anti-immigrant texts are collected by Basile et al. [33]. Similarly, the concept of a topic in toxicity is utilized to gather data indirectly: Zampieri et al. [22] pre-select messages for toxicity annotation according to their topic. Likewise, Hessel et al. [34] analyze topics for possibly toxic (controversial) conversations.

Unintentional bias in toxicity detection occurs as a result of such a topic-based perspective of toxicity – a misleading link between toxicity and a specific topic (Islam, feminism, LGBT, etc.) [35, 36]. This is consistent with the scope of this research, as it is recognized that among toxicity-provoking issues there are acceptable and unacceptable messages. Existing research proposes algorithmic approaches to debiasing the trained models. For example, Xia et al. [37] teach their model to identify two goals: toxicity and the presence of the topic that can provoke toxicity. Zhang et al. [38] do instance re-weighing, and Park et al. [39] construct pseudo-data to align examples' balance. These works frequently focus on a single topic and apply topic-specific approaches.

The Russian language has also been subjected to topic-based toxicity data gathering. For instance, Bogoradnikova et al. [40] build a set of toxicity corpora in Russian for different topics like sport, education, and entertainment.

According to Ousidhoum et al. [41], in current research the biggest disadvantage of topic-oriented toxicity identification is the ad hoc selection of topics: scientists manually choose the small number of dominant themes or according to topics that appear frequently in the data. Banko et al. [42] take a step toward a more systematic approach to the study of toxicity. They propose, for example, a taxonomy of dangerous online behavior. It contains themes that produce toxicity, but they are combined with other toxicity characteristics, such as direction or severity. As far as it was investigated, the only example of a long list of toxicity-inducing topics is the research of Salminen et al. [43], however, the view of the problem is still limited.

Coming to the point of toxicity detection using topic information, external characteristics, or features that are not derived from the text but come from another source, are frequently used to improve text classification. When it comes to toxicity, the phrase ”context-dependent” is frequently used. To put it in another way, a message that appears neutral in one context may become poisonous in another. However, when compared to context-agnostic approaches, several studies that look at this issue find little evidence that the context has any major impact on toxicity detection [44, 45].

There are several methods for introducing additional data into a classification model. For example, such information (e. g., a topic label) can be incorporated as an additional token in BERT's input. This method is used by Wu et al. [46] for data augmentation with BERT. Similarly, for text generation, Xia et al. [47] insert context into BERT. Yu et al. [48] present the following alternative technique. Each sample is represented by the authors as a pair of sentences separated by the [SEP] token, with the second sentence representing the text sample itself and the first sentence being the auxiliary sentence constructed using one of the methods inherited from Sun et al. [49].

2.2 Sensitive topics and inappropriateness

Babakov et al. [11] describe a *sensitive* topic as one that has a high likelihood of triggering a conversation that could damage the reputation of the speaker. It implies that there are no uniformly sensitive topics; the safety is contingent on the conversation's context and purpose. The level of formality, the regulations of the firm that built the chatbot, and the laws of the country in which it works are all examples of context. It's also worth noting that just because a message touches on a sensitive subject doesn't mean it should be banned. Instead, they propose the concept of appropriateness, which refers to acceptable statements on a sensitive topic.

The authors define an *inappropriate* message as one on a sensitive topic that may frustrate the reader or damage the reputation of the speaker. Because this concept is difficult to formalize, they rely on the intuitive sense of appropriateness that is unique to people and shared by people of the same culture. They specifically inquire if a given statement made by a chatbot can hurt the reputation of the firm that created it. As a result, they rely heavily on human judgments to determine whether something is appropriate.

As was mentioned before, toxicity is one type of messages in this category, but inappropriateness includes more than just toxicity. A message that encourages drug use or suicide, for example, can be written in a non-toxic (and even supportive) manner, but its content is unquestionably harmful (Fig. 2.1).

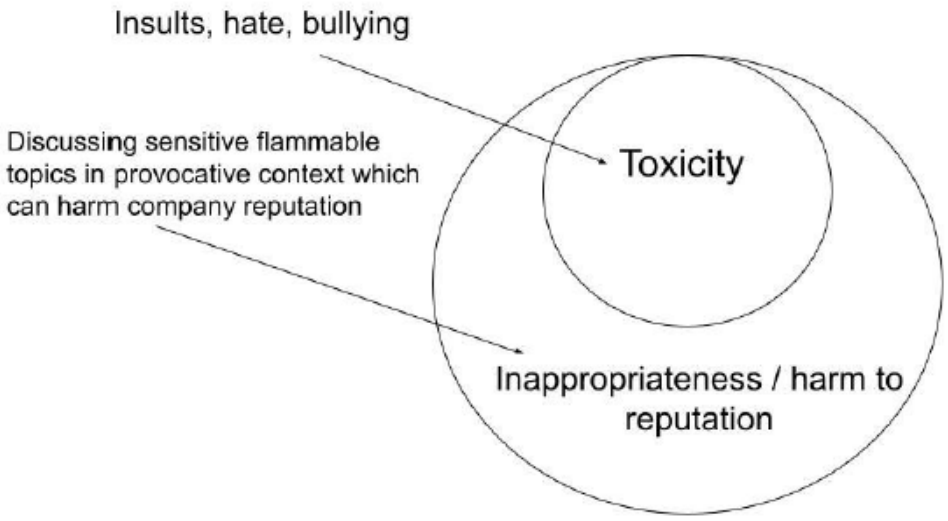

Figure 2.1: Toxicity vs. inappropriateness.

The list of sensitive topics is the following:

- **gambling**;
- **pornography**, description of sexual intercourse;
- **prostitution**;
- **slavery**, human trafficking;
- **suicide**: incitement to suicide, discussion of ways to commit suicide;
- **social injustice** and inequality, social problems, class society;
- **religion**;
- **terrorism**;
- **weapons**;
- **offline crime** (murder, physical assault, kidnapping and other), prison, legal actions;
- **online crime**: breaking of passwords and accounts, viruses, pirated content, stealing of personal information;
- **politics**, military service, past and current military conflicts;
- **body shaming**, people's appearances and clothes;
- **health shaming**, physical and mental disorders, disabilities;
- **drugs**, alcohol, tobacco;
- **racism** and ethnicism;
- **sexual minorities**;
- **sexism**, stereotypes about a particular gender.

They hand-pick this range of delicate themes that frequently lead to inappropriate responses. The collection is diverse: it covers themes relating to potentially risky or harmful actions (such as drugs or suicide), that are be prohibited in the majority of countries (e.g. terrorism, slavery) or themes that can generate aggressive discussion (e.g. politics) and are linked to inequity and

controversy (e.g. minorities), necessitating unique system policies eliminating conversational bias (e.g. response post-processing).

The list of topics is based on proposals and requirements from a significant Russian telecommunications company's legal and public relations departments. It may, for example, be utilized for the moderation of the corporate discussion system or identification of the content that is improper for children, reducing the risk of operational damage.

It is important to note that their main goal is to classify the data as inappropriate, and the sensitive topics are essentially the way of determining inappropriateness, rather than a goal by itself. Hence, they use sensitive topics as the data pre-selection method.

As a result, the inappropriateness labeling procedure is divided into three stages: 1) collection of a dataset of sensitive topic sentences, 2) using this dataset to develop a sensitive topic classifier, and 3) using the classifier to collect sensitive subject texts and categorize them as appropriate or inappropriate. Fig. 2.2 shows it in more detail.

Text	Appropriate	Reason	Sensitive topic
Killing is not a big deal. They have killed no-name ni**a not a politician	✘	Justifying murder	offline crime, racism
This is a violation of the Criminal Law	✔	Shaming the fact of breaking law	offline crime
Religion is the highest level of violence. The believer is forced to believe in infancy.	✘	Insulting the feeling of religious people	religion
Christians, how do you feel about the Gentiles?	✔	Common question about religion	religion
I will earn some money and grab some prostitutes for weekend	✘	Popularisation of prostitution	prostitution
It's good that prostitution is prohibited here	✔	Shaming the prostitution	prostitution

Figure 2.2: Appropriate and inappropriate samples according to the sensitive topics (translated from Russian).

2.3 Classification methods

To solve toxicity detection problems, a variety of methodologies have been used, ranging from linear regression and Support Vector Machines to deep learning. In the best performing systems (for aggression and toxic comment classification), deep learning algorithms such as LSTMs and CNNs were applied [31].

The work of Minaee et al. [50] evaluates the technical contributions, similarities, and strengths of more than 150 deep learning-based text classification models that have been built in recent years. They also present the summary of more than 40 widely used text categorization datasets.

For clarification, they group the models into several categories based on the model architectures:

- Feed-forward networks, which consider the text as a bag of words.
- RNN-based models, where text is viewed as a sequence of words, and which are designed to capture word dependencies and text structures.
- CNN-based models, which are taught to recognize text patterns like significant phrases for text classification.
- Capsule networks, which alleviate the problem of information loss caused by CNN pooling procedures, and have recently been used for text classification.
- The attention mechanism, which is useful in constructing DL-based models since it is effective in identifying correlated terms in the text.
- Memory-augmented networks, which unite neural networks with an external memory system that the models may read and write to.
- Graph neural networks, which capture internal graph structures of natural language like syntactic and semantic parse trees.
- Siamese Neural Networks, which are built for text matching - a special case of text classification.

- Hybrid models unite attention, RNNs, CNNs, and other techniques in to capture local and global aspects of phrases and documents.

- Lastly, they look at modeling techniques other than supervised learning, such as unsupervised learning with autoencoders and adversarial training, as well as reinforcement learning.

Fig. 2.3 indicates how various models performed in news categorization and topic classification tasks. The Italic font here shows that the model is non-deep-learning.

	News Categorization			**Topic Classification**	
Method	AG News	20NEWS	Sogou News	DBpedia	Ohsumed
Hierarchical Log-bilinear Model [221]	-	-	-	-	52
Text GCN [107]	67.61	86.34	-	-	68.36
Simplfied GCN [108]	-	88.50	-	-	68.50
Char-level CNN [50]	90.49	-	95.12	98.45	-
CCCapsNet [76]	92.39	-	97.25	98.72	-
LEAM [84]	92.45	81.91	-	99.02	58.58
fastText [30]	92.50	-	96.80	98.60	55.70
CapsuleNet B [71]	92.60	-	-	-	-
Deep Pyramid CNN [49]	93.13	-	98.16	99.12	-
ULMFiT [216]	94.99	-	-	99.20	-
L MIXED [174]	95.05	-	-	99.30	-
BERT-large [220]	-	-	-	99.32	-
XLNet [156]	95.51	-	-	99.38	-

Figure 2.3: The classification models' accuracy in news categorization and topic classification tasks.

The "Deep Learning" textbook of Goodfellow et al. [51] can also be viewed for better comprehension.

Regarding the text classification for the Russian language, several works also tested various models. Convolutional Neural Networks were used by Gordeev [52] to detect aggressive comments on anonymous imageboards. Andrusyak et al. [53] suggested the unsupervised probabilistic technique with the seed dictionary for categorizing toxic comments published in Ukrainian and Russian languages on YouTube. Makhnytkina et al. [54] investigated several methods for transferring information from English to Russian for the categorization of toxic texts, such as machine translation, domain adaptation, and multilingual models. Smetanin et al. [55] manually labeled the

dataset from Pikabu (Russian social network) and made fine-tuning experiments with the RuBERT and ToxicRuBERT models (Fig. 2.4).

Model	P	R	F_1
RuBERT-TPikabu	0.8335 ± 0.0051	0.8268 ± 0.0028	0.8300 ± 0.0020
ToxicRuBERT-TPikabu	0.8414 ± 0.0050	0.8358 ± 0.0050	$\mathbf{0.8385 \pm 0.0039}$
RuBERT-TPikabu-25KTC	0.8339 ± 0.0048	0.8231 ± 0.0022	0.8283 ± 0.0017
RuBERT-TPikabu-50KTC	0.8170 ± 0.0169	0.8322 ± 0.0031	0.8239 ± 0.0102
RuBERT-TPikabu-75KTC	0.7734 ± 0.0312	0.8334 ± 0.0052	0.7931 ± 0.0286
RuBERT-TPikabu-100KTC	0.7255 ± 0.0393	0.8277 ± 0.0143	0.7417 ± 0.0505
ToxicRuBERT-TPikabu-25KTC	0.8369 ± 0.0079	0.8363 ± 0.0060	0.8364 ± 0.0042
ToxicRuBERT-TPikabu-50KTC	0.8315 ± 0.0104	0.8304 ± 0.0049	0.8308 ± 0.0053
ToxicRuBERT-TPikabu-75KTC	0.8163 ± 0.0053	0.8363 ± 0.0029	0.8256 ± 0.0028
ToxicRuBERT-TPikabu-100KTC	0.7628 ± 0.0331	0.8378 ± 0.0068	0.7857 ± 0.0344
RuBERT-TPikabu-25VKHate	0.8295 ± 0.0042	0.8278 ± 0.0015	0.8286 ± 0.0015
RuBERT-TPikabu-50VKHate	0.8342 ± 0.0038	0.8261 ± 0.0015	0.8301 ± 0.0018
RuBERT-TPikabu-75VKHate	0.8208 ± 0.0053	0.8282 ± 0.0047	0.8243 ± 0.0022
RuBERT-TPikabu-100VKHate	0.8039 ± 0.0063	0.8387 ± 0.0037	0.8194 ± 0.0038

Figure 2.4: The classification metrics of the models.

2.4 Text detoxification methods

After detecting that the text is inappropriate the next stage is its detoxification. Instagram, for example, offers to reformulate toxic comments before posting them.

The task of style transfer involves changing a text so that its content and most of its attributes remain the same while one attribute (style) is modified. The sentiment [56], the bias presence [57], the degree of formality [58] are examples of this attribute. More details of style transfer applications can be found in [59].

Regarding the detoxification task, different groups of researchers have already addressed it [60, 61], as well as a comparable goal of changing text into a more polite form [62].

Yet, all of these works are limited to the English language. Russian text style transfer and text detoxification procedures were investigated only by Dementieva et al. [63], where the authors present the Russian language's new text detoxification study and perform various experiments with the style transfer methods.

Chapter 3

Methodology

3.1 Datasets collection

Babakov et al. [11] obtained two datasets: the sensitive topics dataset and the appropriateness dataset.

The first step in the creation of the datasets was to collect sensitive topics lists, as described in sec. 2.2. After that, they collected the initial pool of texts from a variety of sources, filtered them, then hired crowd workers to manually categorize them for the sensitive topics' presence. They got the information from the following places: 2ch.hk and otvet.mail.ru. The labeling was performed with the use of a crowdsourcing platform Yandex.Toloka.

The topic labeling task is usually handled as a multiple-choice task with the option of selecting several answers: the text and possible topics are provided to the worker, and he is asked to select one or more of them. Since choosing from such a wide range of options is tough as there are 18 sensitive topics, the topics were divided into three groups to make the process easier: 1) gambling, pornography, prostitution, slavery, suicide, social injustice, 2) religion, terrorism, weapons, offline crime, online crime, politics, 3) body shaming, health shaming, drugs, racism, sex minorities, sexism.

Crowd workers were asked to complete the training procedure before labeling the samples. There were 20 questions with pre-determined answers. A worker must finish the training with at least 65 percent right answers to be admitted to labeling. Similarly, pre-defined replies are applied to perform the quality control: one out of ten questions presented to the worker is used for the performance monitoring. At least 25 % of the control questions should be answered correctly, otherwise, he is excluded from labeling in the future, and his most recent responses are deleted. Workers' average performance on control and training tasks ranged from 65 to 70 %.

Then, a classifier was trained that predicts the existence of a sensitive topic in a text after collecting over 10,000 texts on the sensitive subject. Despite the classifier is not great for real-

world applications, the authors believe that, if samples are labeled as belonging to a sensitive topic with high confidence (greater than 0.75 in the testing), then they can be perceived as belonging to the sensitive topic.

The appropriateness labeling was done using the Yandex.Toloka crowdsourcing platform in the same way that the sensitivity topic labeling was done. The workers showed a solid comprehension of the appropriateness concept, despite the fact that it wasn't expressly stated. The training and control tasks' average performance was 75-80 %, demonstrating strong agreement.

As a result, two datasets were collected. Regarding the statistics, there are 25,679 unique samples in the sensitive topics dataset. A crowdsourcing platform was used to label 9,946 samples, their team labeled approximately 1,500 samples, and the remaining samples were obtained using keywords from specialist sources. The crowdsourcing annotation has an average confidence level of 0.995; the average number of annotations per example is 4.3; and the average time to classify one example is 10.8 seconds.

There are 82,063 unique samples in the appropriateness dataset. 8,687 of these samples are also part of the sensitive topics collection and have been given topic labels by hand. The labels for the remaining 73,376 data were generated automatically using a BERT-based topic classification algorithm. The annotation's average confidence is 0.956; the average number of annotations per example is 3.5; and it takes an average of 7 seconds to label one example.

Fig. 3.1 illustrates the number of samples collected on each sensitive topic in both datasets. While considerable effort was made to equalize the distribution of the topics, several of them (drugs, health shaming, and politics) receive significantly more samples in the appropriateness dataset.

A single sample can be applied to multiple topics. 15 % of similar occurrences in the data was found. The order in which topics appear is not random. It denotes the intersection of many topics. "Politics, racism, social injustice," "prostitution, pornography," and "sex minorities, pornography" are the most common co-occurrences. In the topic dataset, however, 13 % of samples do not interact with any sensitive topics. These are examples of topics that were pre-selected for the manual topic labeling with keywords and subsequently labeled as unrelated to the topics of interest. They were included in the dataset so the classifier trained on it would not be based primarily on keywords.

The datasets' samples are mainly single sentences, with the appropriateness dataset's average length being 15 words and the topic dataset's average length being 18 words. The sample length varies between 14 and 21 words depending on the topic. The number of samples of a certain

Sensitive topic	Topic dataset	Appropriateness dataset
total samples	25,679	82,063
religion	4,110	2,869
drugs	3,870	8,618
sex minorities	1,970	754
health shaming	1,744	7,270
politics	1,593	7,650
weapons	1,530	726
suicide	1,420	1,931
gambling	1,393	2,693
pornography	1,289	2,824
social injustice	1,230	5,294
racism	1,156	3,760
online crime	1,058	3,181
offline crime	1,037	2,206
sexism	1,022	3,644
body shaming	715	3,537
prostitution	634	240
terrorism	577	310
slavery	288	442

Figure 3.1: Sensitive topics and appropriateness datasets' sample numbers.

topic and the average number of words per sample for this topic had a strong correlation (Spearman's r of 0.72). It is not clear if this is a coincidence or whether topics with longer sentences are better represented in the data. It may be easier to annotate longer sentences.

3.2 Datasets classification

3.2.1 Logistic regression

Logistic regression is one of the most popular classification techniques used in machine learning. Here, the probability that a given example belongs to the "1" class versus the probability that it belongs to the "0" class is predicted. Specifically, the hypothesis is:

$$\begin{aligned} z &= b + w^T x = b + \sum_{j=1}^{m} w_i * x_i \\ \hat{y} &= h_w(x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \end{aligned} \tag{3.1}$$

where $\sigma(z)$ is Sigmoid function.

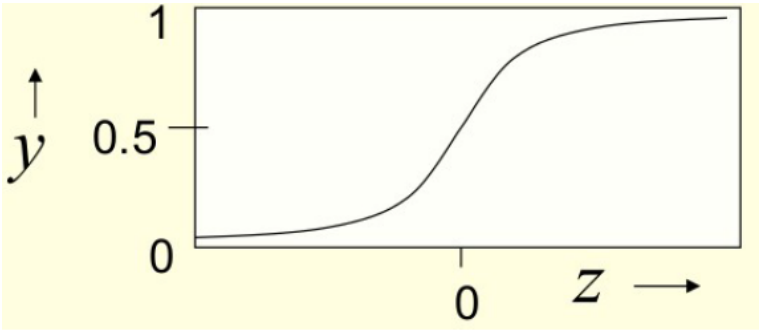

Figure 3.2: Logistic regression function.

3.2.2 BERT

Introduced in the work of Devlin et al. [64], BERT is a transformer-based machine learning model for NLP, which stands for Bidirectional Encoder Representations from Transformers. It is one of the most powerful and well-known language representation models nowadays. It is intended to condition both left and right context in all layers to pre-train deep bidirectional representations from the unlabeled text.

Consequently, the pre-trained BERT model could be fine-tuned with just one additional output layer to provide state-of-the-art models for a wide range of tasks, including question answering and language inference, without requiring significant task-specific architecture changes.

Originally, there were two models for English language: $BERT_{BASE}$ and $BERT_{LARGE}$. In $BERT_{BASE}$, there are 12 transformer encoder layers, 768 hidden size, 12 attention heads, and

110M parameters. In $BERT_{LARGE}$, there are 24 transformer encoder layers, 1024 hidden size, 16 attention heads, and 340M parameters.

As input, the BERT model expects a sequence of tokens (words). There are two special tokens that BERT would expect as input in each sequence of tokens:

- `[CLS]`: the classification token, which is the first token in every sequence.
- `[SEP]`: the token that tells BERT what sequence each token belongs to. It is particularly useful in tasks involving sentence prediction or question answering. If there is only one sequence, the token will be appended to the end of it.

Fig. 3.3 shows BERT's overall pre-training and fine-tuning procedures for the question-answering example. Apart from output layers, both pre-training and fine-tuning apply identical architectures. Models for various downstream tasks are initialized using the same pre-trained model parameters. More details can be found in the paper of Devlin et al. [64].

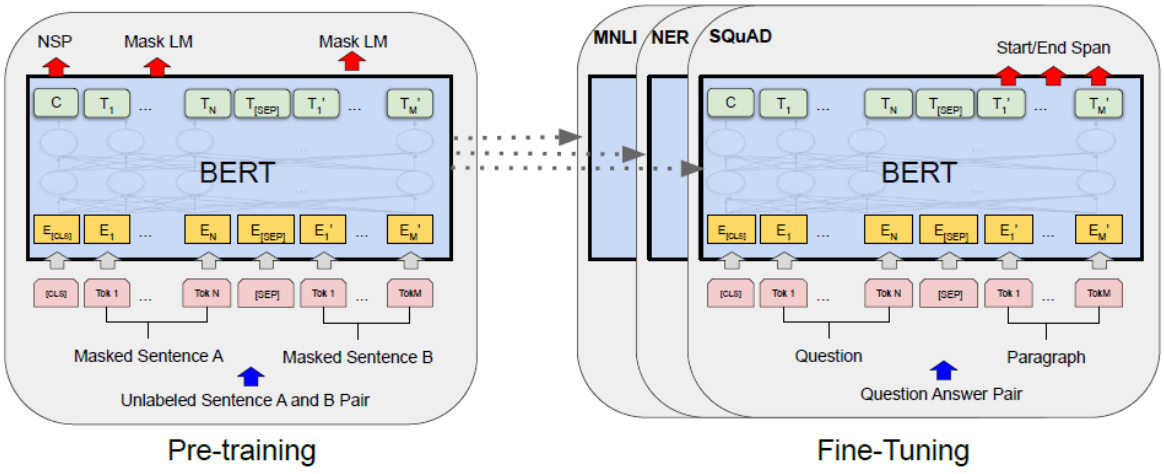

Figure 3.3: BERT pre-training and fine-tuning procedures.

Regarding the model architecture, BERT is a multi-layer bidirectional transformer encoder, and a detailed description of the original transformer implementation can be found in Vaswani et al. [65].

3.3 Text detoxification

GPT-2 [6] is a robust language model that can be used in a large number of Natural Language Processing applications and trained on small task-specific data. There were no such models for the Russian language till quite recently. The ruGPT3 model, which can generate cohesive and comprehensible Russian sentences, was presented as part of the AI Journey competition [66].

The following settings are recommended for the style transfer task [63]:

- **zero-shot**: the model is used exactly as it is (with no fine-tuning). A toxic statement that is going go to be detoxified is considered as the input, with the prefix "Перефразируй" (ru. paraphrase) before and the suffix >>> after for the paraphrasing task denoting. Because ruGPT3 has previously been trained to accomplish this task, the case is similar to paraphrase (Fig. 3.4).

- **few-shot**: the model is used exactly as it is. In contrast to the zero-shot setting, we provide the prefix containing from the parallel dataset $\left\{\left(t_1^X, t_1^Y\right), \ldots, \left(t_n^X, t_n^Y\right)\right\}$ of toxic and neutral sentences in the form "t_i^X >>> t_i^Y". These examples can help the model comprehend that detoxifying paraphrasing is necessary. The input sentence is followed by the parallel sentences, which are going to be detoxified using the prefix "Перефразируй" and the suffix >>> (Fig. 3.5).

- **fine-tuned**: on a parallel dataset $\left\{\left(t_1^X, t_1^Y\right), \ldots, \left(t_n^X, t_n^Y\right)\right\}$, the model is fine-tuned for the paraphrasing job. This involves the model being trained on strings of the kind "t_i^X >>> t_i^Y". Following the training, we provide input to the model in the same manner as in the other cases (Fig. 3.6).

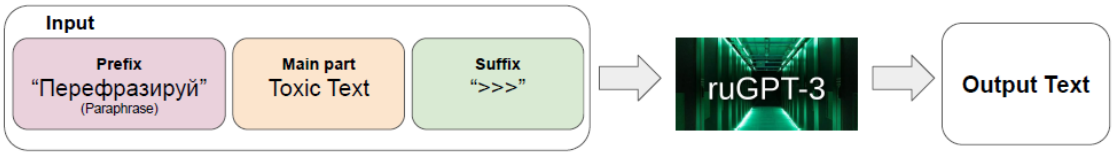

Figure 3.4: The schematic pipeline for the zero-shot setup.

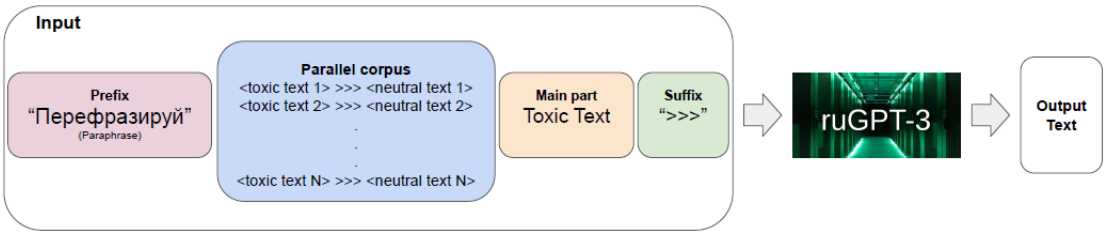

Figure 3.5: The schematic pipeline for the few-shot setup.

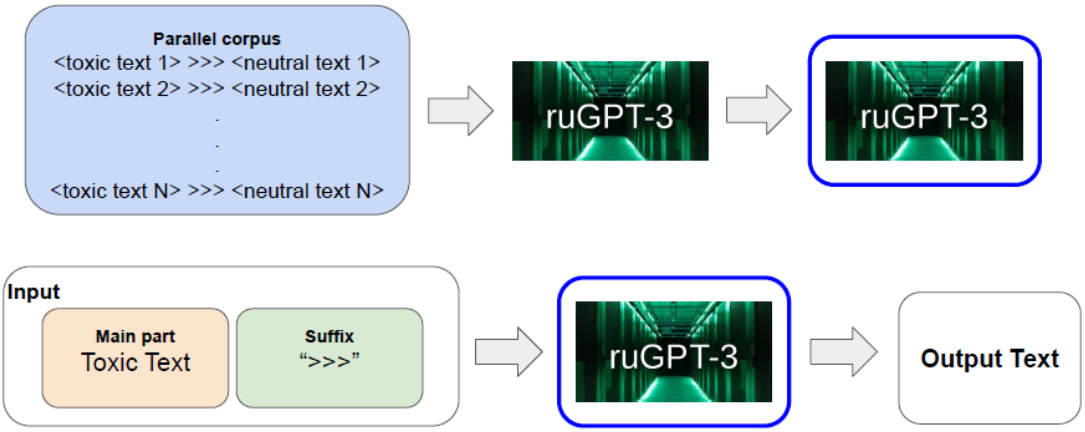

Figure 3.6: The schematic pipeline for the fine-tuned setup.

Chapter 4

Results and Discussion

4.1 Datasets description

All the details about the sensitive topics dataset and the appropriateness dataset were given in the previous sections.

4.2 Datasets classification

4.2.1 Baseline: logistic regression

The logistic regression model and the appropriateness dataset were considered for the first baseline.

The pre-trained `BertTokenizer` that was used at the preprocessing stage can be visualized in this manner (Fig. 4.1):

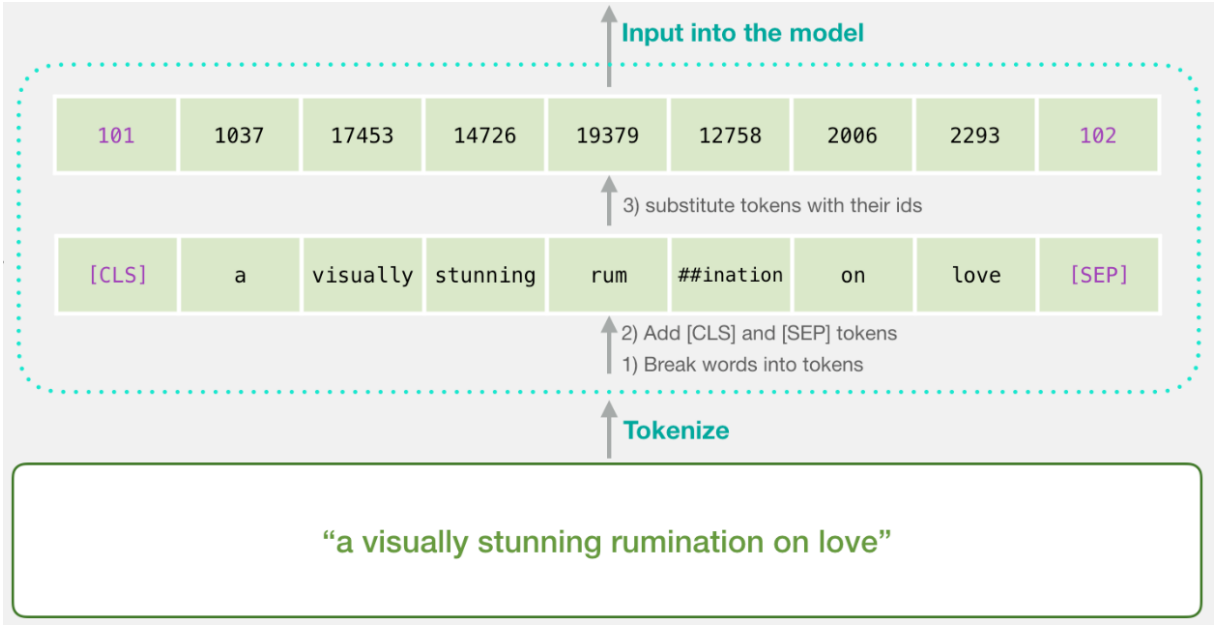

Figure 4.1: BERT tokenizer.

It takes as input the collection of raw sentences, splits them into tokens, adds special tokens

`[SEP]` and `[CLS]`, and encodes each token with its index. The big convenience here is that `BertTokenizer` pads encoded sentences to the same length, so they can be packed into tensor.

Then, tensor with the input indexes is fed into the embedding layer and then into the linear layer. Finally, the softmax function is used to obtain class probabilities for each sentence.

The batch size was equal to 16. The logistic regression model was trained with the Adam optimizer with the constant learning rate 2×10^{-2}. The classifier performance is measured with F1-score, which takes the harmonic mean of a classifier's precision and recall to create a single metric. Its best F1 score on the validation dataset was equal to 0.58.

4.2.2 Baseline: BERT

It was decided to include the simple BERT-based approach in this research to compare it with the first baseline and with the main method.

The BERT model pre-trained by Kuratov et al. [67] was used. The preprocessing step here involves the same tokenization as described in 4.2.1. After token IDs and attention mask are fed into pre-trained BERT, the fixed-dimension hidden state vector of a sentence is obtained. It can be fed into the linear layer of the corresponding dimension and the softmax layer to obtain the class probabilities, as described before.

The batch size was also equal to 16. The model was trained with the learning rate 1×10^{-3}, regularization coefficient 1×10^{-2}, and 500 warm-up steps. It achieved F1-score of 0.81 on the validation dataset.

4.2.3 Main method

The main approach implemented in this research involves four stages.

Multilabel classifier

At this stage, the multilabel classification task on the sensitive topics dataset was considered. It contains 18 sensitive topics; each sentence can belong to either several of them (one or more than one) or none of them. Hence, the topic labels of sentences form 393 combinations and each of them can be viewed as a separate class label. Similar to the binary classification described in 4.2.2, the BERT-based multilabel classifier was trained here. Its trained weights will be used further. The classification report at this stage can be viewed in Fig. 4.2.

	precision	recall	f1-score	support
offline_crime	0.52	0.57	0.54	76
online_crime	0.59	0.57	0.58	23
drugs	0.90	0.90	0.90	58
gambling	0.00	0.00	0.00	2
pornography	0.82	0.39	0.53	128
prostitution	0.80	0.75	0.77	55
slavery	0.71	0.68	0.70	22
suicide	1.00	0.50	0.67	4
terrorism	0.65	0.42	0.51	26
weapons	0.89	0.91	0.90	96
body_shaming	0.88	0.51	0.65	68
health_shaming	0.96	0.62	0.76	72
politics	0.84	0.32	0.46	159
racism	0.85	0.43	0.57	127
religion	0.84	0.83	0.83	63
sexual_minorities	0.88	0.40	0.55	58
sexism	0.68	0.49	0.57	81
social_injustice	0.81	0.18	0.30	114
none	0.60	0.69	0.64	188

Figure 4.2: Classification report.

Weight averaging trick

From the previous stage, we have the linear layer matrix of shape 768×393. By applying max operation to each column and extracting distinct topics, we obtain 786-dimensional embeddings for each of 18 initial topics. We will use these embeddings later.

Additional tokens

Then, we return to the appropriateness dataset. Apart from the ordinary tokens constructed by the BERT tokenizer, we use 18 special tokens denoting dangerous topics. We add to each sentence in the dataset a token corresponding to the topic which this sentence belongs to. Sentence with added topic token may look like `[PL]` `[RS]` `[SL]` `[SEP]` . . . , where `[PL]` denotes politics, `[RS]` denotes racism, etc.

Model	Input Sequence	Label
BERT4TC-S	[CLS] I like this film. [SEP]	{negative, **positive**}
BERT4TC-AQ	[CLS] I like this film. [SEP] What is the result? [SEP]	{negative, **positive**}
BERT4TC-AA	[CLS] I like this film. [SEP] positive [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] negative [SEP]	{**0**, 1}
BERT4TC-AWA	[CLS] I like this film. [SEP] The result is positive. [SEP]	{0, **1**}
	[CLS] I like this film. [SEP] The result is negative. [SEP]	{**0**, 1}

Figure 4.3: Examples of input sequence constructions.

Sentence transformation and final tuning

To improve the above-mentioned method, it was decided to include the approach described in [48] in the experiments. The authors suggest adding auxiliary sentences to original ones to improve the BERT classification accuracy. The example is shown in Fig. 4.3. This approach was applied to the samples in the dataset.

Then, the aforementioned DeepPavlov pre-trained BERT was initialized. To each weight vector in the BERT parameters corresponding to the topic token, we assign the topic embedding value obtained above. Finally, the model was applied for binary classification of the transformed dataset sentences.

The main approach was trained with the same learning parameters as the baseline BERT. It achieved an outstanding F1-score level of 0.98 on validation.

Several approaches shown in Fig. 4.3 were tested and it was investigated that the best quality is provided by the BERT4TC-AWA transformation approach. The reason is that the auxiliary sentence derived from the labels is a pseudo-sentence made up of a categorical label and a few extra words.

The table with results can be seen below:

Model	F1-score
Logistic regression	0.58
BERT	0.81
Main method	0.98

Table 4.1: Experiments results.

4.3 Text detoxification

RuGPT3 comes in three versions: `small` (125m parameters, 2048 context), `medium` (350m parameters, 2048 context), and `large` (760m parameters, 2048 context). Large is used here.

The following hyper-parameters are used by the ruGPT3 model:

- top_k: the integer parameter ≥ 1. Words are produced one by one by transformers (which GPT is), with the following word being chosen from among the top k options, ranked by probability. We use $top_k = 5$.
- top_p: the floating-point parameter in 0-1. The sampling is done by picking from the smallest collection of words possible, where cumulative probability is greater than the probability p, analogous to the top_k parameter. We use $top_p = 0.9$.
- $temperature(t)$: the floating-point parameter ≥ 1. It indicates the model's degree of freedom. Instead of paraphrasing, the model might initiate a conversation at higher temperatures (100), yet at the temperature of approximately 1, it only slightly modifies the phrase. We use $t = 40$.

	text	inappropriate	paraphrased text
8	Кароче как я понял Татьяна это эскортница из русской провинции но её обманули и теперь держат в секс рабстве	1.00	Как я понял Татьяна это эскортница с российской глубинки но её обманули и теперь держат в рабстве.
19	Ну привет Маря, ну рассказывай, сколько сигарет скурила?	0.79	Здравствуйте, Марья, подскажите сколько сигарет вы выкурили? Ну,
49	Если девушки не знакомятся с парнем, то этот парень неликвидный самец и он никому не нужен?	0.93	Девушка не знакомятся с парнем, то этот юноша никудышный самец и он никому не нужен?

Figure 4.4: Text detoxification in Russian.

The overall style transfer accuracy is 0.6, which explains why some words were changed clumsily or stayed the same (4.4 and Fig. 4.5). This happens because the phrases are created from

	text	inappropriate	paraphrased text
8	On the quicks as I got it Tatiana is an escort from da Russian provinces but she wuz deceived n now day are holding her for sex slavery	1.00	As I got it Tatiana is an escort from the Russian countryside but she was deceived and now they are holding her for slavery
19	Well Hi Marya, so tell me, how many cigarettes did you finish?	0.79	Hello, Marya, could you tell how many cigarettes did you smoke? Well,
49	If girls don't get acquainted with a guy, then this guy is an illiquid male and nobody needs him?	0.93	Girl don't get acquainted with a guy, then this guy is useless male and nobody needs him?

Figure 4.5: Text detoxification in English (translated from Russian).

scratch when using the model. We have no control over the content's preservation (the models might occasionally change it completely).

Chapter 5

Conclusion

During this work, the notion of inappropriateness, which is a wider class of poisonous communications than toxicity, was presented. The inappropriate dataset was classified by two baseline models and the main method, the F-1 score of the latter showed good results that gave a significant improvement over baselines. Despite that acceptable results were obtained, other context-awareness approaches can be used to improve the scores.

In addition, data was detoxified with the usage of the recently released RuGPT3 model, and the results were pretty average, as 40 % of the text was paraphrased awkwardly or not changed at all. Further improvements to the model are necessary.

Bibliography

[1] Jigsaw, "Toxic comment classification challenge," https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2018.

[2] ——, "Jigsaw unintended bias in toxicity classification," https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification, 2019.

[3] ——, "Jigsaw multilingual toxic comment classification," https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification, 2020.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[5] L. Breitfeller, E. Ahn, D. Jurgens, and Y. Tsvetkov, "Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 1664–1674.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[7] "Tay (BOT)," https://en.wikipedia.org/wiki/Tay_(bot).

[8] "A South Korean Chatbot Shows Just How Sloppy Tech Companies Can Be With User Data)," https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html.

[9] "Another AI chatbot shown spouting offensive views," https://techcrunch.com/2017/10/24/another-ai-chatbot-shown-spouting-offensive-views/.

[10] "Overcoming data sourcing issues when testing finance virtual assistants," https://www.fintechfutures.com/2020/04/overcoming-data-sourcing-issues-when-testing-finance-virtual-assistants/.

[11] N. Babakov, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko, "Detecting inappropriate messages on sensitive topics that could harm a company's reputation," *arXiv preprint arXiv:2103.05345*, 2021.

[12] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser, "Anticipating safety issues in e2e conversational ai: Framework and tooling," *arXiv preprint arXiv:2107.03451*, 2021.

[13] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper attention to abusive user content moderation," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 1125–1135.

[14] N. Cecillon, V. Labatut, R. Dufour, and G. Linares, "Wac: A corpus of wikipedia conversations for online abuse detection," *arXiv preprint arXiv:2003.06190*, 2020.

[15] E. Fersini, D. Nozza, and P. Rosso, "Overview of the evalita 2018 task on automatic misogyny identification (ami)," *EVALITA Evaluation of NLP and Speech Tools for Italian*, vol. 12, p. 59, 2018.

[16] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "A large-scale semi-supervised dataset for offensive language identification," *arXiv preprint arXiv:2004.14454*, 2020.

[17] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech," *arXiv preprint arXiv:1910.03270*, 2019.

[18] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech," *arXiv preprint arXiv:1909.04251*, 2019.

[19] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "Ethos: an online hate speech detection dataset," *arXiv preprint arXiv:2006.08328*, 2020.

[20] O. De Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," *arXiv preprint arXiv:1809.04444*, 2018.

[21] X. Han and Y. Tsvetkov, "Fortifying toxic speech detectors against veiled toxicity," *arXiv preprint arXiv:2010.03154*, 2020.

[22] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," *arXiv preprint arXiv:1902.09666*, 2019.

[23] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," *arXiv preprint arXiv:1705.09899*, 2017.

[24] A. Lees, D. Borkan, I. Kivlichan, J. Nario, and T. Goyal, "Capturing covertly toxic speech via crowdsourcing," in *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 2021, pp. 14–20.

[25] "Russian Language Toxic Comments," https://www.kaggle.com/datasets/blackmoon/russian-language-toxic-comments, 2019.

[26] "Toxic Russian Comments," https://www.kaggle.com/datasets/alexandersemiletov/toxic-russian-comments, 2021.

[27] N. Zueva, M. Kabirova, and P. Kalaidin, "Reducing unintended identity bias in russian hate speech detection," *arXiv preprint arXiv:2010.11666*, 2020.

[28] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[29] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[30] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[31] E. Dinan, S. Humeau, B. Chintagunta, and J. Weston, "Build it break it fix it for dialogue safety: Robustness from adversarial human attack," *arXiv preprint arXiv:1908.06083*, 2019.

[32] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan, "Recipes for safety in open-domain chatbots," *arXiv preprint arXiv:2010.07079*, 2020.

[33] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti *et al.*, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019, pp. 54–63.

[34] J. Hessel and L. Lee, "Something's brewing! early prediction of controversy-causing posts from discussion features," *arXiv preprint arXiv:1904.07372*, 2019.

[35] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 67–73.

[36] A. Vaidya, F. Mai, and Y. Ning, "Empirical analysis of multi-task learning for reducing model bias in toxic comment detection," *arXiv preprint arXiv:1909.09758*, 2019.

[37] M. Xia, A. Field, and Y. Tsvetkov, "Demoting racial bias in hate speech detection," *arXiv preprint arXiv:2005.12246*, 2020.

[38] G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, and T. Zhao, "Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting," *arXiv preprint arXiv:2004.14088*, 2020.

[39] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection," *arXiv preprint arXiv:1808.07231*, 2018.

[40] D. Bogoradnikova, O. Makhnytkina, A. Matveev, A. Zakharova, and A. Akulov, "Multilingual sentiment analysis and toxicity detection for text messages in russian," in *2021 29th Conference of Open Innovations Association (FRUCT)*. IEEE, 2021, pp. 55–64.

[41] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," *arXiv preprint arXiv:1908.11049*, 2019.

[42] M. Banko, B. MacKeen, and L. Ray, "A unified taxonomy of harmful content," in *Proceedings of the fourth workshop on online abuse and harms*, 2020, pp. 125–137.

[43] J. Salminen, S. Sengün, J. Corporan, S.-g. Jung, and B. J. Jansen, "Topic-driven toxicity: Exploring the relationship between online toxicity and news topics," *PloS one*, vol. 15, no. 2, p. e0228723, 2020.

[44] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity detection: Does context really matter?" *arXiv preprint arXiv:2006.00998*, 2020.

[45] M. Karan and J. Šnajder, "Preemptive toxic language detection in wikipedia comments using thread-level context," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 129–134.

[46] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional bert contextual augmentation," in *International Conference on Computational Science*. Springer, 2019, pp. 84–95.

[47] C. Xia, C. Zhang, H. Nguyen, J. Zhang, and P. Yu, "Cg-bert: Conditional text generation with bert for generalized few-shot intent detection," *arXiv preprint arXiv:2004.01881*, 2020.

[48] S. Yu, J. Su, and D. Luo, "Improving bert-based text classification with auxiliary sentence and domain knowledge," *IEEE Access*, vol. 7, pp. 176 600–176 612, 2019.

[49] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, 2019.

[50] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[52] D. Gordeev, "Detecting state of aggression in sentences using cnn," in *International conference on speech and computer*. Springer, 2016, pp. 240–245.

[53] B. Andrusyak, M. Rimel, and R. Kern, "Detection of abusive speech for mixed sociolects of russian and ukrainian languages." in *RASLAN*, 2018, pp. 77–84.

[54] O. Makhnytkina, A. Matveev, D. Bogoradnikova, I. Lizunova, A. Maltseva, and N. Shilkina, "Detection of toxic language in short text messages," in *International Conference on Speech and Computer*. Springer, 2020, pp. 315–325.

[55] S. Smetanin and M. Komarov, "Share of toxic comments among different topics: The case of russian social networks," in *2021 IEEE 23rd Conference on Business Informatics (CBI)*, vol. 2. IEEE, 2021, pp. 65–70.

[56] I. Melnyk, C. N. d. Santos, K. Wadhawan, I. Padhi, and A. Kumar, "Improved neural text attribute transfer with non-parallel data," *arXiv preprint arXiv:1711.09395*, 2017.

[57] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, "Automatically neutralizing subjective bias in text," in *Proceedings of the aaai conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 480–489.

[58] S. Rao and J. Tetreault, "Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer," *arXiv preprint arXiv:1803.06535*, 2018.

[59] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Computational Linguistics*, vol. 48, no. 1, pp. 155–205, 2022.

[60] C. N. d. Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," *arXiv preprint arXiv:1805.07685*, 2018.

[61] M. Tran, Y. Zhang, and M. Soleymani, "Towards a friendly online community: An unsupervised style transfer framework for profanity redaction," *arXiv preprint arXiv:2011.00403*, 2020.

[62] A. Madaan, A. Setlur, T. Parekh, B. Poczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, and S. Prabhumoye, "Politeness transfer: A tag and generate approach," *arXiv preprint arXiv:2004.14257*, 2020.

[63] D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, and A. Panchenko, "Methods for detoxification of texts for the russian language," *Multimodal Technologies and Interaction*, vol. 5, no. 9, p. 54, 2021.

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[66] "AI Journey competition," https://ai-journey.ru, 2021.

[67] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for russian language," *arXiv preprint arXiv:1905.07213*, 2019.

[68] S. Smetanin, "Toxic comments detection in russian," in *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue*, 2020.