

When a Language Question Is at Stake. A Revisited Approach to Label Sensitive Content

Daria Stetsenko

NASK National Research Institute

Warsaw, Poland

daria.stetsenko@nask.pl

Abstract

Many under-resourced languages require high-quality datasets for specific tasks such as offensive language detection, disinformation, or misinformation identification. However, the intricacies of the content may have a detrimental effect on the annotators. The article aims to revisit an approach of pseudo-labeling sensitive data on the example of Ukrainian tweets covering the Russian-Ukrainian war. Nowadays, this acute topic is in the spotlight of various language manipulations that cause numerous disinformation and profanity on social media platforms. The conducted experiment highlights three main stages of data annotation and underlines the main obstacles during machine annotation. Ultimately, we provide a fundamental statistical analysis of the obtained data, evaluation of models used for pseudo-labeling, and set further guidelines on how the scientists can leverage the corpus to execute more advanced research and extend the existing data samples without annotators' engagement.

1 Introduction

The Russian invasion of Ukraine has been causing thousands of casualties, millions of displaced people, and severe economic and social consequences for many countries. The full-fledged escalation of the conflict broke out on February 24, 2022, when Russian forces trespassed the sovereign country's territory with flying jets and military vehicles (Ellyatt, 2022). The wave of misinformation, panic, and mass hysteria took a toll on millions of Ukrainians during the first days of the invasion. Although the war has been ongoing for over a year, the problem of disinformation, misinformation, and harmful content identification across various social media platforms remains an open issue. The scarcity of well-annotated and verified warfare datasets is the main obstacle to developing high-quality models for offensive speech detection, disinformation, and misinformation classification (Poletto et al., 2021).

The study by Pierri et al. (2023) is of particular interest as the authors examine Twitter accounts' creation and suspension dynamics based on tweets about the Russian-Ukrainian war. The scientists underline the vagueness of Twitter's policies regarding de-platforming. The most common soft-moderation tactics deployed by Twitter include down-ranking (lowering the visibility of certain content in users' feeds), "shadow banning" (hiding content from other users), and warning labels (tagging content as potentially harmful or inaccurate) (Papakyriakopoulos and Goodman, 2022; Ali et al., 2021; Pierri et al., 2022). As a result, some people fall victim to "shadow banning" by an algorithm's miscalculation. It might also apply to Ukrainian accounts that post messages in their native language but get down-ranked due to the incongruence of a moderation model. To implement a well-rounded neural network, one must have high-quality data, such as a labeled dataset for the Ukrainian language. Therefore, the study presents the first and only tagged Ukrainian corpus for offensive language detection in the context of the Russian-Ukrainian war.

The article's main objective is to describe a new data collection and labeling approach. The sensitive content of gathered tweets requires a rigorous algorithm to minimize the subjectivity of human evaluation. Hence a pseudo-labeling technique has been utilized at the second and third stages of the annotation. We also highlight the main challenges and limitations faced at different phases. In the end, some descriptive statistics and general analyses are offered to illustrate the potential and usefulness of the dataset for studying the offensive language in the context of the Russian-Ukrainian war. We hope this dataset will contribute to a better understanding of the offensive context in Ukrainian tweets and will be applied to various types of research on the level of the Russian dataset VoynaSlov and many English-annotated datasets (Chen and Ferrara, 2023; Park

et al., 2022).

The paper is structured in the following way:

1. Outline of the related works on the Russian-Ukrainian war together with available monolingual Ukrainian datasets;
2. A closer look at the three stages of the data collection and annotation process, a description of the challenges that have occurred down the way;
3. General data statistics of the obtained corpus and suggestions for further research works.

2 UA Corpus for Offensive Language Detection

2.1 Data collection

5000 tweets were prior collected through an available Twitter streaming API service. In order to minimize the probability of acquiring tweets unrelated to the topic of war, we initially selected the ten most prominent and unique hashtags (Table 1) which appeared at different periods of military actions in Ukraine.

After the primary analysis of the social media platform, we can conclude that Ukrainians tend to write their tweets in English rather than Ukrainian. It can be addressed as an effort to attract more attention from Western countries to the situation in Ukraine. We also added a few general hashtags to the existing list (#Ukraine, #Russia, #ukraine, #russia) to get a more extensive collection of tweets. Other filters condition that a tweet should not be a reply or retweet, and the language of the content is strictly Ukrainian. The gathered messages cover a period from 09/2022 to 03/2023.

Although we explicitly stated the target language, some tweets were scrapped in Russian and Belarussian. Therefore, at this step, we eliminate duplicates (usually produced by bots), validate the language, and check a tweet's correspondence to the war (it is necessary as we use a couple of general hashtags). Finally, only 2043 tweets remain in the dataset.

2.2 Definitions of the offensive language

Prior to the annotation stages we need to provide a comprehensive definition and criteria of what is considered as an instance of the offensive language use in our dataset. Every year a large amount of studies tackle the problem of offensive, abuse, hate and toxic language (Wiegand et al., 2021; Davidson

Hashtags

#RussiaIsATerroristState #russiaisaterroriststate
 #WarInUkraine #warinukraine
 #Україна #українці
 #BeBraveLikeUkraine #bebravelikeukraine #braveukraine
 #UkraineWar #UkraineRussiaWar
 #StandWithUkraine
 #рiквiйни
 #Putin #путiн
 #СлаваУкраїні #GloryToUkraine
 #FreeLeopards #freeleopards

Table 1: Top ten unique hashtags related to the war in Ukraine.

et al., 2017; Israeli and Tsur, 2022; Saleem et al., 2022). Despite the numerous studies that present an exhaustive outline and a definition of the offensive language, the scientists point out still occurring discrepancies between annotators (Sigurbergsson and Derczynski, 2020; Goffredo et al., 2022; Ruitenbeek et al., 2022; Ross et al., 2017). We strive to minimize inconsistencies in inter-annotator agreement (IAA) by setting a clear-cut demarcation and regularities of what should be regarded as offensive content (Demus et al., 2022).

Sigurbergsson and Derczynski (2020) formulates the offensive language as a phenomenon that varies greatly and ranges from simple obscene language to more severe cases such as life threat, hate, bullying and toxicity. Bretschneider and Peters (2017) states that hate speech, cyberhate and offensive language are umbrella terms used in the context of social media to denote offending or hostile message. Many researchers highlight that it remains hard to distinguish between offensive language and hate speech (Waseem et al., 2017; Sigurbergsson and Derczynski, 2020; Waseem and Hovy, 2016; Stamou et al., 2022). However, there exists some general agreement that hate speech is usually defined as "language that targets a group with the intent to be harmful or to cause social chaos" and can be identified as a subset of offensive language (Sigurbergsson and Derczynski, 2020; Schmidt and Wiegand, 2017). On the other hand, offensive language, is a broader category containing any type of profanity or insult (Sigurbergsson and Derczynski, 2020). As the UA corpus is a collection of annotated tweets gathered from the social media platform, we apply a definition provided by Zampieri et al. (2019a), who determines that a message is offensive if it contains any form of foul language or a targeted offense, which can be stated implicitly

or explicitly. The targeted offense may be insults, threats, and posts containing obscene language.

Zampieri et al. (2019b) introduces general guidelines for offensive language identification, its types and targets. Waseem and Hovy (2016) attempt to give the most rigorous criteria of what should be considered as an offensive message. The researchers highlight ten main points of any offensive tweet: "1) it uses a sexist or racial slur; 2) it attacks a minority; 3) it seeks to silence a minority; 4) it criticizes a minority (without a well founded argument); 5) it promotes, but does not directly use hate speech or violent crime; 6) it criticizes a minority and uses a straw man argument; 7) it blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims; 8) it shows support of problematic hash tags; 9) it negatively stereotypes a minority; 10) it defends xenophobia or sexism; 11) it contains a screen name that is offensive". We add the direct citation from the article as it gives a thorough and concise summary of the Twitter's rules and policies sections on abusive, violent and hateful behaviour. The tweets that contain any marked characteristics become suspended.¹ We modify the criteria in the following way. A tweet is offensive if:

1. it promotes xenophobia, uses sexist or racist slur;
2. it implies the direct attack on a person or a group of people;
3. it promotes violence or abuse (overtly through the profound language or covertly);
4. it promotes misconception or misrepresentations that targets some violence or harm;

We further utilize the defined points as the guidelines for annotators.

2.3 Stages of the annotation process

There are three general scenarios for annotators selection: the subject-matter experts; individuals familiar with the subject background; and a crowdsourcing platform, where the annotators are only known after the process (Poletto et al., 2021). Our major challenge during the recruiting period was the war context. Regardless of whether a person

was inside Ukraine when the invasion started or outside – people perceive the atrocities of war similarly for many reasons: strong national identity, families or relatives that remain in Ukraine, etc. (Слюсаревський, 2022) Nevertheless, we decided to assess the geographical location of annotators at the time of data processing to minimize the probability of biased opinions. We do not exclude those who reside in Ukraine. As a result, 15 people familiar with the topic of war agree to participate voluntarily in the rating procedure. Each of them is provided with guidelines on what should be evaluated as an offensive tweet. Among the sample, 8 participants reside outside Ukraine, and 7 – stay in Ukraine; 5 of them are professional linguists with some prior experience in data annotations, and others are academics from different fields. Women prevail over men in the sample (12 vs. 3).

The whole annotation process is divided into three main iterations. The first stage includes 15 participants who manually annotate 300 tweets. Although the number of tweets is minimal, it is worth highlighting that the war is still ongoing, and new facts or crimes occur daily, which influence peoples' decisions. Moreover, the content of tweets is quite sensitive, which also impacts the general psychological sustainability of people to finish data annotation in one take. Considering the psychological factor, in the second stage, we strive to apply a pseudo-labeling technique (Arazo et al., 2020; Kuligowska and Kowalczyk, 2021) to tag a batch of 700 tweets by fine-tuning RoBERTa (Minixhofer et al., 2022) and ELECTRA (Schweter, 2020) for the Ukrainian language using the Keras library (Gulli and Pal, 2017). As the data is scarce, we obtain some inconsistent and biased labels; hence the three linguists who take part in the first stage and reside outside Ukraine are chosen to check the pseudo-labeled data.

Consequently, we get both manual and machine annotation. Repeating the automatic tagging process for the remaining 1043 tweets, we gather a sample of uncertain messages (tweets whose probability lay between 0.40 to 0.55). Hence, the same three annotators adjudicate the pseudo-labels.

Here we summarize the three iterations completed for the data annotation. Further, we provide more in-depth characteristics and results for every stage.

Results of the Stage I

The approach we acquire at the first stage of the

¹<https://help.twitter.com/en/rules-and-policies/abusive-behavior>, <https://help.twitter.com/en/rules-and-policies/violent-speech>, <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

data annotation is similar to the one described in the study by Ruitenbeek et al. (2022). The researchers offer three criteria for labeling the data, where the first option states "EXPLICIT" if the content expresses profanity unambiguously on the lexical level. The message is "IMPLICIT" when it lacks the overt lexical markers of offensive language. The option "NOT" is chosen when no offense is found. Instead of the three-layer annotation approach, we offer the raters to choose between four categories:

- Offensive language, offensive sense;
- Neutral language, neutral sense;
- Offensive language, neutral sense;
- Neutral language, offensive sense

These labels allow people to make more accurate judgments about the context. Tweets under the labels "Offensive language, offensive sense" and "Neutral language, offensive sense" are straightforward in their semantic manifestation, which can be conveyed through explicit or implicit markers. On the other hand, tweets that fall under the categories "Offensive language, neutral sense" and "Neutral language, neutral sense" carry no offensive meaning but can be externalized through some harsh or inappropriate language.

Fifteen people have completed the first iteration of data labeling. The number of participants appears to be significant; however, it has been agreed to keep a more extensive sample to achieve less biased results considering the nature of the material presented in the dataset. Selected annotators receive a link to the Google Form with a user-friendly interface and guidelines.

When the annotation process is completed, we access the spreadsheet with answers and extract statistics for each tweet. Some examples are presented in Figure 1 and Figure 2. Due to ethical policy we omit revealing the context of the tweets. 31% of participants correctly identify that the first tweet carries a neutral sense, whereas 25% have stated that the message from Figure 2 has no offense.²

Before proceeding to the second stage of the annotation procedure, we measure the IAA (the inter-annotator agreement) to evaluate the general quality of the acquired labels (Artstein, 2017). Since

²In our survey neutral or offensive "sense" equals neutral or offensive "meaning" of a tweet. These notions are used interchangeably here.

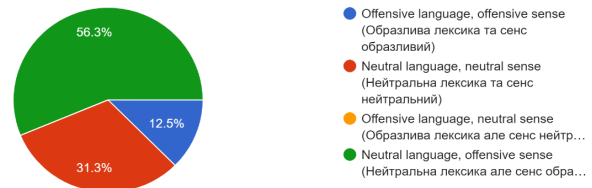


Figure 1: Explicitly neutral tweet that implies no offensive meaning.



Figure 2: Explicitly offensive tweet with profanity words and overtly offensive meaning.

15 participants took part in the labeling process, we used the Fleiss Kappa score to assess the IAA (Fleiss, 1971). Cohen's Kappa is useful if the number of annotators is no more than 2. Hence it is not applicable in our case (Cohen, 1960), while Krippendorff's alpha is more relevant for collections with some missing values (Krippendorff, 1980). We collapse four categories into offensive and neutral based on the tweet's meaning and utilize an open-source statistical Python library to calculate the Fleiss Kappa³. The inter-annotator agreement score at this stage is 0.384, indicating fair agreement between raters.

In the second iteration, we aim to utilize a pseudo-labeling technique for data annotation. This approach has demonstrated rigorous and consistent results in the computer vision domain (Iscen et al., 2019; Xie et al., 2020) and recently gained much attention in NLP research (Ahmed et al., 2011; Li and Yang, 2018). We follow a methodology offered by Kuligowska and Kowalczyk (2021), where authors use the DistilBERT model to distinguish questions from answers.

The 300 manually annotated tweets are applied for fine-tuning four neural network architectures:

1. DistilBERT (multilingual) (baseline model)
2. RoBERTa + BiLSTM

³https://www.statsmodels.org/dev/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

Results after the Stage I			
Model	Recall OF	Recall NON-OFF	F1 Macro
DistilBERT (multilingual)	.20	.95	.35
RoBERTa + BiLSTM	.86	.62	.65
ELECTRA + ReLU Dense layer	.83	.63	.73
ELECTRA + BiLSTM	.71	.82	.69
Results after the Stage II			
RoBERTa + BiLSTM	.59	.89	.65
ELECTRA + ReLU Dense layer	.82	.78	.76
ELECTRA + BiLSTM	.74	.80	.70
Results after the Stage III			
RoBERTa + BiLSTM	.60	.95	.69
ELECTRA + ReLU Dense layer	.73	.90	.72
ELECTRA + BiLSTM	.66	.96	.74

Table 2: Results after each stage.

3. ELECTRA + ReLU Dense layer

4. ELECTRA + BiLSTM

An exhaustive list of the pre-processing steps and hyperparameters is provided in Appendix A for replicability. DistilBERT is trained for 104 languages, so we do not expect it to perform well for this particular task.⁴ On the other hand, we apply a specific type of ELECTRA model (discriminator) trained solely on the Ukrainian data specifically for the text classification tasks⁵, anticipating it outperforms other architectures.

The models are trained on the train split (80%) and evaluated against the non-overlapping test split (20%). At this stage, the architectures are compared using the Recall scores of two classes. The choice of this metric is driven by the imbalance of the data, where 60% of the annotated tweets belong to the neutral class; hence the models are prone to overfit. The Recall score gives insight into the sensitivity or true positive rate prioritized at this stage. We utilize the F1 score for the final iteration as the number of samples increases. Besides, we opt for a more general statistical evaluation provided by the evaluation metric that measures the model’s accuracy. At this stage, the primary objective is to collect more or less solid probabilities for each tweet. Table 2 presents the Recall scores for each model.

Due to the scarcity of data, three out of four

models result in overfitting. ELECTRA + BiLSTM architecture has shown a more rigorous outcome compared to others. Nevertheless, we are unsure about the probabilities assigned for the unlabelled 700 tweets. Therefore, two raters from the previous sample of 15 people are chosen based on their professional training in linguistics and the 0.625 pairwise Cohen’s Kappa score after the first iteration (Landis and Koch, 1977), which indicates substantial agreement. The paper’s corresponding author serves as an adjudicator of the final label.

A batch of 700 tweets has been pre-annotated by the chosen model architecture and verified by three annotators. The second iteration results in a set of 1000 annotated and justified tweets.

Results of the Stage III

We prune the baseline model and evaluate the three transformer models using the same train/test split on the labeled 1000 tweets. Table ?? displays each architecture’s Recall and F1 (on the offensive) scores.

We utilize a simple ELECTRA + ReLU Dense layer model at this stage as it slightly outperforms the previous ELECTRA architecture. We also strive to minimize the overfitting of the BiLSTM layer at the previous stage. Correspondingly we get the probabilities for the 1043 unlabeled tweets. All tweets in the range of [0.40, 0.55] have been submitted for manual verification by the same three raters.

Subsequently, we have obtained the total annotated corpus of 2043 tweets partially labeled by the selected individuals and partially by the neural networks. The final Fleiss Kappa remains close to the one obtained in the second stage - 0.814.

⁴<https://huggingface.co/distilbert-base-multilingual-cased>

⁵<https://huggingface.co/lang-uk/electra-base-ukrainian-cased-discriminator>

The Table 2 describes the evaluation of the three neural networks on the annotated dataset. As we can conclude, the performance of each model has improved; models demonstrate robust and consistent results regardless of the number of performed iterations.

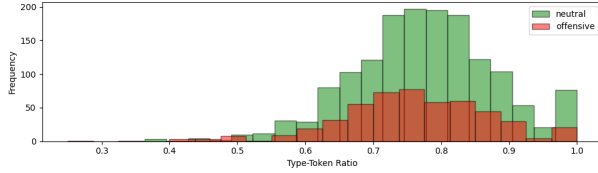


Figure 3: A distribution of type-token ratio in neutral and offensive subsets of the dataset.

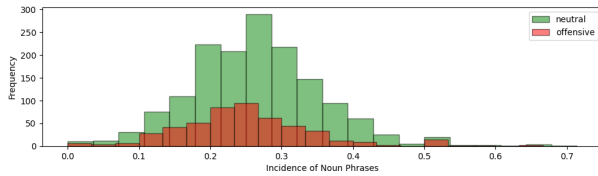


Figure 4: The 25 most frequent words in the subset of offensive tweets (Ukrainian language).

3 Data Statistics

The corpus is available as a public GitHub⁶ repository, enabling further research on offensive language detection and war rhetoric in the Ukrainian language. As Twitter's Terms & Conditions⁷ prohibit any public release of texts or metadata of tweets, we provide tweets' IDs and labels (1 – offensive; 0 – neutral). Scientists can use third-party tools such as Hydrator⁸ or Twarc⁹ to obtain the raw context.

The UA Corpus for Offensive Language Detection in the Context of the Russian-Ukrainian War incorporates 500 offensive tweets and 1543 neutral gathered from 1020 unique users. The Table 3 offers the English translation of the 25 most frequent words in the subset of neutral tweets. Subsequently, the Table 4 lists the 25 most frequent words from the offensive subset. The corresponding pie charts with the Ukrainian equivalents and word-clouds can be found in Appendix A.

⁶The link will be added after the blind review.

⁷<https://developer.twitter.com/en/docs/twitter-api/compliance>

⁸<https://github.com/DocNow/hydrator>

⁹<https://github.com/DocNow/twarc>

We can conclude that the term "the Armed Forces of Ukraine" is equally significant for offensive and neutral tweets. "Ukraine" is used more frequently in the neutral context rather than offensive. The word "Russia" dominates in the offensive tweets and ranks less for neutral. Noticeable that the top five words of the offensive tweets do not incorporate any obscene or profane language.

Word	Statistics
Ukraine (different declinations)	9.4%, 8.5%, 8.2%, 3.2%
The Armed Forces of Ukraine [3CY]	8.0%
glory	6.1%
war (different declinations)	5.1% , 4.1, 2.4%
RF, Russia (different declinations)	4.3%, 3.8%, 3.5%
people	4.1%
day	3.5%
life	3.1%
Ukrainians (different declinations)	3.1%, 2.8%,
the USA	2.8%
anti-aircraft warfare	2.6%
NATO	2.5%
rockets	2.3%
region	2.3%
victory	2.1%
country	2.1%

Table 3: The English translation of the 25 most frequent words among the neutral tweets.

Word	Statistics
Russians (different declinations)	9.9%, 5.1%, 4.3%, 4.1%, 3.7%, 2.8%
The Armed Forces of Ukraine [3CY]	6.1%
Ukraine (different declinations)	5.3%, 4.7%, 3.9%
war	5.1%
country	4.3%
people	4.1%
what	4.1%
glory	3.9%
go f**k yourself	3.7%
rockets	3.4%
want	3.4%
Putin	3.2%
fag***s	3.0%
dumb	3.0%
hate (verb)	2.8%
day	2.8%

Table 4: The English translation of the 25 most frequent words among the offensive tweets.

Moreover, we apply an open-source tool for corpus analysis - the StyloMetrix¹⁰ to extract grammatical features that help to set a boundary between the offensive vs. the neutral language of tweets (Okulska and Zawadzka). For instance, the Figure 3 indicates an aggregated type-token ratio of tweets. Even though the neutral tweets dominate in the dataset, their overall frequency is higher; we can still trace the tendency of offensive tweets to be slightly shorter than the neutral ones. Another example is the incidence of noun phrases (Figure 4).

¹⁰<https://github.com/ZILiAT-NASK/StyloMetrix#readme>

The mean value of NPs in the offensive subset is close to 0.25, whereas the mean of neutral NPs shifts closer to 0.3.

4 Related Works

4.1 Existing datasets on the Russian-Ukrainian war

Since the beginning of the full-scale Russian invasion, many corpora related to the war have been produced to research disinformation, warfare, misinformation, and political discourse. The datasets described in this section primarily aim to provide essential statistical evaluations of the texts related to the war.¹¹ The existing warfare corpora can be divided into two broad categories: multilingual and monolingual, mainly collected via Twitter’s streaming API¹² or other web-scraping tools. For instance, a dataset by [Chen and Ferrara \(2023\)](#) incorporates over 570 million tweets in more than 15 languages. The researchers offer a concise overview of their corpus’s top languages and keywords. Another publicly available multilingual dataset is "UKRUWAR22: A collection of Ukraine-Russia war related tweets," ([GHOSH, 2022](#)) comprising 55186 unique tweets in 57 languages. "Twitter dataset on the Russo-Ukrainian war" ([Shevtsov et al., 2022](#)) is a web-based analytical platform that daily updates the analysis of the volume of suspended/deactivated accounts, popular hashtags, languages, and positive/negative sentiment of tweets. A similar corpus by [Haq et al. \(2022\)](#) includes 1.6 million tweets; while the project is ongoing, it outlines the keywords assessment and language diversity. The listed multilingual datasets contain a profound amount of raw data that can be used to make broad statistical inferences about cross-language and sentiment analysis or as a tool for unsupervised data mining for topic detection, author identification, disinformation, and misinformation pattern extraction.

On the other hand, the monolingual corpora on the Russian-Ukrainian war essentially cover English sources and are significantly underrepresented for other languages. The online English dataset by the Social Media Labs¹³ gives a deeper insight into an alleged chemical attack in Mariupol. The focus was to construct the retweet network and to

identify Ukraine’s seven most retweeted accounts that broadcasted this topic. The corpus by [Fung and Ji \(2022\)](#) is a collection of over 3.5M user posts and comments in Chinese from the popular social platform Weibo. The gathered data can be a rich resource for propaganda and disinformation analysis in China. Another group of researchers created the Twitter dataset, which encompasses only original tweets in the English language, excluding retweets or quotes ([Pohl et al., 2022](#)). The data covers one week before the war and one week after the onset of the Russian invasion. "VoynaSlov" is a corpus that contains only the Russian language texts scraped from Twitter and a Russian social platform VKontakte ([Park et al., 2022](#)). The dataset includes 38M posts subdivided into two groups: state-affiliated texts and notes from independent Russian media outlets. The researchers state that the main objective is to use the obtained data to capture Russian government-backed information manipulation, which can be regarded as disinformation and propaganda.

Despite the plethora of datasets, most lack validation criteria that validate that the gathered texts are related to the topic of war. We assess this drawback while creating a well-grounded monolingual Ukrainian dataset. Moreover, the researchers highlight that the data scraped through the Twitter streaming API is not entirely random, which may result in some biases ([Shevtsov et al., 2022](#); [Pohl et al., 2022](#)). Unfortunately, we cannot escape this shortcoming as the presented dataset is collected through Twitter’s API.

5 Conclusions and Future Work

The study introduces the first Ukrainian dataset for offensive language detection in the context of the Russian-Ukrainian war. We propose a new method for annotating sensitive data using a pseudo-labeling algorithm with transformer models and human validation. In the first iteration, the annotators choose between four labels that capture tweets’ explicit and implicit offensive meaning. Then, the four labels are merged into two categories: offensive and neutral, depending on the context. We apply three main neural network architectures and obtain satisfactory results in the following two stages of data collection. The best-performing architecture in the second stage is ELECTRA + BiLSTM; however, it tends to overfit due to the small corpus size, which consists of only 300 tweets. Therefore,

¹¹<https://conflictmisinfo.org/datasets/>

¹²<https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data>

¹³<https://conflictmisinfo.org/>

we submit 700 automatically annotated tweets for verification to three annotators. In the last stage, we collect the logits from ELECTRA + ReLU Dense layer architecture. If the tweet's probability falls within [0.40, 0.55], its label is adjudicated by the raters. The final corpus comprises 500 offensive tweets and 1543 neutral tweets collected from 1020 unique users.

We present the descriptive statistics of the collected data by extracting the 25 most frequent words from each class and using the StyloMetrix tool to identify some grammatical features that differentiate offensive language from neutral language.

In future work, we plan to enlarge and balance the dataset and develop more robust neural networks for offensive language detection in Ukrainian. We also aim to apply the established criteria and terminology of offensive language to create a general Ukrainian multilabel dataset for abusive and hate speech detection.

Limitations

The dataset has a few limitations worth noticing:

1. **A human factor.** The collected tweets present the warfare content, and as the war is ongoing, people carry some bias, prejudice, and emotions that can influence any judgment, even professionally trained annotators. Hence, there remains room for bias in the obtained labels.
2. **Data labeling.** One can argue that hashtags and emojis play a significant role in data labeling. However, we eliminated them by explicitly mentioning to annotators not to consider them. We adhere to this rule because of the tweets' context. If people were to consider the hashtags, their opinion would have fluctuated even more, and in the end, we would not have achieved any rigorous and agreed annotation.
3. **Twitter API access.** In compliance with Twitter's rules and content-sharing policies¹⁴, we must provide only tweet IDs and labels, which can lead to data loss in further dehydration because some accounts can be suspended or banned at the time of content extraction. Besides, the Twitter stream rate limit may restrict

some content during the data scraping process, consequently bringing that bias to the corpus.

4. **Imbalance of data.** The number of neutral tweets is dominant in the dataset, which can cause incongruence in neural network models and limit their performance. We plan to balance the dataset during the following stages of its development.

Ethics Statement

Scientists who use this dataset need to understand the sensitivity of the context they aim to research. The inferences, conclusions, and statements they can make based on the content of the tweets may have a powerful influence on many people and their opinion on this war. Hence, the researchers need to be objective and rational while delivering their work. Moreover, one has to remember that collected tweets present only a small subset of the Twitter's data. Therefore, the bias and limitations have to be explicitly stated in their work.

Furthermore, we provide the content of tweets, excluding accounts' IDs, retweets, links, or any personal information, only upon explicit request and specifically to scientists for academic purposes. The academics granted the access should comply with our main conditions to not redistribute the corpus to third parties and not publish it as an open-source. Therefore, we satisfy Twitter's regulations on this issue.

References

- Mohammad Salim Ahmed, Latifur Khan, and Nikunj C Oza. 2011. Pseudo-label generation for multi-label text classification. In *CIDU*, pages 60–75.
- Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the effect of deplatforming on social networks. In *13th acm web science conference 2021*, pages 187–195.
- Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

¹⁴<https://help.twitter.com/en/rules-and-policies/twitter-rules>

- Uwe Bretschneider and Ralf Peters. 2017. [Detecting offensive statements towards foreigners in social media](#).
- Emily Chen and Emilio Ferrara. 2023. [Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between ukraine and russia](#).
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. [Detox: A comprehensive dataset for German offensive language and conversation analysis](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Holly Ellyatt. 2022. Russian forces invade ukraine. *cnn.com*. <https://www.cnn.com/2022/02/24/russian-forces-invade-ukraine>.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Yi R. Fung and Heng Ji. 2022. [A weibo dataset for the 2022 russo-ukrainian crisis](#).
- SATYAJIT GHOSH. 2022. [Ukrwar22: A collection of ukraine-russia war related tweets](#).
- Pierpaolo Goffredo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. [Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Antonio Gulli and Sujit Pal. 2017. *Deep learning with Keras*. Packt Publishing Ltd.
- Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Tristan Braud, and Pan Hui. 2022. [Twitter dataset for 2022 russo-ukrainian crisis](#).
- Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079.
- Abraham Israeli and Oren Tsur. 2022. [Free speech or free hate speech? analyzing the proliferation of hate speech in parler](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Klaus Krippendorff. 1980. Validity in content analysis.
- Karolina Kuligowska and Bartłomiej Kowalczyk. 2021. Pseudo-labeling with transformers for improving question answering systems. *Procedia Computer Science*, 192:1162–1169.
- J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.
- Ximing Li and Bo Yang. 2018. A pseudo label based dataless naive bayes algorithm for text classification with seed words. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1908–1917.
- Benjamin Minixhofer, Fabian Paischer, and Navid Rekasaz. 2022. [WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.
- Inez Okulska and Anna Zawadzka. Styles with benefits. the stylometric vectors for stylistic and semantic text classification of small-scale datasets and different sample length.
- Orestis Papakyriakopoulos and Ellen Goodman. 2022. The impact of twitter labels on misinformation spread and user engagement: Lessons from trump’s election tweets. In *Proceedings of the ACM Web Conference 2022*, pages 2541–2551.
- Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. [Challenges and opportunities in information manipulation detection: An examination of wartime russian media](#).
- Francesco Pierri, Luca Luceri, and Emilio Ferrara. 2022. [How does twitter account moderation work? dynamics of account creation and suspension during major geopolitical events](#).
- Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. [Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine](#). In *Proceedings of the 15th ACM Web Science Conference 2023*.
- Janina Pohl, Moritz Vincent Seiler, Dennis Assenmacher, and Christian Grimme. 2022. [A twitter streaming dataset collected before and after the onset of the war between russia and ukraine in 2022](#). Available at SSRN.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.
- Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. “zo grof !”: A comprehensive corpus for offensive and abusive language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 40–56, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Haji Mohammad Saleem, Jana Kurrek, and Derek Ruths. 2022. Enriching abusive language detection with community context. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 131–142, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Stefan Schweter. 2020. *Ukrainian electra model*.
- Alexander Shevtsov, Christos Tzagkarakis, Despoina Antonakaki, Polyvios Pratikakis, and Sotiris Ioannidis. 2022. Twitter dataset on the russo-ukrainian war.
- Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.
- Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTXEL with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Zeera Waseem, Thomas Davidson, Dana Warrmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.
- Zeera Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval).
- Микола Миколайович Слюсаревський. 2022. Соціально-психологічний стан українського суспільства в умовах повномасштабного російського вторгнення: нагальні виклики і відповіді. *Вісник Національної академії педагогічних наук України*, 4(1).

A Appendix A: Guidelines for reproducibility.

The data cleaning and pre-processing for the second and third iterations:

1. lowercasing of all words
2. all users' mentions were eliminated
3. all URLs were deleted
4. emojis were excluded
5. hashtags with the following were removed
6. extra blank spaces were replaced with a single space
7. extra blank new lines were removed

Models' hyperparameters:

- **DistilBERT (multilingual)**

Layers:

trf_model(input_word_ids, attention_mask)

Flatten layer

Dense layer (1, activation= 'sigmoid')

Hyperparameters: epochs = 4; batch size = 32; Adam optimizer with learning rate = 5e-5.

- **RoBERTa + BiLSTM**

Layers:

trf_model(input_word_ids, attention_mask)

SpatialDropout1D(0.3)

Bidirectional(LSTM(128,
return_sequences=True))

Dropout(0.5)

Bidirectional(LSTM(128,
return_sequences=True))

Dense(128, activation='relu')

Dense(1, activation='sigmoid')

Hyperparameters: epochs = 3; batch size = 16; Adam optimizer with learning rate = 3e-5.

- **ELECTRA + Dense layer with ReLU activation**

Layers:

trf_model(input_word_ids, attention_mask)

Flatten layer

Dense(128, activation="relu")

Dense layer (1, activation= 'sigmoid')

Hyperparameters: epochs = 3; batch size = 16; Adam optimizer with learning rate = 3e-5.

- **ELECTRA + BiLSTM**

Layers:

trf_model(input_word_ids, attention_mask)

SpatialDropout1D(0.3)

Bidirectional(LSTM(128,return_sequences=True))

Dropout(0.5)

Bidirectional(LSTM(128,return_sequences=True))

Dense(128, activation='relu')

Dense(1, activation='sigmoid')

Hyperparameters: epochs = 3; batch size = 16; Adam optimizer with learning rate = 2e-5.

A Appendix B: General statistics of the dataset.

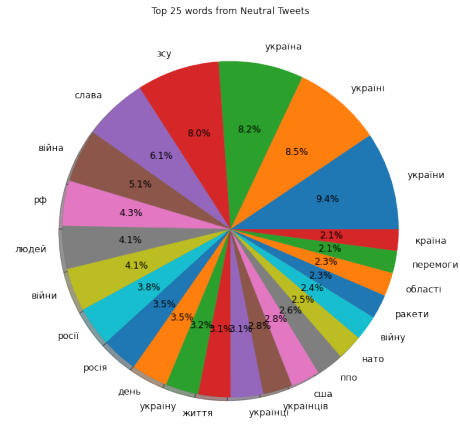


Figure 5: The 25 most frequent words in the subset of neutral tweets (statistics).

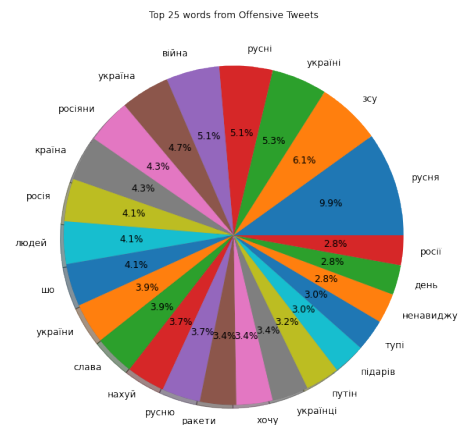


Figure 6: The 25 most frequent words in the subset of offensive tweets (statistics).