data

MDPI

Data Descriptor

Expert-Annotated Dataset to Study Cyberbullying in Polish Language

Michal Ptaszynski [1,*,†], Agata Pieciukiewicz [2,†], Pawel Dybala [3,†], Pawel Skrzek [4,†], Kamil Soliwoda [4,†], Marcin Fortuna [4,5,†], Gniewosz Leliwa [4,†] and Michal Wroczynski [4,†]

1 Text Information Processing Laboratory, Kitami Institute of Technology, Kitami 090-8507, Japan
2 Polish-Japanese Academy of Information Technology, 02-008 Warszawa, Poland
3 Institute of Middle and Far Eastern Studies, Faculty of International and Political Studies, Jagiellonian University, 30-059 Kraków, Poland
4 Samurai Labs, Aleja Zwycięstwa 96/98, 81-451 Gdynia, Poland; marcin.fortuna@samurailabs.ai (M.F.)
5 Institute of English and American Studies, University of Gdańsk, ul. Bażyńskiego 8, 80-309 Gdańsk, Poland
* Correspondence: michal@mail.kitami-it.ac.jp
† These authors contributed equally to this work.

Abstract: We introduce the first dataset of harmful and offensive language collected from the Polish Internet. This dataset was meticulously curated to facilitate the exploration of harmful online phenomena such as cyberbullying and hate speech, which have exhibited a significant surge both within the Polish Internet as well as globally. The dataset was systematically collected and then annotated using two approaches. First, it was annotated by two proficient layperson volunteers, operating under the guidance of a specialist in the language of cyberbullying and hate speech. To enhance the precision of the annotations, a secondary round of annotations was carried out by a team of adept annotators with specialized long-term expertise in cyberbullying and hate speech annotations. This second phase was further overseen by an experienced annotator, acting as a super-annotator. In its initial application, the dataset was leveraged for the categorization of cyberbullying instances in the Polish language. Specifically, the dataset serves as the foundation for two distinct tasks: (1) a binary classification that segregates harmful and non-harmful messages and (2) a multi-class classification that distinguishes between two variations of harmful content (cyberbullying and hate speech), as well as a non-harmful category. Alongside the dataset itself, we also provide the models that showed satisfying classification performance. These models are made accessible for third-party use in constructing cyberbullying prevention systems.

Keywords: cyberbullying; hate speech; abusive language; offensive language; toxic language; automatic cyberbullying detection; polish language

Citation: Ptaszynski, M.; Pieciukiewicz, A.; Dybala, P.; Skrzek, P.; Soliwoda, K.; Fortuna, M.; Leliwa, G.; Wroczynski, M. Expert-Annotated Dataset to Study Cyberbullying in Polish Language. *Data* **2024**, *9*, 1. https://doi.org/10.3390/data9010001

Academic Editors: Robertas Damaševičius and Keke Chen

Received: 25 July 2023
Revised: 2 December 2023
Accepted: 8 December 2023
Published: 20 December 2023

1. Background and Summary

While the issue of demeaning and defaming individuals through Internet-based communication has persisted since the inception of online communication, it was especially the emergence of novel handheld mobile devices, such as smartphones and tablet computers, that further accentuated this problem. These devices facilitate Internet access not only within the confines of homes, workplaces, or educational institutions but also during commutes, thereby exacerbating the challenge. Particularly noteworthy is the past decade, marked by the rapid proliferation of social networking services (SNS) such as Facebook and Twitter, which has brought into focus the prevalence of unethical conduct within online environments. This conduct notably impairs the mental well-being of adults and, more significantly, younger users and children. Central to this discourse is the phenomenon of cyberbullying (CB), characterized by the exploitation of open digital communication channels, including Internet forums and SNS, to disseminate harmful and distressing content about specific individuals, often targeting minors and students [1] (see

also section "Cyberbullying—a working definition" for further explanation of the problem of cyberbullying).

In response to this challenge, researchers worldwide have initiated an in-depth exploration of this issue, aiming to achieve an automated identification of such harmful online content. Such identified instances, when detected, need to be reported to SNS providers for in-depth analysis and subsequent removal. Accumulating knowledge spanning more than a decade [2–5], researchers have assembled a comprehensive foundation of understanding concerning this matter, predominantly pertaining to languages prevalent in highly developed nations, such as the United States and Japan. Regrettably, the same level of advancement remains notably lacking for less-resourced languages such as the Polish language. Through the dissemination of the dataset introduced herein and the preliminary formulation of classification models derived from said dataset, our objective resides in addressing and bridging this existing research gap.

The growing tendencies of polarization and radicalization in Poland, which started around 2015 [6] and have proliferated on the Polish Internet [7–9], necessitated the need for more active studies on the language of polarization among Polish researchers. This resulted in a number of datasets and shared tasks being proposed. The first attempt to study cyberbullying and hate speech in the Polish language with the use of data science, artificial intelligence, and natural language processing techniques was presented by Troszyński et al. in 2017 [10], with their dataset being released for a wider public in 2021[1]. The first open shared task focused on studying cyberbullying and hate-speech in a more open-science-based approach was the PolEval 2019 shared task on automatic cyberbullying detection in Polish Twitter [11,12]. The data collected for that task also became the basis for the improved dataset presented in this data descriptor. The early version of that dataset is also available on the HuggingFace datasets webpage[2]. Some data in the Polish language[3] was also included in the compilation of non-English hate-speech datasets compiled by Röttger et al. [13]. A more recent attempt was presented by Okulska et al. [14] in their BAN-PL dataset of banned harmful and offensive content collected from the Wykop.pl web service.

The dataset presented in this data descriptor, coupled with the included classification tools, furnishes an avenue for researchers and practitioners operating in the domain of AI to assess the efficacy of their proprietary classification techniques in determining the categorization of Internet entries within the realm of cyberbullying discourse. The dataset encompasses tweets sourced from openly accessible Twitter discussions. Given that a substantial portion of the challenge pertaining to automated cyberbullying detection often hinges upon the meticulous selection and crafting of features ([4,5]), the tweets are provided in their raw form with minimal preprocessing. Any preprocessing implemented is predominantly reserved for instances where the exposure of personal details about a private person could be revealed to the public.

The primary objective of employing the dataset centers on the classification of tweets, distinguishing them into two categories: cyberbullying/harmful and non-cyberbullying/non-harmful, with the utmost emphasis on achieving optimal precision, recall, F-score, and accuracy. An ancillary sub-task involves the differentiation between distinct forms of detrimental content, specifically cyberbullying (CB) and hate speech (HS), in addition to other forms of non-harmful material.

The remainder of this data descriptor is structured as follows. To begin, we expound upon the procedure by which the dataset was acquired. Subsequently, we explain the process of annotation, encompassing our operational definition of cyberbullying and the guidelines for annotation that underpinned the training of the annotators. Third, we conduct an exhaustive analysis of the generated dataset. This examination encompasses both comprehensive statistical analysis and more intricate instance-specific analysis. Following this, we detail the inaugural task in which the dataset was employed. Specifically, we delineate two classification tasks: (1) the overarching categorization of harmful content and (2) the differentiation between two distinct forms of harmful as well as non-harmful

content. Moreover, we put forth the default evaluation metrics and introduce exemplary AI models that were developed employing the dataset. Additionally, we present the dataset's original iteration alongside the refined annotations contributed by expert annotators. Concluding, we summarize this data descriptor and outline imminent plans and trajectories for advancing the dataset in the immediate future.

2. Methods

2.1. Data Collection

To collect the data, we applied the capabilities of the standard Twitter application programming interface (API)[4]. This API presented several inherent constraints, necessitating strategic workarounds. For instance, the API enforces limitations on the volume of requests permissible within a 15-minute interval, as well as the number of tweets accessible per individual request. Our data download procedure conscientiously respected the limitations. When the request limit was exhausted, the download script adeptly awaited the commencement of another download window.

Employing the python-twitter library[5] facilitated seamless interaction with the Twitter API. However, a distinct challenge emerged in the form of temporal constraints governing tweet searches. Within the confines of the standard Twitter API, users are allowed to retrieve tweets spanning the past 7 days only. Consequently, our efforts to compile responses to tweets emanating from our initial anchor accounts were impeded by this temporal restriction.

To archive and manage the data harvested from Twitter, we employed the pymongo library[6] to integrate it into MongoDB. Twitter provides tweet data in JavaScript object notation (JSON) format. This choice expedited the subsequent data handling processes with enhanced convenience.

The Python script was employed to retrieve tweets from nineteen designated official Polish Twitter accounts. These accounts were selected based on their prominence as the foremost Polish Twitter accounts during the year 2017[7]. The criteria for prominence encompass accounts with the largest followership, those exhibiting rapid follower growth, those accumulating substantial user engagement, those frequently mentioned in interactions, and those demonstrating prolific tweeting activities. Specifically, our initial focus was directed towards the subsequent set of accounts: @tvn24, `@MTVPolska`, `@lewy_official`, `@sikorskiradek`, `@Pontifex_pl`, `@donaldtusk`, `@BoniekZibi`, `@NewsweekPolska`, `@AndrzejDuda`, `@lis_tomasz`, `@tvp_info`, `@pisorgpl`, `@K_Stanowski`, `@R_A_Ziemkiewicz`, `@Platforma_org`, `@RyszardPetru`, `@RadioMaryja`, `@rzeczpospolita`, `@PR24_pl`.

Additionally, we gathered responses to tweets originating from the accounts referenced earlier (spanning the preceding 7 days). Our cumulative collection encompassed more than 101 thousand tweets procured from 22,687 individual accounts, as denoted by the `screen_name` attribute within the Twitter API. To initialize subsequent investigations, we adopted a selection process utilizing bash random functions, resulting in the random designation of ten accounts as the initial focal points.

Subsequently, adhering to the same methodology as previously outlined, we proceeded to acquire tweets from these ten designated accounts. Additionally, we procured all responses to their respective tweets, utilizing the Twitter search application programming interface (API), subject to a temporal constraint of the past 7 days. Through this iterative approach, we successfully amassed a corpus of 23,223 tweets originating from Polish accounts, earmarked for subsequent detailed analysis. The data procurement process culminated on the 20 November 2018. This collection of 23,223 tweets forms the foundational dataset that serves as the basis of the presentation within this paper.

2.2. Data Preprocessing and Filtering

Given that the original conversation threads were not traceable in our initial dataset, as the official API did not furnish such contextual information, we treated each tweet as an individual entity.

Initially, we employed a randomization procedure to alter the tweet sequence in the dataset. This was aimed to mitigate the potential anchoring bias [15] during annotations. The intent was to curtail the likelihood of human annotators assigning identical scores to multiple messages when encountering consecutive tweets from the same account.

Subsequently, tweets containing URLs were systematically excluded from the dataset. This was prompted by the recognition that URLs often occupy valuable character space within tweets, thereby constricting their textual content. In practice, this frequently resulted in tweets being truncated midway through sentences or featuring an abundance of *ad hoc* abbreviations. Next, tweets sharing identical content were systematically removed, a process that effectively eradicated a substantial portion of duplicates. Tweets solely comprising at-marks (@) or hashtags (#) were also removed. These elements, while serving as integral components of social media communication, lack inherent linguistic value as complete entities and instead function as detached keywords. To further refine the dataset, tweets containing less than five words were eliminated, as were those composed in languages other than Polish. These measures resulted in a remaining corpus of 11,041 tweets. From this collection, a subset of 1000 tweets was randomly extracted for utilization as a test dataset, while the remaining 10,041 tweets constituted the training dataset. This larger subset was earmarked for subsequent application in AI predictive models devised for the purpose of detecting cyberbullying. A comprehensive, step-by-step account of the preprocessing procedure and a thorough analysis outlining the attrition of tweets at each stage are presented below.

1. Removed tweets containing URLs, retaining solely the tweet text without any accompanying meta-data or timestamps. This resulted in a retention of 15,357 out of 23,223 tweets, representing 66.13% of the total.
2. Removed exact duplicates, leading to the retention of 15,255 tweets; 102 tweets were eliminated, constituting 0.44% of the entire dataset.
3. Removed tweets comprising solely @atmarks and #hashtags. Consequently, 15,223 tweets were retained, with 32 tweets, or 0.14% of the total, being removed.
4. Removed tweets that, apart from @atmarks or #hashtags, consisted solely of a single word, a few words, or emojis:
    (a) Removed tweets consisting of just one word, resulting in the retention of 14,492 tweets. A total of 731 tweets, corresponding to 3.1% of the dataset, were removed this way.
    (b) Removed tweets containing only two words, leading to the preservation of 13,238 tweets. The removal of 1254 tweets constituted 5.4% of the total.
    (c) Removed tweets comprising solely three words, yielding the retention of 12,226 tweets, and 1012 tweets, constituting 4.4% of the dataset, were deleted.
    (d) Removed tweets with only four words, resulting in 11,135 tweets being retained, and 1091 tweets, representing 4.7% of the total, were removed.

Following the aforementioned operations, our dataset comprised 11,135 tweets, characterized by a word count of five or more, with exclusions made for @atmarks or #hashtags.

The rationale underlying the exclusion of brief tweets was as follows.

1. A tweet of insufficient length poses a challenge for human annotators, as the limited contextual information hinders comprehensive assessment, thereby leading to a proliferation of ambiguous annotations.
2. Moreover, from the perspective of machine learning (ML) models, a broader spectrum of content (features) contributes to enhanced training. This, in turn, suggests that lengthier sentences offer a conducive environment for training more precise ML models. While it is plausible to conceive of brevity encompassing instances of aggression, we assume that a system trained on more extensive data will inherently encompass the shorter tweet instances as well.

Within the remaining subset of 11,135 tweets, a discernible fraction was composed in a language distinct from Polish, predominantly in English. To address this particular

issue, we harnessed the Text::Guess::Language Perl module[8], renowned for its capability to ascertain the language of a sentence by referencing the top 1000 words of that given language. The preliminary manual scrutiny of a limited tweet sample unveiled that the module occasionally yielded erroneous assessments, categorizing Polish tweets as Slovak or Hungarian. This was attributed to the atypical phrasings of account names and hashtags periodically incorporated in the tweets. However, its performance remained accurate in detecting English-authored tweets without misjudgment. Consequently, as a pragmatic approach, we opted to exclude all English-authored tweets, thereby preserving only the corpus of Polish-authored tweets. Following this ultimate preprocessing step, our dataset was distilled to encompass 11,041 tweets. From this refined collection, 10,041 tweets were selected for training purposes, while the remaining 1000 tweets were used for testing.

In tandem with the dataset release, we are unveiling the Perl scripts that facilitated the removal of English-written tweets from the Polish dataset (onlypolish.pl), alongside scripts for the curation of tweets solely comprising @atmarks or #hashtags (extractnohashatmarks.pl).

3. Data Records

The final list of all released files with their explanations was represented in Figure 1. Below, we describe how the dataset was created and explain the data records it contains.

file/folder name	explanation
/v1/	Folder containing original version of the dataset.
├ v1_test.tsv	Original test set.
├ v1_test_anonymized.tsv	Anonymized test set.
├ v1_training.tsv	Original training set.
└ v1_training_anonymized.tsv	anonymized test set
/v2/	Folder containing version of the dataset with improved annotations.
├ v2_test.tsv	Improved test set.
├ v2_test_anonymized.tsv	Anonymized improved test set.
├ v2_training.tsv	Improved training set.
└ v2_training_anonymized.tsv	Anonymized improved training set.
/scripts/	Folder containing scripts used in postprocessing of the dataset.
├ extractnohashatmarks.pl	Perl script to discard tweets that contain only @atmarks or #hashtags.
├ onlypolish.pl	Perl script used to discard tweets in English from Polish data.
├ preprocess_LEMSNERS.py	Python script to preprocess the dataset according to the second-best method for
│	the classic ML model.
└ preprocess_TOKPOS.py	Python script to preprocess the dataset according to the best method for the classic ML model.
/bert-base-polish-cyberbullying/	Folder containing all files of the best-performing classification model trained on the Polish cyberbullying dataset.
├ config.json	File containing configuration settings of the model.
├ pytorch_model.bin	Model file in binary form.
├ README.md	Readme file for the model.
├ special_tokens_map.json	Special tokens required by the model.
├ tokenizer_config.json	Original file required by the model.
├ training_args.bin	Binary file with training arguments for the model.
└ vocab.txt	Vocabulary file of the model.

Figure 1. Explanation of folder structure of all data records.

3.1. Data Annotation

3.1.1. Cyberbullying—A Working Definition

To create the annotation guidelines for the categorization of the acquired tweets, our initial step involved the formulation of a comprehensive working definition of cyberbullying. Notably, while a spectrum of overarching definitions concerning this issue exists, a consensus among the majority of these definitions (See: [1]) agrees to the following.

> Cyberbullying happens when modern technology, including hardware, such as desktop or tablet computers, or, more recently, smartphones, in combination with software, such as Social Networking Services (later: SNS, e.g., Twitter, Facebook, Instagram, etc.), is used in a repeated, hostile and, in many times, deliberate attempt to embarrass or shame a private person by sending messages, consisting of text or images, with contents that is malicious and harmful for the victim, such as, shaming the person's appearance or body posture, or revealing the person's private information (address, phone number, photos, etc.).

Moreover, social science research on cyberbullying [16] confirms the existence of both shared characteristics and distinctions between cyberbullying and in-person bullying. These dissimilarities contribute to the heightened complexity of addressing cyberbullying as an issue. The commonalities aligning it with bullying encompass the involvement of a **peer group** , mirroring the role of classmates in traditional bullying and the association of friends within social network groups. This often results in overlapping dynamics. Additionally, the **repetitiveness** of bullying behaviors, particularly in the digital realm, can be more frequent than in face-to-face interactions. The **imbalance of power** , wherein one individual or a small group becomes subjected to harassment from a disproportionately larger number of aggressors and their supporters, further parallels these phenomena.

An optimal approach would involve the capability to analyze data within a broader framework, such as the threading of conversations on the Twitter platform. Regrettably, the API's limitations hinder the possibility of conversation grouping. Consequently, for this dataset, each tweet is regarded in isolation. This methodology aligns with the preceding studies, wherein individual online entries were treated as distinct instances [1]. In the future, however, it is imperative to explore avenues for automated tweet clustering to facilitate the annotation of roles participants take in cyberbullying scenarios, encompassing the roles of victims, aggressors, and passive observers (supporters, defenders, etc.).

3.1.2. Annotation Guidelines

In order to enhance the efficacy of annotators in executing their task and to curtail the potential subjective influence stemming from individual annotators, we formulated a comprehensive set of guidelines tailored to guide the annotation process of harmful information in tweets. The guidelines encompass the subsequent categories. The diagram of all categories involved in the annotation process is represented in Figure 2.

Guidelines in English:

- Phishing, disclosure or threat of disclosure of private information (Tel. number, e-mail, address, account name, school name/number, class at school, private identification number (PESEL), credit card number, etc.)
- Personal attack ("Kill yourself, bitch!", etc.)
- Threats ("I will find you and I will kill you", etc.)
- Blackmail ("I will tell everyone where you live if you do not pay me", etc.)
- Mocking/ridiculing ("Look how fat this guy is", "you pimple-face", etc.)
- Gossip/insinuations ("Hey, apparently he's a zoophiliac!", etc.)
- The accumulation of profanity (single profane and vulgar words appear in conversations fairly often, but a longer "bundle" can be considered as harmful)
- Various combinations of all of the above

Guidelines in Polish:

- Wyłudzanie, ujawnienie lub groźba ujawnienia prywatnych informacji (Numer tel., e-mail, adres, nazwa konta, nazwa/numer szkoły, klasy, PESEL, karta kredytowa, itd.)
- Atak personalny ("Powieś się, gnoju!", etc.)
- Groźba ("znajdę cię i zajebię", etc.)
- Szantaż ("powiem wszystkim gdzie mieszkasz, jeśli mi nie zapłacisz", etc.)
- Szyderstwo/wyśmiewanie ("Patrzcie na tego grubasa", "ty pryszczata mordo", etc.)
- Plotka/insynuacja ("Ej, podobno to zoofil!", etc.)

- Nagromadzenie wulgaryzmów (pojedyncze występują dość często, ale ich nagromadzenie może być potraktowane jako niepożądane)
- Kombinacje powyższych

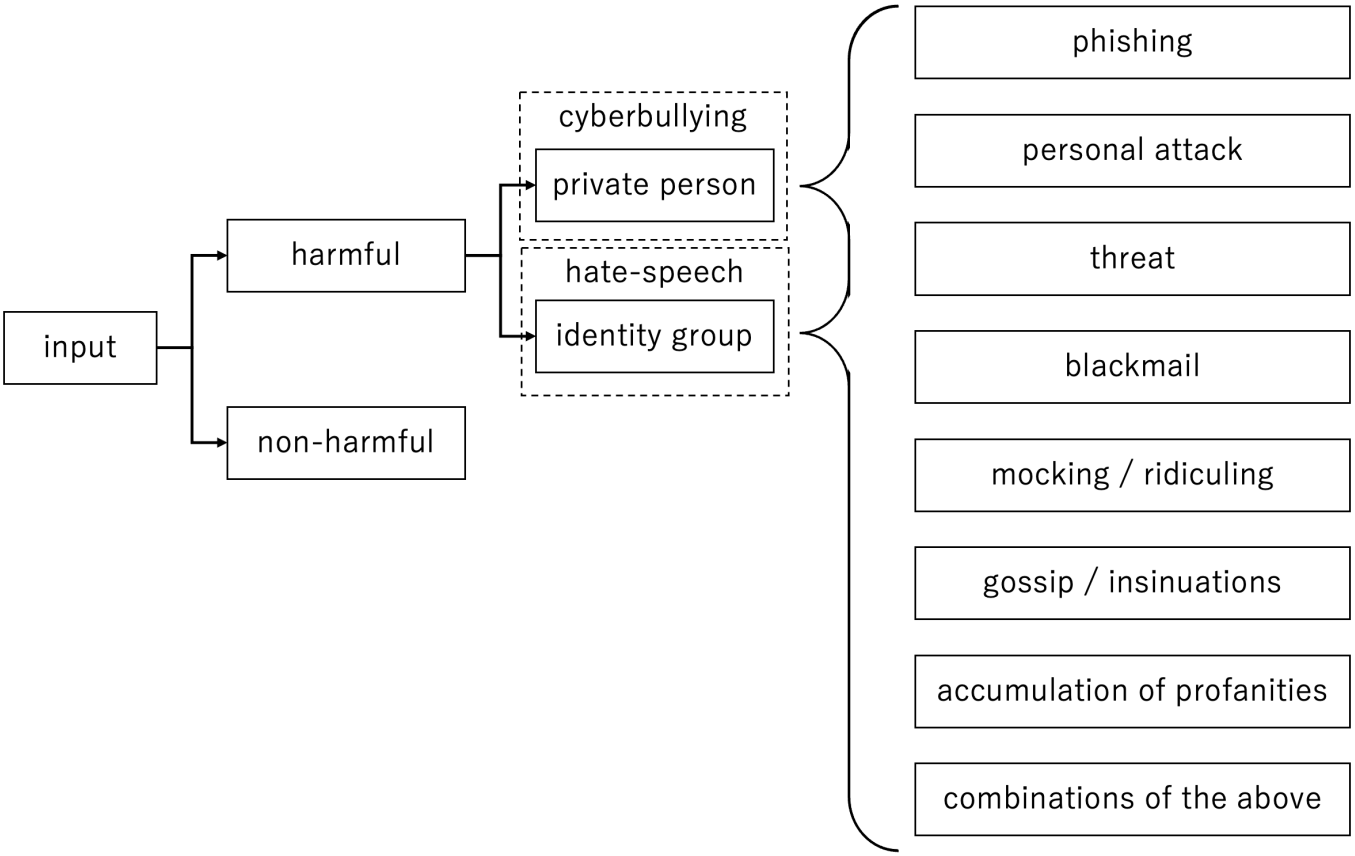

Figure 2. Diagram of all categories involved in the annotation process.

3.1.3. The Scope of the Collection of Tweets

In the context of cyberbullying, the focus predominantly centers on private individuals. As a result, our dataset exclusively comprises tweets derived from private Twitter accounts. Tweets originating from public accounts, such as those associated with politicians and celebrities, were intentionally omitted. This exclusion stems from the inherent potential exposure of public figures to critique and personal attacks due to the nature of their professions. Moreover, these public figures often intentionally engage in actions that may incite criticism as a strategy to augment their public visibility. While acknowledging that even public figures can experience private feelings of being offended, it remains noteworthy that individuals in the public eye possess a repertoire of resources to address such issues. This might encompass mechanisms such as deploying employees to flag instances of abuse within the Twitter platform, exerting influence through various channels, or even resorting to legal actions against hostile users.

3.1.4. Various Categories of Harmful Language

In spite of confining the search scope to the accounts of private persons, an inherent potential exists for the inclusion of injurious tweets directed towards public figures within said compilation. As a countermeasure, we decided to include within our annotations all tweets that do not embody instances of cyberbullying specifically, yet manifest other harm through alternative wordings. This encompasses instances of hate speech, racial prejudice, and sexism that do not target an individual or a small group (e.g., not referencing "you" or "a selected few from the class") but rather pertain to public figures or broader communal entities (e.g., "gays and lesbians" or "Paki" (Pakistanis)/"ciapaty" in Polish).

3.1.5. Annotation Process

The annotators were exclusively provided with the textual content of the tweets and proceeded to annotate on a tweet-by-tweet basis. Each tweet underwent annotation by a minimum of two and a maximum of three trained layperson annotators, accompanied by a single expert annotator. The pool of layperson annotators, comprising seven individuals, all of the female gender and within their early twenties, underwent training for the detection of cyberbullying and hate speech, following the guidelines elucidated in this section. The expert annotator, a male in his late thirties, possessing a decade of research experience in the realm of cyberbullying and cyberbullying detection, constituted the sole representative from the expert domain.

Subsequent to the completion of annotations by the layperson annotators, the expert annotator meticulously reviewed all annotations, subsequently endorsing or rectifying them. The annotations encompassed three distinct types of information: (A) harmfulness score, (B) specific tag if possible to specify, and (C) specific phrases if possible to specify in the text. We represent (A) and (B) in Tables 1 and 2 below.

Table 1. Harmfulness score used in annotation.

Score	Label Type
0	non-harmful
1	cyberbullying
2	hate-speech and other harmful contents

Table 2. Specific tag for harmful entries used in annotation.

Abbr.	Full Descr.	Explanation
pry	prywatne	disclosure or threat of disclosure of private information, phishing
atk	atak	personal attack
gro	grozba	threat
sza	szantaz	blackmail
szy	szyderstwo	mocking/ridiculing
plo	plotka	gossip/insinuations
wul	wulgaryzmy	accumulation of profanity and vulgarities
szy, wul, pry	(etc.)	various combinations of the above

3.1.6. Examples of Annotated Tweets

Displayed in Table 3 is a selection of examples included in the dataset. As the dataset encompassed tweets originating from diverse private accounts, the annotators underwent training to facilitate the annotation of tweets impartial to their political inclinations. Consequently, each tweet was appraised with a designated category of harmfulness. This assessment spans tweets exhibiting anti-right sentiments (e.g., tweet examples 2, 4, and 6), as well as anti-left sentiments (e.g., examples 5), as well as those with an indeterminate recipient (e.g., example 1). Notably, certain tweets were tainted with typographical errors (e.g., example 5, wherein "endekdu" erroneously supplants "endeku", stemming from "National Democracy supporter"; example 10, with "czulem" instead of "czułem" and "głow" instead of "głowa"). Evident within this corpus are tweets featuring profane vocabulary; however, their exclusion from the harmful category derives from their lack of targeted aim toward individuals or collectives (e.g., example 12, featuring "dupa", translating to "ass"). In contrast, a subset of tweets, despite the absence of explicit targets, embodies an endorsement of illicit substance consumption. As a result, these instances are characterized as harmful (e.g., example 3).

Table 3. Examples of tweets with annotated information.

No.	Tweet Contents	Score	Specific Tag	Phrases
1.	"Jebać, jebać, nic się nie bać." Eng.: *Fuck them, fuck them, fear nothing.*	2		jebać
2.	"Ty się oszuście i kłamco od nas odpierdol. My sobie damy radę bez twoich kłamstw" Eng.: *Fuck off you cheat and liar. We can do without your lies*	2	atk, wul, atk	oszuście, kłamco, odpierdol się
3.	"Picie to przyszłość, piękna i nieznana, szukajcie zapomnienia, próbujcie różnych trunków, łączcie środki psychoaktywne, Powodzenia" Eng.: *Drinking is the future, beautiful and unknown, seek oblivion, try different drinks, combine psychoactive drugs, Good luck*	2		
4.	"Ale ty jesteś żenujący. Dno metr mułu." Eng.: *You're such an embarrassment. A bottom with a meter of mud*	2		
5.	"Jaki hejt. Hejt to jest na Tuska, Komorowskiego, przecież to nie Polacy, śmierć wrogom ojczyzny. Obudź się" Eng.: *What hate. The hate is on Tusk, Komorowski, they are not Poles, death to the enemies of the motherland. Wake up*	2	gro	śmierć wrogom ojczyzny
6.	"Wio endekdu, ścierwa, zdrajcy, szubienica" Eng.: *Out you ND-supporters scum, traitors, gallows*	1	szy	ścierwa, endek, szubienica, zdrajcy
7.	"Jeszcze was zjemy i wysramy" Eng.: *We'll still eat you up and shit you out*	1	atk, gro	zjemy, wysramy, jeszcze
8.	"A ty wieś kretynie CONTI jest Acta2 i czego dotyczy? Najpierw przeczytaj a potem się wypowiadaj." Eng.: *And do you know you moron what is Acta2 and what does it apply to? Read first and then speak up.*	1	atk	kretynie
9.	"Ty pajacu, zmień sobie herb na pusty łeb." Eng.: *You clown, change your coat of arms to an empty head.*	1	atk	pajacu, pusty łeb
10.	"jak ja się źle czulem jak byłem dzieckiem w kościele to głow mala, szopka do kwadratu, nie mogłem tego wytrzymać" Eng.: *how I felt so bad when I was a child in the church I just cannot wrap my head around it, circus squared, I could not stand it*	0		
11.	"Kiedy Christina wychodzi za mąż" Eng.: *When Christina is getting married*	0		
12.	"kot też się załapał na fotke, a raczej jego dupa :)" Eng.: *the cat was also caught in the photo, or rather his ass*	0		

3.2. Dataset Analysis and Discussion

3.2.1. General Statistical Analysis

A general statistical overview of the dataset was represented in Table 4. The dataset encompassed a total of 11,041 tweets, of which 10,041 were allocated to the training set and 1000 to the test set. The evaluations conducted by layperson annotators exhibited a significant concurrence in most annotations, yielding an overall agreement rate of 91.38%. A minimal subset of tweets, numbering 84 (0.76%), remained untagged by either annotator. This percentage of concurrence is notably high. However, it is noteworthy that this high agreement predominantly emerged from the non-harmful tweet category, constituting a substantial portion of the dataset (approximately 89.76%). Within the subset of harmful tweets, the annotators achieved complete consensus for designating the "cyberbullying" class in

only 106 instances (0.96%) and the "hate-speech" class in only 73 tweets (0.66%). Interestingly, even among the tweets characterized by unanimous agreement, several instances were subsequently reassigned to different classes following expert annotator review.

Table 4. General statistics of the dataset.

	#	% of All	% of Set
Overall # of Tweets	11,041	100.00%	
# of Tweets Annotator 1 was unable to tag	38	0.34%	
# of Tweets Annotator 2 was unable to tag	46	0.42%	
# of Tweets where Annotators agreed	10,089	91.38%	
# of Tweets where Annotators agreed for 0	9910	89.76%	
# of Tweets where Annotators agreed for 1	106	0.96%	
# of Tweets where Annotators agreed for 2	73	0.66%	
# of Tweets where Annotators disagreed	952	8.62%	
# of final 0	10,056	91.08%	
# of final 1	278	2.52%	
# of final 2	707	6.40%	
# of all harmful	985	8.92%	
Training set	10,041	90.94%	
# of final 0	9190	83.24%	91.52%
# of final 1	253	2.29%	2.52%
# of final 2	598	5.42%	5.96%
# of all harmful	851	7.71%	8.48%
Test set	1000	9.06%	
# of final 0	866	7.84%	86.60%
# of final 1	25	0.23%	2.50%
# of final 2	109	0.99%	10.90%
# of all harmful	134	1.21%	13.40%

In general, the individuals trained as layperson annotators exhibited a reasonable degree of confidence in categorizing tweets as non-harmful, even when these tweets incorporated explicit language. Unfortunately, the layperson annotators showed a much lower level of capability in identifying the specific elements of harmfulness or bullying. This observation substantiates the necessity of expert annotations in tackling specific issues such as cyberbullying. This notion was underscored by the extensive research of Ptaszynski and Masui (2018) [1], spanning a decade, despite the prevalence of studies employing laypeople en masse, including undergraduate students or Mechanical Turk workers, for annotation purposes [17,18]. In terms of comparing the training and test datasets, the latter exhibited a slightly elevated proportion of harmful tweets (8.48% for the training set vs. 13.40% for the test set). Furthermore, there were 82 instances wherein one or both annotators encountered challenges in providing annotations. Subsequently, these instances were addressed by the expert overseeing the annotation process.

As an additional analysis, we compared the top thirty words from the harmful and non-harmful groups separately, extracted from the training dataset using standard term frequency with inverse document frequency (TF-IDF)[9]. Predictably, for the non-harmful group, none of the words with the highest TF-IDF were harmful, with the majority of them being related to soccer, due to this topic being prevalent in the discussions within the extracted messages. On the contrary, for the harmful group, out of thirty top scoring words, only three were not directly related to harmful context ("kredyt", eng. *loan* and "PAN", referring to *Polish Academy of Sciences*, "własne" eng. [your] *own*). Most were direct slurs and vulgarities ("pajacu" eng. [you] *moron*, "debil" eng. *idiot*, "debila" eng [of that] *idiot*, "świnia", eng. *pig*,), words that although not being harmful by definition in practice, are usually used in harmful context (e.g., "psychiatry", eng. *psychiatrist* often used in the context: [you should go see a] *psychiatrist*, or "IQ", often used in such contexts as [you have a low] *IQ*), or words with a specific negative political connotation ("pisowska" or

"pisowski", eng. [you]*PiS-supporting*- [ADJ]), or addressing a public person being the object of a slur ("Bolka" which in Polish alt-right extremist messages refers to former Polish president Lech Walesa, or "Gasiuk" referring to Kamila Gasiuk-Pihowicz, a lawyer and a politician of the Civic Platform party). The top thirty words for the harmful and non-harmful group extracted from the training dataset using TF-IDF are represented in Table 5. A visualization of these statistics was represented in wordclouds in Figure 3.

Figure 3. Word clouds of the top thirty most significant harmful (**right**) and non-harmful (**left**) words/tokens according to tf-idf.

Table 5. Top thirty words for harmful and non-harmful group extracted from training dataset using TF-IDF.

Top #	Harmful	TF-IDF Score	Non-Harmful	TF-IDF Score
1	pajacu	0.0007611	XD	0.000482
2	debil	0.0004497	😜	0.000462
3	IQ	0.0004151	:-)	0.000319
4	debila	0.0003805	jeśli	0.000306
5	kradli	0.0003460	:(	0.000258
6	psychiatry	0.0003114	?:)	0.000258
7	debilu	0.0002422	rok	0.000238
8	kłamco	0.0002422	miałam	0.000231
9	pisowska	0.0002422	😎	0.000231
10	tępa	0.0002422	!!!	0.000224
11	bura	0.0002076	??	0.000217
12	jebie	0.0002076	U	0.000211
13	kutasie	0.0002076	wykonawcy	0.000211
14	leczyć	0.0002076	klubu	0.000197
15	Bolka	0.0001730	Skoro	0.000183
16	idiotę	0.0001730	kontrakt	0.000183
17	kredyt	0.0001730	klub	0.000177
18	parchu	0.0001730	super	0.000177
19	pisowski	0.0001730	wg	0.000177
20	profilowym	0.0001730	TW	0.000163
21	swastyka	0.0001730	trenera	0.000163
22	własne	0.0001730	trzy	0.000163
23	zjebie	0.0001730	źle	0.000163
24	łże	0.0001730	trener	0.000156
25	świnia	0.0001730	💪	0.000156
26	Gasiuk	0.0001384	???	0.000149
27	Kamilo	0.0001384	zdjęcie	0.000149
28	Katynia	0.0001384	Dzięki	0.000143
29	Ona	0.0001384	ani	0.000143
30	PAN	0.0001384	piłki	0.000143

3.2.2. Kappa Coefficient with Modified Quadratic Weights for Inter-Annotator Agreement in Cyberbullying Scenario

To obtain a better grasp on how well the annotators agreed with one another, we calculated Cohen's Kappa coefficient [19] among all pairs of annotators. However, standard Kappa assumes that all categories are unrelated, which is untrue in our case. Therefore, we used weighted Kappa [20], in which an ordered distance is assumed between the categories, which is then used to calculate class weight. Moreover, we specifically used weighted Kappa with not linear but quadratic weights [21], because the distance between category 1 (cyberbullying) and category 2 (hate speech) is much smaller than each of them to 0 (non-harmful). However, this way Kappa would become too high, as these settings would assume that there is at least some closeness between category 0 and one of the harmful categories. In fact, the weight for category pairs 0 and 1 would be the same as for category pairs 1 and 2, which is highly unrealistic. Therefore, to account for the fact that (1) the two harmful categories are closer to each other than the non-harmful category and (2) the non-harmful category is in fact unrelated to the other two, we modified the weights of the category pair 0 and 1. The difference between the standard and modified quadratic weights is represented in Table 6.

Table 6. Standard and modified quadratic weights. The modification compared to the standard quadratic weights was highlighted in bold red font.

		Standard Quadratic Weights			**Modified Quadratic Weights**		
		Categories			**Categories**		
		0	1	2	0	1	2
categories	0	1	**0.75**	0	1	**0**	0
	1	**0.75**	1	0.75	**0**	1	0.75
	2	0	0.75	1	0	0.75	1

The inter-annotator agreements of annotators of the dataset, together with all confusion matrices required for reproducibility, are represented in Table 7. The Kappa values, regardless of whether they were standard Kappa, weighted Kappa, or the proposed Kappa with modified quadratic weights, were around 0.3, which suggests *fair agreement*. It is not high, however, such lower agreements are expected for laypeople, as mentioned in previous studies [22]. This is especially true for multi-class tasks, which are also difficult, such as the annotation of cyberbullying and other harmful and harm-related data.

Table 7. Inter-annotator agreements for annotators of the dataset.

		Observed				**Expected**			
		Annotator 1				**Annotator 1**			
	class	0	1	2	sum	0	1	2	sum
Annotator 2	0	9845	70	191	10,106	9550.86	222.24	332.90	10,106
	1	199	106	97	402	379.92	8.84	13.24	402
	2	313	65	73	451	426.23	9.92	14.86	451
	sum	10,357	241	361	10,959	10,357	241	361	10,959
								Kappa =	0.325
								weighted Kappa =	0.329
								weighted Kappa with modified quadratic weights =	0.378

3.2.3. Discussion on Specific Tweet Examples

The annotation process yielded a spectrum of valuable insights as documented by the annotators. The annotators often identified that the semantic nuances of the majority of tweets were contextually dependent. When contextual clarity was lacking, the assessment of

these tweets within the designated categories—particularly the harmful category—proved challenging. To mitigate this issue, a comprehensive analysis of the complete interaction between Twitter users was necessary. Such an analysis would illuminate the contextual nuances in which a specific tweet was disseminated. To address this issue, the tweets could be clustered within conversational threads. In the future, our focus will include the formulation of an automated technique for such a coherent clustering of tweets into meaningful threads. This objective can be realized through the incorporation of distinct meta-information. Specifically, the inclusion of details concerning the target tweet to which a given message is directed (facilitated by the API's in_reply_to_status_id) or capitalizing on user quotations (@user) that commonly preface tweets, often serving as responses, together with the temporal intervals between successive tweets, could be harnessed to enhance the confidence levels associated with identifying response tweets.

When addressing tweets pertaining to authoritative figures or public personalities, instances in which a tweet solely conveyed an opinion without involving derogatory language or defamation were predominantly categorized as non-harmful by most annotators. This classification stemmed from the prevailing societal understanding that the mere expression of an opinion is not inherently subject to punitive measures. The annotators also emphasized the necessity for maintaining a discerning perspective, one that can distinguish personal convictions from critiques aimed at authoritative entities, thereby ensuring objectivity throughout the annotation process. This can be achieved through the systematic retraining of annotators or through the assignment of task-specialized annotators adept at impartial analysis. Furthermore, although clear contrasts were observable between linguistic attributes used by proponents of specific political orientations (e.g., "lemingi"/"lemmings" or "lefties" versus "pisiory"/"PiS-supporters" or "right wingers"), overarching linguistic patterns emerged on both ideological sides, transcending political subjects.

3.2.4. Examples of Tweets with Additional Explanations of Reasoning Behind Annotation

Not Harmful

"500+ bardzo na plus jednak ten rzad wykorzystał dorobek poprzednich rządów do swojego populizmu chorego"
Eng.: *500+ is a plus, however, this government has used the achievements of previous governments for its sick populism*
"Mamy do czynienia z najgorszym prezydentem RP w historii. Kropka."
Eng.: *We are dealing with the worst president of the Republic of Poland in history. Period.*

Both above samples considered a general opinion. Score: 0.

"I kurwa mamy ta wolność"
Eng.: *And we fucking have this freedom*

Although the utilization of explicit language ("kurwa") might suggest a robustly offensive tone, the phrase inherently lacks any indication of harmful conduct. Consequently, the tweet was categorized as non-harmful. Score: 0.

"Matka Boska była półką i Jezus też."
Eng.: *The Mother of God was a shelf* [misspelling of "Polish"] *and so was Jesus.*

While initially appearing as a blasphemous statement, the consequential detrimental impact primarily stems from a spellchecker malfunction (with "Polka" being erroneously corrected to "półka"). Score: 0.

"Biało-Czerwoni brawo, brawo, brawo! Zbigniew Boniek i Adam Nawałka - wyrazy szacunku. Robert Lewandowski-wielkie podziękowania!
Eng.: *Bravo White-Reds, bravo, bravo! Zbigniew Boniek and Adam Nawałka-respect. Robert Lewandowski-many thanks!*
"WISŁA KRAKÓW !! brawo za dzisiejszy mecz :)"
Eng.: *WISLA KRAKOW [soccer team name] !! Bravo for today's match :)*

Score: 0.

Cyberbullying

"[tel. no. anonymized] w Bułgarii numer ten uważany jest za przeklęty ponieważ podobno każdy z jego właścicieli umierali po kilku dniach"
Eng.: [tel. no. anonymized] in Bulgaria, this number is considered cursed because apparently each of its owners died after a few days

Regarded as conceivably falling within the purview of phishing endeavors, an attempt to validate if a provided telephone number is in fact cursed, it introduces the potential for unknowingly endangering oneself by calling the number. This, in turn, elevates the vulnerability to personal information compromise. Therefore the attributed risk factor is quantified with a score of 1.

"Tu stary chuju PZPRowski zajmij się swoimi komuchami z PiSu."
Eng.: *You old PZPR prick, go to your PiS commies.*

A common assault involving allegations of endorsing communism is observed. Although the inception of this incident stems from a reaction to a publicly accessible profile, its manifestation appears to be targeted towards an individual recipient. Thus, a quantified assessment yields a score of 1.

Hate-Speech/Other Harmful

"Rozumiem, że jutro w sejmie powie to pani protestującym. Załgane pisowskie skurwysyny."
Eng.: *I understand that you will tell that to the protesters tomorrow at the meeting of the Diet. Lying PiS motherfuckers.*

Accumulation of profanity. Score: 2.

"Was, gnidy powinno się zaorać na metr w głąb i grubo posypać niegaszonym wapnem. A dla pewności zbombardować napalmem."
Eng.: *You, scum, should be plowed a meter deep and sprinkled coarsely with quicklime. And to be sure bombard with napalm.*

A typical case of hate-speech consisting of over-exaggerated death threats aimed at a public person. Score: 2.

"KAIN TEŻ ZABIŁ BRATA ALE NIE ŚWIĘTOWAŁ TEGO CO MIESIĄC I NIE STAWIAŁ POMNIKÓW NA TĘ OKOLICZNOŚĆ."
Eng.: *KAIN KILLED BROTHER AS WELL BUT DIDN'T CELEBRATE EVERY MONTH, AND DIDN'T ESTABLISH MONUMENTS FOR THIS CIRCUMSTANCE.*

An illustrative case of contextually dependent derisive behavior targeting a prominent public figure is observed. Although the specific identity of the subject of mockery remains implicit, rendering computational inference challenging, it readily resonates with an adult audience that is well-versed in the political landscape of Poland. The provided score for this instance stands at 2.

"MILIONY POLAKÓW CZEKA NA BADANIA PSYCHIATRYCZNE LISA PO WPISACH WIDAĆ NIE ZRUWNOWARZENIE PSYCHICZNE I CIĄGŁĄ DEPRESJE"
Eng.: *MILLIONS OF POLES ARE WAITING FOR LIS'S PSYCHIATRIC TESTS, AFTER THE ENTRIES, THERE CAN BE SEEN MENTAL UNEASE AND CONSTANT DEPRESSION*

The Twitter post, initially targeted at an identifiable public figure (a television presenter), transgresses privacy boundaries and may be categorized as a form of public defamation. Additionally, the reference to the necessity for psychiatric evaluation and the invocation of psychological ailment (specifically, depression), conditions typically evaluated and diagnosed by a qualified psychiatrist, trespasses into the realm of personal affairs. In this context, these references are employed pejoratively. Evaluation Score: 2 (defamation of a public figure).

"UK jest coraz bardziej faszystowska, ale widocznie jeszcze nie wystarczająco faszystowska aby powitać"
Eng.: *The UK is increasingly fascist, but apparently not yet fascist enough to welcome you*
"Elo swastyka na ryju kiedy będzie, sorry że ciągle pytam?"
Eng.: *Howdy, swastika in your pig-face-when? sorry I keep asking.*

While the primary content of the tweet predominantly conveys a broad unfavorable sentiment directed towards the United Kingdom, it concurrently centers on an individual of particular significance (initially identified as a conservative public figure), attributing to this individual the label of "a fascist." The assigned score for this instance is 2, indicative of the act of ascribing a public figure with the accusation of espousing fascist ideologies.

"Lzy ogromne, kiedyś usunąłem ciążę, nie mów nikomu"
Eng.: *Enormous tears, once I terminated my pregnancy, do not tell anyone*

Regarded as a harmful Twitter post crafted with the intent of eliciting provocation, as evidenced by the use of the phrase "*nie mów nikomu*" (translated: "don't tell anyone"), given the tweet's accessibility to the general public. This instance is situated within a socially contentious subject matter, namely, abortion. The assigned score for this tweet is 2.

4. Technical Validation

4.1. Description of Tasks

The goal of the preliminary task designed for the dataset is to classify if an online entry should be considered harmful (cyberbullying, hate-speech) or non-harmful. The specific objective revolves around categorizing the supplied tweets into the domains of cyberbullying/harmful and non-cyberbullying/non-harmful, aiming to attain optimal precision, recall, balanced F-score, and accuracy metrics. This task was further divided into two distinct sub-tasks.

4.1.1. Task 1—Harmful vs. Non-Harmful

Within this task, the goal is to classify tweets of a normal/non-harmful nature (class: 0) and tweets containing various forms of harmful content (class: 1), which encompasses cyberbullying, hate speech, and other interconnected phenomena.

4.1.2. Task 2—Type of Harmfulness

In this task, the goal is to classify three tweet categories, denoted as follows: 0 (non-harmful), 1 (cyberbullying), and 2 (hate-speech). Notably, a range of definitions existed for both cyberbullying and hate-speech, with some even classifying these phenomena under a common category. The precise criteria employed for annotating instances of cyberbullying and hate-speech have been meticulously developed over the course of a decade's worth of research (see [1]). However, the primary and definitive criterion for differentiation hinges on whether the harmful message is directed at a private individual or individuals (cyberbullying) or towards a public individual, entity, or larger group (hate-speech). The supplementary distinct definitions and guidelines implemented in the creation of the dataset are explained in preceding sections.

4.2. Evaluation Measures

The scoring criteria for the first task were based on standard Precision (P), Recall (R), balanced F-score (F_1), and Accuracy (A), calculated according to the following Equations (1)–(4), respectively, with the conditions of being the highest in terms of true positives (TP), true negatives (TN), and the lowest in terms of false positivse (FP) and false negatives (FN) being considered the winner. If the F-scores were equal for more than two compared models, the one with higher accuracy would be declared better. Furthermore, the results closest to the break-even point of precision and recall are given priority.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The evaluation of the second task's performance was grounded in two metrics: the micro-average F-score (microF) and the macro-average F-score (macroF). The calculation of the micro-average F-score aligns with the conventional F-score equation, yet is based on micro-averaged precision and recall, calculated according to the Equations (5) and (6), respectively. In a parallel manner, the macro-average F-score is computed, leaning on the macro-averaged precision and recall, calculated according to Equations (7) and (8), respectively. The criterion for a win is primarily obtaining the highest microF. This metric ensures equitability by treating all instances uniformly, where the number of samples across classes could differ. In situations where a model yields identical microF outcomes, precedence is awarded to the model showcasing superior macroF values. The supplementary macroF, while not uniformly treating instances, proffers deeper insights by considering the equality of all classes.

$$P_{micro} = \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$R_{micro} = \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i} \tag{6}$$

$$P_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i} \tag{7}$$

$$R_{macro} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i} \tag{8}$$

Macro F-score and Micro F-score were not calculated in a standard way. Typically, Macro F-score is calculated based only on all the F-scores, while Micro F-score is calculated only on the basis of the sums of all samples. This way there is only insight into P and R for each class (macroF) or overall (microF), and the scores are grouped as F-scores. This hinders the more general insight into the relations between P and R. Thus, instead, we used the proposed method of calculation specifically to provide insight into how close each score was to the break-even point in case of similar F-scores and accuracies. However, despite the difference in the calculation, the overall tendency in the final macroF and microF remains the same as for the standard way of calculation of those scores, thus assuring fairness in evaluation.

4.3. Initial Task Participants

A total of fourteen submissions were received for the initial installation of the challenge. All the participating groups endeavored to address the initial task, characterized as computationally less challenging and involving the binary classification of tweets into harmful and non-harmful categories. In contrast, the subsequent task, entailing a three-class classification problem, garnered a more limited engagement with only eight endeavors. Herein, we provide a concise overview of the methodologies introduced by each participating team.

4.4. Results

4.4.1. Baselines

The dataset exhibited an imbalance in class distribution, resulting in varying proportions for each class (refer to Table 4). To ensure an objective assessment of the task participants' model performance in classification, we initially established a series of baselines.

The initial set of baselines encompassed basic classifiers that generated scores not based on any data-specific insights.

A. Classifier only assigning a score of 0.
B. Classifier only assigning a score of 1.
C. Classifier only assigning a score of 2 (restricted to Task 2).
D. Classifier only random scores: 0 or 1 (for Task 1).
E. Classifier only random scores: 0, 1, or 2 (for Task 2).

Consequently, the performance of all baseline classifiers exhibited low performance. Concerning Task 1, baseline A (constant value of 0) yielded an F1 score of 0, a foreseeable outcome, confirming that it is not possible to consider the problem as too simple. Baseline D (random prediction) similarly obtained an F1 score of 0, additionally confirming that it is not possible to solve the task of cyberbullying detection by depending on mere chance. Baseline B (constant value of 1), by its inherent nature, effectively captured the entirety of harmful instances, manifesting a recall of 100%, albeit at the cost of an exceedingly low precision (13.4%), consequently yielding a considerably diminished F-score (23.63%). The outcomes for the simple baselines for Task 1 are reported in Table 8.

Table 8. Results of simple baselines for Task 1.

Task 1	P	R	F1	A
Baseline A	0.00 %	0.00 %	0.00 %	86.60 %
Baseline B	13.40 %	100.00 %	23.63 %	13.40 %
Baseline D	0.00 %	0.00 %	0.00 %	86.60 %

Concerning the second task, analogously to task 1, baselines B (constant 1), C (constant 2), and E (random) attained notably low performance scores. Baseline A (constant 0) achieved a substantial microF score (86.6%), primarily due to its automatic success in non-harmful instances, which constituted the dataset's majority. However, macroF, serving as a more comprehensive evaluation metric, demonstrated a considerably diminished value (30.94%). The outcomes pertaining to the elementary baselines for Task 2 are documented in Table 9.

Table 9. Results of simple baselines for Task 2.

Task 2	microF	macroF
Baseline A	86.60 %	30.94 %
Baseline B	2.50 %	1.63 %
Baseline C	10.90 %	6.55 %
Baseline E	31.20 %	31.16 %

4.4.2. Results for Task Participants

Task 1

In Task 1, among fourteen initial submissions, a total of nine distinct teams emerged: n-waves, Plex, Inc., Warsaw University of Technology, Sigmoidal, CVTimeline, AGH and UJ, IPI PAN, UWr, and an independent entity. Certain teams opted for multiple system proposals, notably Sigmoidal (3 submissions) and the independent entity (3), alongside CVTimeline (2). The participants harnessed an array of techniques, frequently drawing upon readily available OpenSource solutions that were then adapted and trained to align with the Polish language and the provided dataset. The various approaches included established methodologies such as fast.ai/ULMFiT (http://nlp.fast.ai/, accessed on 7 December 2023), SentencePiece (https://github.com/google/sentencepiece, accessed on 7 December 2023), BERT (https://github.com/google-research/bert, accessed on 7 December 2023), tpot (https://github.com/EpistasisLab/tpot, accessed on 7 December 2023), spaCy (https://spacy.io/api/textcategorizer, accessed on 7 December 2023), fasttext (https://fasttext.cc/, accessed on 7 December 2023), Flair (https://github.com/zalandoresearch/flair, accessed on 7 December 2023), neural networks

(especially involving a gated recurrent unit (GRU)), or conventional ML models such as support vector machines (SVM). Moreover, there were instances of innovative methods, such as Przetak (https://github.com/mciura/przetak, accessed on 7 December 2023). Evidently, the most successful strategy stemmed from the recent ULMFiT/fast.ai paradigm, adeptly applied by the n-waves team. Following closely was the initially proposed Przetak method by Plex, Inc., securing the second position, while the third spot was obtained by a fusion of ULMFiT/fast.ai, SentencePiece and the BranchingAttention model. The summary of the results of all participating teams in Task 1 are reported in Table 10.

Task 2

In the second task, among a total of eight submissions, there emerged five distinct entries. Among the teams that submitted multiple proposals, the independent group presented three and Sigmoidal presented two. The strategies that demonstrated notable effectiveness in the second task encompassed: SVM—proposed by the independent researcher Maciej Biesek, a combination of an ensemble of classifiers from spaCy combined with tpot and BERT—devised by the Sigmoidal team, and the utilization of fasttext—as employed by the AGH and UJ team. A comprehensive presentation of the results attained by all participating teams in Task 2 is reported in Table 11. Interestingly, although the participants frequently introduced novel methodologies, the majority of these methodologies predominantly hinged on lexical information encapsulated by terms (words, tokens, word embeddings, etc.). Interestingly, the incorporation of more advanced feature engineering, involving elements such as parts-of-speech, named entities, or semantic features, remained unexplored by all participants.

Table 10. Results from participants for Task 1. Highest results for each column in bold type font.

Submission Author(s)	Affiliation	Name of the Submitted System	Precision	Recall	F-Score	Accuracy
Piotr Czapla, Marcin Kardas	n-waves	n-waves ULMFiT	**66.67%**	**52.24%**	**58.58%**	**90.10%**
Marcin Ciura	Plex, Inc.	Przetak	66.35%	51.49%	57.98%	90.00%
Tomasz Pietruszka	Warsaw University of Technology	ULMFiT + SentencePiece + BranchingAttention	52.90%	54.48%	53.68%	87.40%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble spacy + tpot + BERT	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble + fastai	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team (Renard Korzeniowski, Przemysław Sadownik, Rafał Rolczyński, Tomasz Korbak, Marcin Możejko, Krystyna Gajczyk)	Sigmoidal	ensenble spacy + tpot	43.09%	58.21%	49.52%	84.10%
Rafal Pronko	CVTimeline	Rafal	41.08%	56.72%	47.65%	83.30%
Rafal Pronko	CVTimeline	Rafal	41.38%	53.73%	46.75%	83.60%
Maciej Biesek		model1-svm	60.49%	36.57%	45.58%	88.30%
Krzysztof Wróbel	AGH, UJ	fasttext	58.11%	32.09%	41.35%	87.80%
Katarzyna Krasnowska, Alina Wróblewska	IPI PAN	SCWAD-CB	51.90%	30.60%	38.50%	86.90%
Maciej Biesek		model2-gru	63.83%	22.39%	33.15%	87.90%
Maciej Biesek		model3-flair	81.82%	13.43%	23.08%	88.00%
Jakub Kuczkowiak	UWr	Task 6: Automatic cyberbullying detection	17.41%	32.09%	22.57%	70.50%

Table 11. Results from participants for Task 2. Highest results for each column in bold type font.

Submission Author(s)	Affiliation	Name of the Submitted System	Micro-Average F-Score	Macro-Average F-Score
Maciej Biesek		model1-svm	**87.60%**	**51.75%**
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble spacy + tpot + BERT	87.10%	46.45%
Krzysztof Wróbel	AGH, UJ	fasttext	86.80%	47.22%
Maciej Biesek		model3-flair	86.80%	45.05%
Katarzyna Krasnowska, Alina Wróblewska	IPI PAN	SCWAD-CB	83.70%	49.47%
Maciej Biesek		model2-gru	**78.80%**	**49.15%**
Jakub Kuczkowiak	UWr	Task 6: Automatic cyberbullying detection	70.40%	37.59%
Sigmoidal Team (Renard Korzeniowski, Przemyslaw Sadowski, Rafal Rolczynski, Tomasz Korbak, Marcin Mozejko, Krystyna Gajczyk)	Sigmoidal	ensamble + fastai	61.60%	39.64%

5. Dataset Re-Annotation with Cyberbullying Experts

As the original agreements of annotations were not satisfyingly high, suggesting possible mistakes, especially misses in annotations, we additionally performed a re-annotation of the whole dataset with the help of expert annotators experienced in annotating cyberbullying and other online harm-related data.

5.1. Details Concerning the Annotation Team

There were ten annotators (nine female and one male) and three supervisors (female) who worked on the data. The annotators were from 21 to 29 years of age. The annotators had background in linguistics and/or psychology. All of them had experience in working with annotation of harmful content; however, their experience varied. The most experienced person worked as a professional annotator of cyberbullying data for 2 years, while the least experienced one worked for 2 months. On average, annotators had the experience of approximately 10–12 months. The workload was also varied. Two people worked only on 1500 examples, while the highest number of examples per person was 8937.

5.2. Inter-Annotator Agreement among Experts

Inter-annotator agreements for expert annotators with confusion matrices required for reproducibility are represented in Table 12. The general agreements among experts were much higher than for laypeople (see previous sections) and reached 0.451 for standard Kappa, 0.517 for weighted Kappa, and 0.495 for the proposed Kappa with modified quadratic weights, all of which suggest *moderate agreement*. The results for the proposed Kappa also show how the measure corrects the overoptimistic assumption suggested by the traditional weighted Kappa.

Table 12. Inter-annotator agreements for expert annotators.

		Observed				**Expected**			
		Annotator 1				**Annotator 1**			
		0	1	2	sum	0	1	2	sum
Annotator	0	9052	697	184	9933	8380.69	1059.17	493.14	9933
	1	196	342	127	665	561.07	70.91	33.02	665
	2	65	138	237	440	371.24	46.92	21.84	440
	sum	9313	1177	548	11,038	9313	1177	548	11,038
								Kappa =	0.451
								weighted Kappa =	0.517
								weighted Kappa with modified quadratic weights =	0.495

5.3. General Overview of Re-Annotated Dataset

The overall number of tweets the final re-annotated dataset contained was 11,038 (three tweets were rejected from the original dataset due to near identical similarity), with 10,038 included in the training set and 1000 in the test set. The overview of the re-annotated dataset is represented in Table 13.

Similarly to the layperson annotations, in general, it can be said that expert annotators were able to specify with high confidence that a tweet is not harmful (even if it contained some vulgar words). However, differently from layperson annotators, experts were more consistent with both general tags (harmful/non-harmful) and specific tags (cyberbullying vs hate speech). This confirms, as mentioned in previous sections, that for tasks such as cyberbullying, expert annotation is required.

As for specific findings, for over eighty percent of the cases, all three annotators agreed, while there was no case where all three annotators disagreed. Interestingly, annotation lead by all experts resulted in assigning twice as many harmful labels than in annotation performed by trained laypeople. This confirms previous findings that experts are more sensitive in annotation [23]. It also suggests that when the annotation is entrusted to laypeople, half of the cyberbullying cases will be lost due to annotator mistakes or insufficient training.

Table 13. General statistics of the re-annotated dataset

	#	**% of All**	**% of Set**
# of Tweets where all Annotators agreed	8894	80.58%	
# of Tweets where at least two out of three Annotators agreed	11,038	100.00%	
# of final 0	9268	83.96%	
# of final 1	1356	12.28%	
# of final 2	414	3.75%	
# of all harmful	1770	16.04%	
Training set	10,038	90.94%	
# of final 0	8457	76.62%	84.25%
# of final 1	1217	11.03%	12.12%
# of final 2	364	3.30%	3.63%
# of all harmful	1581	14.32%	15.75%
Test set	1000	9.06%	
# of final 0	811	7.35%	81.10%
# of final 1	139	1.26%	13.90%
# of final 2	50	0.45%	5.00%
# of all harmful	189	1.71%	18.90%

5.4. Automatic Classification of Re-Annotated Dataset

For the experiments with the re-annotated dataset, we used both classic machine learning (ML) models and novel neural network-based models.

For the experiments with the re-annotated dataset using classic ML algorithms, we followed Eronen et al.'s [24] procedure and compared a number of classifiers with various preprocessing methods. In particular, we compared the classifiers including: multinomial naïve Bayes, Bernoulli-based naïve Bayes, linear SVM, SVM with stochastic gradient descent (SGD) optimization, logistic regression, logistic regression with conjugate gradient descent (CGD) optimization, random forest, AdaBoost, and XGBoost. From the various preprocessing and feature extraction methods, we applied only the ones for which Eronen et al.'s [24] extensive comparison concluded that they usually reached the highest scores. Specifically, we used a selected combinations of features containing tokens (TOK), lemmas (LEM), parts of speech (POS), and named entities (NER), as such features have been previously applied in cyberbullying and hate-speech detection [24,25], additionally supported by deleting stop words (stop) or characters other than alphanumerics (alpha). The list of all applied classifiers and all applied feature sets with their representative results for the

re-annotated dataset in macro F-score is represented in Table 14. We also used the same baselines as for the previous version of the dataset.

Table 14. Results (macro F-score) for classic ML models with various preprocessing methods. The five best scores are in bold type font.

Classifier	LEM	LEMNERalpha	LEMSNERS	LEMSNERSalpha	LEMSPOSS	LEMSPOSSstop	LEMSPOSSstopalpha	LEMalpha	LEMstop	LEMstopalpha	TOK	TOKPOS	TOKPOSSstopalpha	TOKSNERS	TOKSNERSalpha	TOKSNERSstop	TOKSNERSstopalpha	TOKSPOSS	TOKSPOSSalpha	TOKSPOSSstop
AdaBoost	0.148	0.561	0.590	0.157	0.579	0.579	0.156	0.158	0.149	0.557	0.576	0.543	0.549	0.154	0.571	0.568	0.152	0.157	0.161	0.153
BernoulliNB	0.155	0.521	0.485	0.155	0.485	0.485	0.155	0.155	0.156	0.498	0.464	0.458	0.474	0.161	0.489	0.464	0.159	0.161	0.158	0.158
CGD	0.152	0.552	0.540	0.152	0.548	0.516	0.169	0.152	0.157	0.508	0.518	0.513	0.524	0.153	0.525	0.494	0.153	0.155	0.156	0.155
Linear_SVM	0.171	0.645	**0.706**	0.215	**0.701**	0.679	0.244	0.215	0.181	0.641	**0.691**	**0.706**	0.631	0.163	0.663	**0.682**	0.209	0.169	0.195	0.178
LogisticRegression	0.152	0.552	0.540	0.152	0.548	0.516	0.169	0.152	0.157	0.508	0.518	0.513	0.524	0.153	0.525	0.494	0.153	0.155	0.156	0.155
MultinomialNB	0.157	0.479	0.469	0.155	0.447	0.447	0.156	0.155	0.156	0.499	0.458	0.453	0.458	0.158	0.464	0.469	0.155	0.160	0.159	0.160
RandomForest	0.152	0.584	0.559	0.146	0.550	0.563	0.162	0.148	0.142	0.620	0.565	0.553	0.578	0.148	0.584	0.602	0.170	0.151	0.154	0.152
SGD	0.143	0.599	0.612	0.146	0.596	0.589	0.147	0.143	0.148	0.611	0.596	0.580	0.579	0.147	0.580	0.589	0.152	0.147	0.150	0.148
XGBoost	0.174	0.604	0.598	0.160	0.573	0.590	0.173	0.160	0.165	0.567	0.613	0.575	0.584	0.164	0.602	0.608	0.158	0.161	0.160	0.165

For the experiments with neural networks, we specifically focused on transformer-based [26] models for the Polish language available on HuggingFace[10]. In particular, we compared the following models:

- Polbert-Polish BERT–dkleczek/bert-base-polish-uncased-v1
- distilHerBERT–BartekK/distilHerBERT-base-cased
- bert-base-pl-cased–Geotrend/bert-base-pl-cased
- distilbert-base-pl-cased–Geotrend/distilbert-base-pl-cased
- HerBERT–allegro/herbert-base-cased
- PolBERTa–marrrcin/PolBERTa-base-polish-cased-v1
- TrelBERT – deepsense-ai/trelbert

5.5. Results

5.5.1. Baselines

The results for the baselines on the re-annotated dataset for Task 1 are represented in Table 15. As the dataset was not balanced, it can be seen that a model that assigns 0 and 1 values randomly, as well as a model that assumes all instances are non-harmful, would still obtain approximately a 44% performance. The results of the baselines are comparable to their respective results from before re-annotation; yet, due to a larger number of harmful samples this time, the results for baseline B (always 1) are slightly higher (a 23% to 32% increase).

Table 15. Results for baselines on re-annotated dataset for Task 1.

Baselines	P_1	R_1	F1_1	P_0	R_0	F1_0	Macro_F1
always 0	0	0	0	0.81	1	0.9	0.4478189
always 1	0.19	1	0.32	0	0	0	0.1589571
random 0/1	0.18	0.47	0.25	0.8	0.49	0.61	0.4306366

5.5.2. Classic ML Models

The initial results for all applied classifiers and all applied feature sets with their representative results for the re-annotated dataset in macro F-score is represented in Table 14. From the above, we analyzed the classifiers that achieved first five best results in more detail. These are summarized in Table 16. The first five best scores were achieved all by the support vector machine classifier with linear function. This confirms a strong performance for SVMs in NLP-related tasks. Unfortunately, all those models struggled with the classification of class 1 (harmful), while all had no problems with classifying class 0 (non-harmful). This result might come from the fact that the dataset was imbalanced, with the harmful class accounting only for approximately 15% of the dataset. A general positive conclusion is that re-annotation caused a general increase in the results. This could come from both

gain in the quality of the dataset annotation and doubling in the number of samples (from approximately 8% to 15%). Together with the dataset, we are also releasing preprocessing scripts for the two kinds of preprocessing that obtained the highest scores when trained on linear SVM, namely, tokens combined with parts of speech (TOKPOS) and lemmas with added named entities (LEMSNERS).

Table 16. Detailed results for the five best models (linear SVM). Highest results for each column in bold type font.

Feature Set	P_1	R_1	F1_1	P_0	R_0	F1_0	Macro_F1	Acc
TOKPOS	**0.617**	0.429	0.506	0.875	**0.937**	**0.905**	**0.706**	**0.841**
LEMSNERS	0.558	**0.482**	**0.517**	**0.882**	0.910	0.896	**0.706**	0.829
LEMSPOSS	0.545	0.476	0.508	0.880	0.906	0.893	0.701	0.825
TOK	0.581	0.414	0.483	0.871	0.930	0.899	0.691	0.832
TOKSNERSstop	0.558	0.403	0.468	0.868	0.925	0.896	0.682	0.826

5.5.3. Transformer-Based Models

As for the applied deep learning (DL) models, specifically Transformer-based models, or bidirectional encoder representations from transformers (BERT) models, all models except one were better than all the classic ML models. The best model, Polbert, achieved a 0.794 macro average F1 score, which is a strong baseline for future attempts. Interestingly, TrelBERT, which was additionally trained on millions of tweets and achieved high scores on the Polish KLEJ benchmark, although still scoring very high, did not achieve the highest scores here. It is difficult to estimate the reason for the lower score, since models pretrained on Twitter have been shown to work better on datasets based on Twitter [27]; however, a number of unpublished internal evaluations suggest that since Tweets are usually short and contain simplistic language due to the character limitations[28], even after the extension of character limitation from 140 to 280 introduced in 2017 by Twitter due to users "cramming" too many abbreviations and slang into tweets[11], the knowledge embedded in Twitter-based models is usually more shallow and limited, which could influence the performance for tasks requiring more context-based knowledge and the ability to process a wider/deeper context and implicit language, such as the task of automatic cyberbullying detection.

Unfortunately, all transformer-based models showed a similar bias for the more numerous class (non-harmful) and achieved over 90% for the F1 score for that class. This again supports the need to collect more balanced datasets in the future. Thus, in reality, the results for specifically detecting cyberbullying (class 1) were unsatisfying, reaching barely above 50% precision for the best performing model. This leaves sufficient space for improvement in the future.

The results for all tested transformer-based models are represented in Table 17.

Table 17. The results for all tested transformer-based models on re-annotated dataset. Highest results for each column in bold type font.

Model	P_1	R_1	F1_1	F1_0	F1_Macro	Acc
distilbert-base-pl-cased–Geotrend/distilbert-base-pl-cased	0.217	0.719	0.333	0.906	0.620	0.836
distilHerBERT–BartekK/distilHerBERT-base-cased	0.402	0.704	0.512	0.915	0.713	0.855
PolBERTa–marrrcin/PolBERTa-base-polish-cased-v1	0.413	0.703	0.520	0.915	0.718	0.856
HerBERT–allegro/herbert-base-cased	0.407	0.748	0.527	0.919	0.723	0.862
bert-base-pl-cased–Geotrend/bert-base-pl-cased	0.471	0.659	0.549	0.913	0.731	0.854
TrelBERT–deepsense-ai/trelbert	0.471	0.864	0.609	0.933	0.771	0.886
Polbert-Polish BERT–dkleczek/bert-base-polish-uncased-v1	**0.561**	**0.785**	**0.654**	**0.933**	**0.794**	**0.888**

6. Usage Notes

To contribute to solving the recently growing problem of cyberbullying and hate speech appearing on the Internet, we presented the first dataset to study cyberbullying in the Polish language for the development of automatic cyberbullying detection methods.

The dataset allows the users to try their classification methods to determine whether an Internet entry (e.g., a tweet) is classifiable as harmful (cyberbullying/hate-speech) or non-harmful. The entries contain tweets collected from openly available Twitter discussions and were provided as such, with minimal preprocessing.

The dataset can be used in several ways. First, it can be used as a linguistic resource to qualitatively study how cyberbullying and related phenomena are expressed online in the Polish language, similarly to what we have demonstrated in the above sections. Second, it can be used to create new cyberbullying detection tools for the Polish language. The first attempts are described in the above sections. Third, the best-performing model for automatic cyberbullying detection in the Polish language, released together with the dataset, can be freely applied in cyberbullying detection and prevention architectures for social networking services.

The tools required to run the accompanied scripts include the Perl programming language, the `Text::Guess::Language`[12] Perl package for language detection, and the Python 3.9.7. programming language, as well as the following Python packages: `pandas`[13], `spaCy`[14], `NumPy`[15], `transformers`[16].

7. Contributions

In recent years, there has been an increase in releases of datasets containing entries annotated for cyberbullying, toxicity, offensiveness, etc., such as the ones listed in the https://hatespeechdata.com repository (accessed on 13 Decemebr 2023). However, most of those datasets are either (1) annotated by laypeople, (2) using non-standardized taxonomies, or (3) are published without proper quality control. In the majority, more than one of those problems, if not all, apply to most of the available datasets. On the contrary, datasets collected and annotated by experts with proper quality control are scarce. Moreover, if they exist, they are available in high-resource languages such as English. The presented dataset was specifically collected and annotated by experts, with proper quality control (thus, two versions of annotations), and is available for a language that is not as popular as English. Therefore, the dataset presents a valuable contribution and can be reused and analyzed both in linguistic research as well as for training ML models. We were further encouraged to publish the dataset by the increasing tendency for inflammatory entries on the Polish Internet. Thus, with this dataset, we wish to provide two specific contributions. First, various researchers will be able to use this expert-annotated dataset to study cyberbullying in the Polish language. Second, with this dataset, we wish to encourage other researchers to publish their datasets in this topic, especially in low-resource and underrepresented languages.

8. Limitations

We acknowledge the following limitations of the study and the provided dataset.

1. **Small overall size of the dataset**: One of the primary limitations of this study is the relatively small size of the dataset used for cyberbullying and hate speech detection. This limited dataset size may affect the generalizability of the findings and could lead to potential overfitting of machine learning models. A larger and more diverse dataset would enhance the robustness of the analysis and improve the reliability of the results.
2. **Imbalance of classes**: Another significant limitation is the class imbalance present in the dataset. The majority of instances in the dataset belong to the non-harmful class, while instances of cyberbullying and hate speech are underrepresented. A class imbalance can bias the model towards the majority class, potentially leading to

lower sensitivity in detecting harmful content. Techniques such as oversampling or undersampling may be employed to address this issue, but it remains a limitation.

3. **Labeling and annotation subjectivity**: The process of labeling and annotating instances of cyberbullying and hate speech involves a degree of subjectivity. Different annotators may interpret content differently, leading to potential inconsistencies in the dataset. While efforts were made to ensure inter-annotator agreement, for example, by employing cyberbullying experts in annotations, this inherent subjectivity remains a potential limitation in the quality of the dataset.
4. **Temporal dynamics**: Cyberbullying and hate speech evolve over time, influenced by changes in language, cultural norms, and online platforms. This study may not capture the latest trends and variations in cyberbullying and hate speech due to the static nature of the dataset. Longitudinal data collection and analysis would provide a more comprehensive understanding of these phenomena.
5. **Contextual understanding**: Detecting cyberbullying and hate speech often requires a deep understanding of context, sarcasm, irony, and cultural references. This study may not fully account for the nuanced contextual cues that humans use to identify harmful content. Developing more context-aware models is an ongoing challenge in this field.
6. **Ethical considerations**: The study primarily focuses on detection and classification but does not address the broader ethical implications of content moderation and censorship. Ethical concerns surrounding freedom of speech and the potential for false positives or unintended consequences are important considerations not covered comprehensively in this research.
7. **Generalization to other languages and platforms**: This study may be limited in its generalizability to languages and online platforms beyond those covered in the dataset. Variations in language use and online behavior across different linguistic and cultural contexts are not fully explored in this research.
8. **Evolution of countermeasures**: The study assumes a static environment regarding countermeasures against cyberbullying and hate speech. However, efforts to combat harmful content are continually evolving, and the study does not account for potential changes in content moderation policies, algorithms, or user behaviors.

In conclusion, while this study provides a valuable dataset for the studies into cyberbullying and hate speech detection, it is essential to acknowledge the above limitations. The future research should aim to address these limitations to build a more comprehensive understanding of the challenges and nuances associated with identifying and mitigating harmful online content.

Author Contributions: M.P. preprocessed the data, conceived and conducted the experiments, analyzed the results, and wrote the first draft of the manuscript; A.P. collected the data and advised in the creation of the guidelines for data annotation; P.D. helped create the guidelines for data annotation, managed the first version of annotation, and analyzed the results; P.S. managed the data re-annotation process and supervised annotation quality; K.S. adjusted the data for re-annotation and oversaw the annotation pipeline; M.F. helped writing parts of the manuscript; G.L. advised on the taxonomies, guidelines, and annotation pipeline; M.W. oversaw the re-annotation process. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All files are available at Zenodo [29]. All the source code necessary to manipulate the dataset is released together with the dataset.

Conflicts of Interest: The authors declare no competing interests.

Abbreviations

AI	Artificial intelligence
API	Application programming interface
BERT	Bidirectional encoder representations from transformers
CB	Cyberbullying
CGD	Conjugate gradient descent
DL	Deep learning
GRU	Gated recurrent unit
HS	Hate speech
JSON	JavaScript object notation
ML	Machine learning
SGD	Stochastic gradient descent
SNS	Social networking services
SVM	Support vector machines
TF-IDF	Term frequency with inverse document frequency

Notes

1 https://huggingface.co/datasets/hate_speech_pl (accessed on 7 December 2023).
2 https://huggingface.co/datasets/poleval2019_cyberbullying (accessed on 7 December 2023).
3 https://huggingface.co/datasets/Paul/hatecheck-polish (accessed on 7 December 2023).
4 https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html (accessed on 7 December 2023).
5 https://github.com/bear/python-twitter/ (accessed on 7 December 2023).
6 https://github.com/mongodb/mongo-python-driver (accessed on 7 December 2023).
7 as per: https://www.sotrender.com/blog/pl/2018/01/twitter-w-polsce-2017-infografika/ (accessed on 7 December 2023).
8 https://metacpan.org/pod/Text::Guess::Language (accessed on 7 December 2023).
9 https://en.wikipedia.org/wiki/Tf-idf (accessed on 7 December 2023).
10 https://huggingface.co (accessed on 7 December 2023).
11 https://blog.twitter.com/engineering/en_us/topics/insights/2017/Our-Discovery-of-Cramming (accessed on 7 December 2023).
12 https://metacpan.org/pod/Text::Guess::Language (accessed on 7 December 2023).
13 https://pandas.pydata.org (accessed on 7 December 2023).
14 https://spacy.io (accessed on 7 December 2023).
15 https://numpy.org (accessed on 7 December 2023).
16 https://pypi.org/project/transformers/ (accessed on 7 December 2023).

References

1. Ptaszynski, M.E.; Masui, F. *Automatic Cyberbullying Detection: Emerging Research and Opportunities*; IGI Global Publishing: Hershey, PA, USA, 2018. ISBN 9781522552499.
2. Ptaszynski, M.; Dybala, P.; Matsuba, T.; Masui, F.; Rzepka, R.; Araki, K. Machine Learning and Affect Analysis Against Cyber-Bullying. In Proceedings of the Thirty Sixth Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB-10), Leicester, UK, 29 March–1 April 2010; pp. 7–16.
3. Ptaszynski, M.; Dybala, P.; Matsuba, T.; Masui, F.; Rzepka, R.; Araki, K.; Momouchi, Y. In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis. *Int. J. Comput. Linguist. Res.* **2010**, 1, 135–154.
4. Ptaszynski, M.; Kalevi, J.; Eronen, K.; Masui, F. Learning Deep on Cyberbullying is Always Better Than Brute Force. In Proceedings of the IJCAI 2017 3rd Workshop on Linguistic and Cognitive Approaches to Dialogue Agents (LaCATODA 2017), Melbourne, Australia, 19–25 August 2017.
5. Ptaszynski, M.; Masui, F.; Kimura, Y.; Rzepka, R.; Araki, K. Brute Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection. *J. Assoc. Inf. Syst.* **2019**, *20*, 1075–1127. [CrossRef]
6. Tworzecki, H. Poland: A Case of Top-Down Polarization. *Ann. Am. Acad. Political Soc. Sci.* **2019**, *681*, 97–119. [CrossRef] https://doi.org/10.1177/0002716218809322.
7. Bilewicz, M.; Soral, W. Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychol.* **2020**, *41*, 3–33. [CrossRef]
8. Domalewska, D. Disinformation and Polarization in the Online Debate During the 2020 Presidential Election in Poland. *Saf. Def.* **2021**, *7*, 14–24.
9. Moulin-Stozek, M. Trends of Radicalization. D3.2 Country Report June 2021. Conducted under the Horizon 2020 project 'De-Radicalisation in Europe and Beyond: Detect, Resolve, Re-integrate'(959198). 2021. Available online: https://dradproject.com/?publications=trends-of-radicalisation-in-poland (accessed on 7 December 2023).

10. Troszyński, M.; Wawer, A. Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych. *Przegląd Socjologii Jakościowej* **2017**, *13*, 62–80. [CrossRef]
11. Ptaszynski, M.; Pieciukiewicz, A.; Dybała, P. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. In *Proceedings of the PolEval 2019 Workshop*; Institute of Computer Sciences: Warsaw, Poland, 2019.
12. Kobylinski, Ł.; Ogrodniczuk, M.; Kocon, J.; Marcinczuk, M.; Smywinski-Pohl, A.; Wołk, K.; Koržinek, D.; Ptaszynski, M.; Pieciukiewicz, A.; Dybała, P. PolEval 2019—The next chapter in evaluating Natural Language Processing tools for Polish. In Proceedings of 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland, 17–19 May 2019.
13. Röttger, P.; Seelawi, H.; Nozza, D.; Talat, Z.; Vidgen, B. MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Detection Models. *arXiv* **2022**, arXiv:2206.09917.
14. Okulska, I.; Głąbińska, K.; Kołos, A.; Karlińska, A.; Wiśnios, E.; Nowakowski, A.; Ellerik, P.; Prałat, A. Ban-pl: A novel polish dataset of banned harmful and offensive content from wykop.pl web service. *arXiv* **2023**, arXiv:2308.10592.
15. Tversky, A.; Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **1974**, *185*, 1124–1131. [CrossRef]
16. Dooley, J.J.; Pyżalski, J.; Cross, D. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Z. Psychol./J. Psychol.* **2009**, *217*, 182–188. [CrossRef]
17. Cano, E.; He, Y.; Liu, K.; Zhao, J. A Weakly Supervised Bayesian Model for Violence Detection in Social Media. In Proceedings of the In Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan, 14–18 October 2013.
18. Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; Picard, R. Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying. *ACM Trans. Intell. Interact. Syst.* **2012**, *2*, 1–30. [CrossRef]
19. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
20. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213. [CrossRef]
21. Fleiss, J.L.; Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. [CrossRef]
22. Ptaszynski, M.; Zasko-Zielinska, M.; Marcinczuk, M.; Leliwa, G.; Fortuna, M.; Soliwoda, K.; Dziublewska, I.; Hubert, O.; Skrzek, P.; Piesiewicz, J.; et al. Looking for Razors and Needles in a Haystack: Multifaceted Analysis of Suicidal Declarations on Social Media—A Pragmalinguistic Approach. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11759. [CrossRef] [PubMed]
23. Ptaszyński, M.; Leliwa, G.; Piech, M.; Smywiński-Pohl, A. Cyberbullying Detection–Technical Report 2/2018, Department of Computer Science AGH, University of Science and Technology. *arXiv* **2018**, arXiv:1808.00926.
24. Eronen, J.; Ptaszynski, M.; Masui, F.; Smywiński-Pohl, A.; Leliwa, G.; Wroczynski, M. Improving classifier training efficiency for automatic cyberbullying detection with Feature Density. *Inf. Process. Manag.* **2021**, *58*, 102616. [CrossRef]
25. Mastromattei, M.; Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Comput. Sci.* **2022**, *8*, e859. [CrossRef]
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
27. Qudar, M.M.A.; Mago, V. Tweetbert: A pretrained language representation model for twitter text analysis. *arXiv* **2020**, arXiv:2010.11091.
28. Boot, A.B.; Tjong Kim Sang, E.; Dijkstra, K.; Zwaan, R.A. How character limit affects language usage in tweets. *Palgrave Commun.* **2019**, *5*, 1–13. [CrossRef]
29. Ptaszynski, M.; Pieciukiewicz, A.; Dybala, P.; Skrzek, P.; Soliwoda, K.; Fortuna, M.; Leliwa, G.; Wroczynski, M. Expert-Annotated Dataset to Study Cyberbullying in Polish Language. *Zenodo.* Version v1. [CrossRef]