Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter

Michal Ptaszynski (Kitami Institute of Technology),
Agata Pieciukiewicz (Polish-Japanese Academy of Information Technology),
Paweł Dybała (Jagiellonian University in Kraków)

Abstract

In this paper we describe the first dataset for the Polish language containing annotations of harmful and toxic language. The dataset was created to study harmful Internet phenomena such as cyberbullying and hate speech, which recently dramatically gain on numbers in Polish Internet as well as worldwide. The dataset was automatically collected from Polish Twitter accounts and annotated by both layperson volunteers under the supervision of a cyberbullying and hate-speech expert. Together with the dataset we propose the first open shared task for Polish to utilize the dataset in classification of such harmful phenomena. In particular, we propose two subtasks: 1) binary classification of harmful and non-harmful tweets, and 2) multiclass classification between two types of harmful information (cyberbullying and hate-speech), and other. The first installment of the shared task became a success by reaching fourteen overall submissions, hence proving a high demand for research applying such data.

Keywords

cyberbullying, automatic cyberbullying detection, hate-speech, natural language processing, machine learning

1. Introduction

Although the problem of humiliating and slandering people with the use of Internet communication measures has existed almost as long as the communication via the Internet between people itself, the appearance of new handheld mobile devices, such as smartphones and tablet computers, which allow using the Internet not only at home, work or school but also in commute, has further intensified the problem. Especially recent decade, during which

Social Networking Services (SNS) such as Facebook and Twitter, rapidly grew in popularity, has brought to light the problem of unethical behaviors in Internet environments, which has been greatly impairing public mental health in adults and, for the most, in younger users and children. It is the problem of *cyberbullying* (CB), defined as exploitation of open online means of communication, such as Internet forum boards, or SNS to convey harmful and disturbing information about private individuals, often children and students.

To deal with the problem, researchers around the world have begun to study the problem with a goal of automatic detection of Internet entries containing harmful information and reporting them to SNS service providers for further analysis and deletion. After ten years of research (Ptaszynski et al. 2010b,a, Nitta et al. 2013a,b, Hatakeyama et al. 2015, Ptaszynski et al. 2015, Lempa et al. 2015, Hatakeyama et al. 2016a, Ptaszynski et al. 2016a, Hatakeyama et al. 2016b, Ptaszynski et al. 2016b, 2017, 2018, Ptaszynski and Masui 2018, Ptaszynski et al. 2019), a sufficient knowledge base on this problem has been collected for languages of well-developed countries, such as the US, or Japan. Unfortunately, still close to nothing in this matter has been done for the Polish language. With the presented here dataset and the initial experiments performed with the dataset, we aim at filling this gap.

The dataset, as well as open shared task supplementing the dataset, allows the users to try their classification methods to determine whether an Internet entry is classifiable as part of cyberbullying narration or not. The entries contain tweets collected from openly available Twitter discussions. Since much of the problem of automatic cyberbullying detection often relies on feature selection and feature engineering (Ptaszynski et al. 2017, 2019), the tweets are be provided as such, with minimal preprocessing. The preprocessing, if used, is applied mostly for cases when information about a private person is revealed to the public.

The goal of the main task is to classify the tweets into cyberbullying/harmful and non-cyberbullying/non-harmful with the highest possible Precision, Recall, balanced F-score and Accuracy. In an additional subtask, the goal is to differentiate between various types of harmful information, in particular cyberbullying and hate-speech, and non-harmful[1].

The rest of the paper is organized in the following way. In Section 2 we describe how the data for the dataset was collected. In Section 3 we explain the whole annotation process, including our working definition of cyberbullying and guidelines for annotation used in training the annotators. In Section 4 we perform an in-depth analysis of the created dataset, which includes both general statistical analysis as well as deeper example-based specific analysis. In Section 5 we describe the task we propose together with the dataset, in particular two subtasks for classification of 1) harmful information in general and 2) two specific types of harmful information. We also propose the default means for evaluation and introduce the participants that took part in the first installment of the shared task. In Section 4 we present the results of the participants in comparison to a number of baselines. Finally, in Sections 7 and 8 we conclude the paper and set up plans and directions for the near future.

[1]The dataset, together with the two subtasks proposed for it, is available under the following URL: `https://github.com/ptaszynski/cyberbullying-Polish`

2. Data Collection and Preprocessing

2.1. Collection

In order to collect the data, we used Standard Twitter API[2]. It has a number of limitations, which we had to work around. For example, the number of requests per 15-minute window and the number of tweets that could be downloaded in one request is limited by Twitter API. We respected those limits, and after exhausting the limit of requests the download script simply waited for another download window. Twitter API was used via the python-twitter library (`https://github.com/bear/python-twitter/`). Another obstacle was the time limit for searching tweets. In Standard (non-paid) Twitter API the user is allowed to search for tweets from past 7 days. That is why we were not able to collect all answers to tweets made from our initial starting accounts. Our script saved data received from Twitter in MongoDB using the pymongo library (`https://github.com/mongodb/mongo-python-driver`). Twitter provides tweet data in JSON format, so the use of a document database was convenient for further handling of data.

The script, written in Python, has been used to download tweets from nineteen official Polish Twitter accounts. Those accounts were chosen as the most popular Polish Twitter accounts in the year 2017[3]. By popular we understand those with the largest number of observers, those with a rapidly growing number of observers, those who collected the most user activity, those most often mentioned and those who themselves tweeted most often. In particular, we initially looked at the following accounts: `@tvn24`, `@MTVPolska`, `@lewy_official`, `@sikorskiradek`, `@Pontifex_pl`, `@donaldtusk`, `@BoniekZibi`, `@NewsweekPolska`, `@PR24_pl`, `@tvp_info`, `@rzeczpospolita`, `@AndrzejDuda`, `@lis_tomasz`, `@K_Stanowski`, `@R_A_Ziemkiewicz`, `@pisorgpl`, `@Platforma_org`, `@RadioMaryja`, `@RyszardPetru`.

In addition to tweets from those accounts, we have collected answers to any tweets from the accounts mentioned above (from the past 7 days). In total, we have received over 101 thousand tweets from 22,687 accounts (as identified by screen_name property in the Twitter API). Using bash random functions ten accounts were randomly selected to become the starting point for further work.

Next, using the same script as before, we downloaded tweets from these 10 accounts and all answers to their tweets that we were able to find using the Twitter Search API (again, limited to the past 7 days). Using this procedure we have selected 23,223 tweets from Polish accounts for further analysis. Data downloading was finished on 20.11.2018. (Last downloaded tweet was created at 18:12:32). These 23,223 tweets became the base for the dataset presented in this paper.

[2] `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html`

[3] According to `https://www.sotrender.com/blog/pl/2018/01/twitter-w-polsce-2017-infografika/`

2.2. Preprocessing and Filtering

Since in this initial dataset, we did not follow the conversation threads (as the official Twitter API does not provide such information), we considered each tweet separately.

At first, we randomized the order of tweets in the dataset to get rid of any consecutive tweets from the same account. This would help decrease the anchoring bias (Tversky and Kahneman 1974) in annotations since when a human annotator reads tweets from the same account they could become prone to assigning the same score to many messages.

Next, we got rid of all tweets containing URLs. This was done due to the fact that URLs often take space and limit the contents of the tweets, which in practice often resulted in tweets being cut in the middle of the sentence or with a large number of *ad hoc* abbreviations. Next, we got rid of all tweets which were exactly the same in contents, which eliminated most of the duplications. Tweets consisting only of at-marks (@) or hashtags (#) were also removed, as they do not convey any intrinsic linguistic value as a whole, but rather are used as unrelated keywords. Finally, we removed tweets with less than five words and those written in languages other than Polish. This left us with 11,041 tweets. From this group, we randomly extracted 1,000 tweets to be used as test data and the rest (10,041) was used as training data. The exact step-by-step preprocessing procedure and analysis of how many tweets were discarded at each time is presented below.

1. Deleted tweets with URLs and retaining only the text of tweets (no meta-data, timestamps, etc.) (retained: 15,357/23,223, or 66.13% of all).
2. Deleted exact duplicates (retained:15,255, only 102 deleted, or 0.44% of all).
3. Deleted tweets containing only @atmarks and #hashtags (retained: 15,223, only 32 deleted, or 0.14% of all).
4. Deleted tweets that, except @atmarks or #hashtags consist of only a single or a few words or emoji, etc.:
   (a) Deleted tweets with only one word (retained: 14,492 tweets, 731 deleted, or 3.1% of all).
   (b) Deleted tweets with only two words (retained: 13,238 tweets, 1254 deleted, or 5.4% of all).
   (c) Deleted tweets with only three words (retained: 12,226 tweets, 1012 deleted, or 4.4% of all).
   (d) Deleted tweets with only four words (retained: 11,135 tweets, 1091 deleted, or 4.7% of all).

After the above operations, we were left with 11,135 tweets containing five or more words, not counting @atmarks or #hashtags.

The reasoning behind deleting short tweets was the following:

1. For a human annotator a tweet that is too short will contain an insufficient amount of context, and thus will be difficult to appraise, thus creating many ambiguous annotations.

2. It is also better for machine learning models to have more contents (features) to train on, which also suggests longer sentences will help training more accurate machine learning models. Although one can imagine short tweets also containing aggression, we can assume that if a system is trained on a larger data it will also be able to cover shorter tweets.

In the remaining 11,135 tweets we also noticed a few samples written in a language other than Polish, mostly in English. To solve this problem we used a Text::Guess::Language Perl module[4], which detects the language of a sentence based on top 1000 words from that language. Initial manual analysis of a small sample of tweets revealed that the module sometimes erroneously guessed tweets written in Polish as written in Slovak or Hungarian, due to strangely sounding account names (@atmarks) and #hashtags sometimes used in tweets, but was never wrong when detecting tweets written in English. Therefore as a rule of thumb, we discarded only all English tweets, which in practice left us with only tweets written in Polish. After this final preprocessing operation we were left with 11,041 tweets, from which we used as training data 10,041 tweets and as 1000 tweets as test data.

Together with the dataset we also released a short Perl script used to discard tweets in English from Polish data (onlypolish.pl), as well as tweets that contain only @atmarks or #hashtags (extractnohashatmarks.pl).

3. Annotation Schema

3.1. Cyberbullying – a Working Definition

To develop the annotation schema for annotating the downloaded tweets we firstly prepared our working definition of cyberbullying. Although there is a number of general definitions of the problem, most definitions (Ptaszynski and Masui 2018) agree that

> *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.)*

Also, social science studies (Dooley et al. 2009) agree that there are both similarities between cyberbullying and traditional face-to-face bullying, as well as differences, which make cyberbullying a problem more difficult to mitigate. Similarities, which make the problem classifiable as a kind of bullying, include: peer group, such as classmates in face-to-face bullying and friends from groups on SNS, which in reality also often overlap; repetitiveness of bullying

[4] `https://metacpan.org/pod/Text::Guess::Language`

acts, which especially on the Internet, could occur more often than in face-to-face bullying; imbalance of power, where one person or a small group becomes bullied by an overwhelming number of bullies and their supporters.

It would be ideal to be able to analyze the data in a wider context, such as threads of conversations on Twitter. Unfortunately, Twitter API does not allow for grouping of conversations, thus in this dataset, we consider each tweet separately. This approach is also similar to all of the previous studies, where each Internet entry was considered as a separate example (Ptaszynski and Masui 2018). In future, however, it is desirable to find a way to automatically group the tweets into conversations to be able to annotate roles of participants in cyberbullying, such as a victim, bully, or bystanders (supporters, defenders).

3.2. Annotation Guidelines

To help annotators perform their task efficiently and to limit the subjective bias of each annotator, we prepared the guidelines for annotations of tweets for harmful information. The guidelines include the following:

English version		**Polish version**
phishing, disclosure or threat of disclosure of private information (phone number, e-mail, address, account name, school name/number, class at school, private identification number (PESEL), credit card number, etc.)	—	wyłudzanie, ujawnienie lub groźba ujawnienia prywatnych informacji (numer tel., e-mail, adres, nazwa konta, nazwa/numer szkoły, klasy, PESEL, karta kredytowa, itd.)
personal attack (“Kill yourself, bitch!”, etc.)	—	atak personalny (“Powieś się, gnoju!”, etc.)
threats (“I will find you and I will kill you”, etc.)	—	groźby (“znajdę cię i zajebię”, etc.)
blackmail (“I will tell everyone where you live if you do not pay me”, etc.)	—	szantaże (“powiem wszystkim gdzie mieszkasz, jeśli mi nie zapłacisz”, etc.)
mocking/ridiculing (“Look how fat this guy is”, “you pimple-face”, etc.)	—	szyderstwa/wyśmiewanie (“Patrzcie na tego grubasa”, “ty pryszczata mordo”, etc.)
gossip/insinuations (“Hey, apparently he’s a zoophiliac!”, etc.)	—	plotki/insynuacje (“Ej, podobno to zoofil!”, etc.)
the accumulation of profanity (single profane and vulgar words appear in conversations fairly often, but a longer “bundle” can be considered as harmful)	—	nagromadzenie wulgaryzmów (pojedyncze występują dość często, ale ich nagromadzenie może być potraktowane jako niepożądane)
various combinations of all of the above	—	różne kombinacje wszystkich powyższych

The scope of the collection of tweets

Cyberbullying is usually addressed at private individuals, thus for the dataset, we used only tweets from private Twitter accounts. We did not include tweets from public accounts (politicians, celebrities) since these are usually from the definition exposed to criticism and personal attacks due to their profession, and often provoke themselves such criticism to raise their popularity. There is no doubt that a public person might also feel privately offended, but even in such case, public persons have the means to deal with such a problem (e.g., employees who massively report abuses in the Twitter system, exert pressure in a number of different ways, even sue an aggressive user).

Harmful, but not cyberbullying

Despite limiting the scope of search to private accounts, there is always a possibility that a harmful tweet addressed at a public person will appear in such collection, Therefore, we decided to also annotate all tweets that do not represent cyberbullying, but are harmful in any other way, e.g., represent hate speech, racism, sexism, but are not addressed at a private person, or a specific small group (e.g. not "you" or "a few people from the class"), but rather a public person, or a specific community in general (e.g., "gays and lesbians", or "Paki" (Pakistanians)/"ciapaty" in Polish).

3.3. Annotation Process

Annotators were provided with only the contents of the tweets and performed annotation one tweet at a time. Each tweet was annotated by at least two, at most three layperson annotators and one expert annotator. Layperson annotators were trained for cyberbullying and hate-speech detection with the guidelines described in this section. Layperson annotators were a group of seven people, all female, in their early twenties. The one expert annotator was a male in his late thirties with a 10-year experience in research on cyberbullying and cyberbullying detection.

After layperson annotators performed their annotations, the expert annotator looked through all annotations and either approved or corrected them. The annotations consisted of the following type of information:

A) harmfulness score:

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

B) specific tag if possible to specify (see next page)

C) specific phrases if possible to specify in the text.

Abbreviation	Full description	Explanation
pry	prywatne	disclosure or threat of disclosure of private information, phishing
atk	atak	personal attack
gro	groźba	threat
sza	szantaż	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

3.4. Examples of Tweets with Annotations

In Table 1 we show a number of examples. Since the dataset contained tweets from various private sources, the annotators were trained to annotate the tweets regardless of their political sentiments. Thus one can see tweets with assigned harmfulness score for both anti-alt-right (Example 2, 4, 6) and anti-left (Example 5), as well as of unknown addressee (Example 1). Some tweets contained typos (Example 5, "endekdu" instead of "endeku" – from "National Democracy supporter"; Example 10 "czulem" instead of "czułem", "głow" instead of "głowa"). Some tweets, which, although contained vulgar vocabulary, were not considered harmful as were not directed at a particular person or a group (Example 12, "dupa"/"ass"). On the other hand, some tweets, although also not being directed at anyone in particular, were encouraging the use of illegal substances, thus were considered as harmful (Example 3).

4. Dataset Analysis and Discussion

4.1. General Statistical Analysis

The overall number of tweets the final dataset contained was 11,041 with 10,041 included in the training set and 1000 in the test set. The layperson annotators agreed upon most of the annotations, with overall 91.38% of agreements, with a very small number of tweets which either of the annotators was unable to tag (84, or 0.76%). This was is a high percentage of agreements, however, this high percentage was mostly due to the fact that most of the annotators agreed upon non-harmful tweets, which comprised most of the dataset (over 89.76%). Among the final number of harmful tweets, the annotators fully agreed on the cyberbullying class (1) for only 106 (0.96%) and on the hate-speech class (2) for only 73 tweets (0.66%). Moreover, even some of the tweets with full agreement ended up being corrected by the expert annotator to other class, thus making the agreements even weaker. In general, it can be said, that layperson annotators can specify with fair amount of confidence that a tweet is not harmful (even if it contains some vulgar words), and to some extent can spot when there is something wrong or undesirable with the tweet, but in majority struggle to specify, what exactly is the reason of a tweet being undesirable. This provides proof that

for specific problems such as cyberbullying, an expert annotation is required. This has been highlighted by Ptaszynski and Masui (2018) for their ten years of research, although there has been a large number of research applying in annotation laypeople, such as undergraduate students or Mechanical Turk workers (Cano et al. 2013, Dinakar et al. 2012). When it comes to the comparison between the training and the test set, the latter contained a slightly higher ratio of harmful tweets (8.48% for training set vs. 13.40% for test set), which could cause some problems for the participants.

Table 1: Examples of tweets with annotated information

No.	Tweet contents	Score	Tag	Phrases
1	@anonymized_account Jebać, jebać, nic się nie bać.	2		jebać
2	@anonymized_account @anonymized_account Ty się oszuście i kłamco od nas odpierdol. My sobie damy radę bez twoich kłamstw	2	atk, wul, atk	oszuście, kłamco, odpierdol się
3	Picie to przyszłość, piękna i nieznana, szukajcie zapomnienia, próbujcie różnych trunków, łączcie środki psychoaktywne, Powodzenia	2		
4	@anonymized_account Ale ty jesteś żenujący. Dno metr mułu.	2		
5	@anonymized_account Jaki hejt. Hejt to jest na Tuska, Komorowskiego, przecież to nie Polacy, śmierć wrogom ojczyzny. Obudź się	2	gro	śmierć wrogom ojczyzny
6	@anonymized_account Wio endekdu, ścierwa, zdrajcy, szubienica	1	szy	ścierwa, endek, szubienica, zdrajcy
7	@anonymized_account Jeszcze was zjemy i wysramy	1	atk, gro	zjemy, wysramy, jeszcze
8	@anonymized_account @anonymized_account A ty wieś kretynie CONTI jest Acta2 i czego dotyczy? Najpierw przeczytaj a potem się wypowiadaj.	1	atk	kretynie
9	@anonymized_account @anonymized_account Ty pajacu, zmień sobie herb na pusty łeb.	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	0		
11	Kiedy Christina wychodzi za mąż	0		
12	@anonymized_account kot też się załapał na fotke, a raczej jego dupa :)	0		

Table 2: General statistics of the dataset

	#	% of all	% of set
Overall # of tweets	11041	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	10089	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 retweets (RT) which slipped through	709		
# of final 0	10056	91.08%	
# of final 1	278	2.52%	
# of final 2	707	6.40%	
# of all harmful	985	8.92%	
Training set	10041	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%

Apart from the above statistics, there was also a fairly large number of retweets that slipped through both the data preparation process as well as a later annotation (709 or 6.42%). All of those tweets were not official retweets, but tweet quotations starting with a short comment "RT". This situation will need to be taken into consideration when creating the second, improved version of the dataset in the future.

4.2. Discussion on Specific Tweet Examples

The whole annotation process provided a number of valuable insights reported by the annotators. For example, many annotators noticed that the meaning of most tweets depended on the context, and when the context was unclear, it was difficult to evaluate them in the given categories (especially for the harmful category). The entire conversation between Twitter users would facilitate better assessment, and show the context in which the given tweet was published. This problem could be solved by clustering tweets into conversation threads. We will propose a method for automatic clustering of tweets into coherent threads. This could be done by incorporating, a specific meta-information about at which tweet the message is addressed at, provided by the API (`in_reply_to_status_id`), or taking additional

advantage of user quotations (`@user`), which appear at the beginning of tweets usually as responses, together with time between tweets, which could additionally suggest the tweet being a response with the higher confidence the shorter the time between tweets.

When it comes to the tweets regarding the authorities or public figures, in cases where the tweet represented only an opinion without insult or defamation, most annotators assigned them with the non-harmful label. This was due to the general common sense that expressing an opinion is not punishable in itself. The annotators also highlighted the need for constant awareness for separating one's own worldviews from the criticisms about the authorities in order to preserve the objectivity during the annotation process. Also, although there was a clear difference between the language of supporters of the right and the left wing (e.g., "lemingi"/"lemmings" vs. "pisiory"/"PiS-supporters"), more general patterns appeared on both sides, also outside of the political topics.

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

Not harmful

1. "500+ bardzo na plus jednak ten rzad wykorzystał dorobek poprzednich rządów do swojego populizmu chorego"
2. "Mamy do czynienia z najgorszym prezydentem RP w historii. Kropka."
   — Both samples considered a general opinion. Score: 0.
3. "@anonymized_account I kurwa mamy ta wolność"
   — Despite the appearance of a strong vulgar language ("kurwa"), the expression itself does not indicate any punishable harmful action, thus the tweet was annotated as non-harmful. Score: 0.
4. "@anonymized_account @anonymized_account @anonymized_account Matka Boska była półką i Jezus też."
   — Although at a first glance this tweet might look like a blasphemy, the harmful effect is caused mostly by an error of a spellchecker ("Polka" wrongly corrected to "półka"). Score: 0.
5. "Biało-Czerwoni brawo, brawo, brawo! Zbigniew Boniek i Adam Nawałka - wyrazy szacunku. Robert Lewandowski - wielkie podziękowania!"
   — Score: 0.
6. "@anonymized_account WISŁA KRAKÓW !! brawo za dzisiejszy mecz :)"
   — Score: 0.

Cyberbullying

1. "[tel. no. anonymized] w Bułgarii numer ten uważany jest za przeklęty ponieważ podobno każdy z jego właścicieli umierali po kilku dniach"

— Considered a type of phishing. By trying to check whether a given telephone number is actually cursed (and it is fair to assume there will be such people) it is possible to fall into a trap and risk of being robbed of personal information. Score: 1.

2. "@anonymized_account Tu stary chuju PZPRowski zajmij się swoimi komuchami z PiSu."
   — A typical attack with accusations of supporting communism. Despite originally being a response to a public account, it looks like an attack at a specific person. Score: 1.

Hate-speech/other harmful

1. "@anonymized_account @anonymized_account @anonymized_account Rozumiem, że jutro w sejmie powie to pani protestującym. Załgane pisowskie skurwysyny."
   — Accumulation of profanity. Score: 2.
2. "@anonymized_account @anonymized_account Was, gnidy powinno się zaorać na metr w głąb i grubo posypać niegaszonym wapnem. A dla pewności zbombardować napalmem."
   — A typical case of hate-speech consisting of over exaggerated death threats aimed at a public person. Score: 2.
3. "KAIN TEŻ ZABIŁ BRATA ALE NIE ŚWIĘTOWAŁ TEGO CO MIESIĄC I NIE STAWIAŁ POMNIKÓW NA TĘ OKOLICZNOŚĆ."
   — A good example of a context-dependent mocking/ridiculing of a public person. Although the name of the object of ridicule is not explicitly stated, thus it is difficult to grasp by a computational method, it is easy to understand by a normal person following political events in Poland. Score: 2.
4. "@anonymized_account MILIONY POLAKÓW CZEKA NA BADANIA PSYCHIATRYCZNE LISA PO WPISACH WIDAĆ NIE ZRUWNOWARZENIE PSYCHICZNE I CIĄGŁĄ DEPRESJE"
   — The tweet, originally directed at a specific public person (a TV announcer), violates privacy and can be considered a public slander. Moreover, the expression of the need for a psychiatric checkup and the use of mental illness (depression), which typically are diagnosed by a psychiatrist, is in the area of a private matter, while in this case are used as a slur. Score: 2 (slandering of a public person).
5. "@anonymized_account Elo swastyka na ryju kiedy będzie, sorry że ciągle pytam?"
   — Although the tweet mainly represents a general negative opinion about the UK, it also focuses on a specific person (originally a right-wing publicist) accusing him of fascism. Score: 2 (accusing a public person of fascism).
6. "@anonymized_account Lzy ogromne, kiedyś usunąłem ciążę, nie mów nikomu"
   — Considered a harmful tweet written with an aim of provocation, signaled by the phrase "nie mów nikomu", since the tweet is publicly viewable, in a socially controversial topic (abortion). Score: 2.

5. Task Description

In the pilot task designed for the dataset, the participants are to determine whether an Internet entry is classifiable as harmful (cyberbullying, hate-speech) or not. The particular goal is to classify the provided tweets into cyberbullying/harmful and non-cyberbullying/non-harmful with the highest possible Precision, Recall, balanced F-score and Accuracy. There are two subtasks.

5.1. Description of Subtasks

Subtask 1. Harmful vs non-harmful

In this subtask, the participants are to distinguish between normal/non-harmful tweets (class: 0) and tweets that contain any kind of harmful information (class: 1). This includes cyberbullying, hate speech and related phenomena.

Subtask 2. Type of harmfulness

In this subtask, the participants are to distinguish between three classes of tweets: 0 (non-harmful), 1 (cyberbullying), 2 (hate-speech). There are various definitions of both cyberbullying and hate-speech, some of them even putting those two phenomena in the same group. The specific conditions on which we based our annotations for both cyberbullying and hate-speech have been worked out during ten years of research (Ptaszynski and Masui 2018). However, the main and definitive condition to distinguish the two is whether the harmful action is addressed towards a private person(s) (cyberbullying), or a public person/entity/larger group (hate-speech). Other specific definitions and guidelines applied in creation were described in Section 3.

5.2. Evaluation

The scoring for the first subtask is done based on standard Precision (P), Recall (R), Balanced F-score (F1) and Accuracy (A), on the basis of the numbers of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), according to the below equations. The winning condition would be to have the highest balanced F-score. However, in the case of F-score equal for two or more participants, the one with higher Accuracy would be considered as the winner. Furthermore, in case of the same F-score and Accuracy, a priority shall be given to the results as close as possible to BEP (break-even-point of Precision and Recall).

$$Precision = \frac{TP}{TP + FP}$$

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

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

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

The scoring for the second subtask is based on two measures, namely, Micro-Average F-score (*microF*) and Macro-Average F-score (*macroF*). Micro-Average F-score is calculated similarly as in standard equation for F-score, but on the basis of Micro-Averaged Precision and Recall, which are calculated according to the below equations. Macro-Average F-score is calculated on the basis of Macro-Averaged Precision and Recall, which are calculated according to the following equations. The winning condition would mean at first the highest *microF*. This measure treats all instances equally, which is a fair approach since the number of instances is different for each class. However, in the case of equal results for

5.3. Task Participants

There were fourteen overall submissions to the task sent by nine unique teams. All the submitting teams attempted to solve the first subtask (6-1), which was a computationally simpler problem of binary classification of tweets into harmful and non-harmful, while there were only eight attempts at solving the second subtask (6-2), which was the three-class classification problem. Below we briefly describe the systems proposed by each team. All teams and submitted systems were summarized in Tables 4 and 5. Below we present short descriptions of teams and systems, for which the authors decided to describe their systems in this volume.

Korzeniowski et al. (2019) from Sigmoidal team presented three approaches, namely, fine-tuning of a pre-trained ULMFiT, fine-tuning a pre-trained BERT model, and using the TPOT library to find the optimal pipeline. The last of the proposed approaches, namely, TPOT with a logistic regression classifier with non-trivial feature engineering, scored as second in the Subtask 6-2 (detection of different types of harmful information).

Wróbel (2019) after, firstly preprocessing the data, tested two classifiers, namely, Flair trained on character-based language model and FastText.

Prońko (2019) compared some of the popular text classification models, such as Ngrams and MLP, word embedding and sepCNN, and Flair with different embeddings, in combination with LSTM and GRU with word embeddings trained from scratch.

Ciura (2019) applied Przetak, a tool wich identifies abusive and vulgar speech in Polish, to detect cyberbullying. Przetak is a dynamically-linked library written in Go, which uses logistic regression over character 5-grams of words. This approach scored as second in the first subtask (6-1) on detection of any type of harmful information.

Biesek (2019) presented three approaches with different architectures and level of complexity, namely, a standard machine learning SVM classifier with TF-IDF, a bidirectional GRU network, and a deep Flair framework model with Contextual String Embeddings. The model applying SVM outperformed all other submissions for Subtask 6-2.

Krasnowska-Kieraś and Wróblewska (2019) proposed a simple neural network setup with various feature sets, including LASER embeddings, stylistic qualifiers signalling various informal modifications (e.g., vulgar, colloquial, depreciating, etc.), a list of offensive words, and character n-grams. To supplement for the imbalanced data samples they also divided separate tweets into separate sentences with full stop, and added artificially created back-translations (Polish-Russian-Polish, etc.) of tweets containing insufficient number of classes.

Czapla et al. (2019) from n-waves team based their approach on transfer learning, which uses large amounts of unlabelled text to reduce data necessary for a target task. They also showed that initial weights of language model play an important role in model performance on the target task, and proposed a mechanism to test if the sampled initial weights are suitable for the target task. Their solution proposed for Subtask 6.1 achieved state-of-the-art performance and took first place.

6. Results of First Shared Task for Automatic Cyberbullying Detection in Polish Twitter

6.1. Baselines

The dataset was not balanced, namely, the ratio of each class was different (see Table 1). Therefore to get a more objective view on how participants of the task managed to classify the data, we first prepared a number of simple baselines.

The first set of baselines consisted of simple classifiers assigning scores without any insight into data:

A. classifier always assigning score 0

B. classifier always assigning score 1

C. classifier always assigning score 2 (only for Subtask 2)

D. classifier assigning random score: 0/1 (for Subtask 1)

E. classifier assigning random score: 0/1/2 (for Subtask 2).

As a result, all simple baselines scored very low. For Subtask 1, baseline A (always 0) scored $F1 = 0$, which was predictable and simply means it is not possible to simply disregard the problem as too easy. Baseline D (random) also scored F1=0, which additionally means that it is not possible to solve to problem of cyberbullying detection by simply flipping a coin. Baseline B (always 1), by the definition, was able to catch all harmful samples (Recall = 100%), but such a simplistic assumption results in a very low Precision (13.4%), thus causing the F-score to be also very low (23.63%).

As for the second subtask, for the same reasons as in subtask 1, baselines B (always 1), C (always 2), and E (random) also achieved very low scores. Baseline A (always 0), achieved a high *microF* (86.6%) due to automatically winning for non-harmful cases, which were the

majority in the dataset. However, *macroF* provided a sufficient clarification of the score, is in fact very low (30.94%).

Table 3: Results of simple baselines for Subtask 1

Subtask 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%

Table 4: Results of simple baselines for Subtask 2

Subtask 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%

6.2. Results of Task Participants

Subtask 6-1

In the first subtask, out of fourteen submissions, there were nine unique teams: n-waves, Warsaw University of Technology, Sigmoidal, CVTimeline, AGH & UJ, IPI PAN, UWr, and two independent researchers. Some teams submitted more than one system proposal, in particular: Sigmoidal (3 submissions), independent (3 by one researcher), CVTimeline (2). Participants used a number of various techniques, usually widely available open source solutions, trained and modified to match the Polish language and the provided dataset when it was required. Some of the methods used applied, e.g., fast.ai/ULMFiT (`http://nlp.fast.ai/`), SentencePiece (`https://github.com/google/sentencepiece`), BERT (`https://github.com/google-research/bert`), tpot (`https://github.com/EpistasisLab/tpot`), spaCy (`https://spacy.io/api/textcategorizer`), fasttext (`https://fasttext.cc/`), Flair (`https://github.com/zalandoresearch/flair`), neural networks (in particular with GRU) or more traditional SVM. There were also original methods, such as Przetak (`https://github.com/mciura/przetak`). The most effective approach was based on recently released ULMFiT/fast.ai, applied for the task by the n-waves team. The originally proposed Przetak was second-best, while third place achieved a combination of ULMFiT/fast.ai, SentencePiece and BranchingAttention model. The results for of all teams participating in Subtask 6-1 were represented in Table 5.

Subtask 6-2

In the second subtask, out of eight submissions, there were five unique submissions. The teams that submitted more than one proposal were: independent (3 submissions) and Sigmoidal (2). Methods that were the most successful for the second subtask were based on: svm (winning method proposed by independent researcher Maciej Biesek), a combination of ensemble of classifiers from spaCy with tpot and BERT (by Sigmoidal team), and fasttext (by the AGH & UJ team). The results for of all teams participating in Subtask 6-2 were represented in Table 6. Interestingly, although the participants often applied new techniques, most of them applied only lexical information represented by words (words, tokens, word embeddings, etc.), while none of the participants attempted more sophisticated feature engineering and incorporate other features such as parts-of-speech, named entities, or semantic features.

7. Conclusions

We presented the first dataset in the Polish language, together with an open shared task for automatic cyberbullying detection, to contribute to solving the recently growing problem of cyberbullying and hate-speech appearing on the Internet.

The dataset, together with the open shared task supplementing 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 only applied preprocessing was for anonymization of mentions so private persons mentioned in tweets were not revealed to the public.

The goal of the main subtask was to classify the tweets into harmful (cyberbullying or hate-speech) and non-harmful with the highest possible Precision, Recall, balanced F-score and Accuracy. In an additional subtask, the goal was to differentiate between various types of harmful information, in particular cyberbullying and hate-speech, as well as non-harmful.

There were fourteen submissions from nine unique teams. All submissions attempted to solve the first binary classification subtask, while only eight submissions were for the second subtask. The participants mostly used widely available solutions for text classification, such as fast.ai/ULMFiT, SentencePiece, BERT, spaCy, fasttext, or more traditional SVM. Original methods were in minority, although appeared quite successful. Best methods were based, either on recently proposed solutions (fast.ai) or original methods (Przetak) for the first subtask, as well as more traditional machine learning methods (SVM) for the second subtask.

8. Future Work

As this was the first task of this kind for the Polish language, and one of the few first in general, we acknowledge that there is room for improvement. In particular, we plan on enlarging the dataset. At this time the dataset contains 11 thousand tweets with only about 9% of harmful

Table 5: Results of participants for Subtask 6-1

Submission author(s)	Affiliation	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	independent	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, Przemysław Sadowski, Rafał Rolczyński, Tomasz Korbak, Marcin Możejko, Krystyna Gajczyk)	Sigmoidal	ensamble spacy + tpot + BERT	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team	Sigmoidal	ensamble + fastai	52.71%	50.75%	51.71%	87.30%
Sigmoidal Team	Sigmoidal	ensenble spacy + tpot	43.09%	58.21%	49.52%	84.10%
Rafał Prońko	CVTimeline	Rafal	41.08%	56.72%	47.65%	83.30%
Rafał Prońko	CVTimeline	Rafal	41.38%	53.73%	46.75%	83.60%
Maciej Biesek	independent	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-Kieraś, Alina Wróblewska	IPI PAN	SCWAD-CB	51.90%	30.60%	38.50%	86.90%
Maciej Biesek	independent	model2-gru	63.83%	22.39%	33.15%	87.90%
Maciej Biesek	independent	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 6: Results of participants for Subtask 6-2

Submission author(s)	**Affiliation**	**Name of the submitted system**	**Micro-Average F-score**	**Macro-Average F-score**
Maciej Biesek	independent	**model1-svm**	**87.60%**	**51.75%**
Sigmoidal Team (Renard Korzeniowski, Przemysław Sadowski, Rafał Rolczynski, Tomasz Korbak, Marcin Możejko, 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-Kieraś, Alina Wróblewska	IPI PAN	SCWAD-CB	83.70%	49.47%
Maciej Biesek	independent	model2-gru	78.80%	49.15%
Jakub Kuczkowiak	UWr	Task 6: Automatic cyberbullying detection	70.40%	37.59%
Sigmoidal Team	Sigmoidal	ensamble + fastai	61.60%	39.64%

ones. In the future, we plan to at least double the size to contain at least a comparable number of harmful tweets, as in research in other languages (Ptaszynski and Masui 2018, Cano et al. 2013, Dinakar et al. 2012). We also need to improve the procedure for the preprocessing of the dataset to make sure no noise or redundant information is contained. In particular, the present dataset contained a number of unofficial retweets (tweets starting with RT). A thorough analysis also revealed some remaining tweets with unusual URLs, which slipped through the URL filtering stage.

Moreover, in a future version of the dataset we also plan to annotate on the tweets roles of participants in cyberbullying, such as: 1) victim, 2) bully and 3) bystanders (3–1 bully-supporter, and 3–2 victim–defender) to get a wider grasp on the problem of bullying as a process taking place on the Internet.

Finally, when it comes to the classification methods, although the participants used new widely available techniques, only lexical information was applied (words, tokens, word embeddings, etc.). Since it has been shown that a thorough feature engineering is useful in cyberbullying detection (Ptaszynski et al. 2017), we encourage future participants to incorporate other features, except words/tokens, e.g., parts-of-speech, named entities, or semantic features.

References

Biesek M. (2019). *Comparison of Traditional Machine Learning Approach and Deep Learning Models in Automatic Cyberbullying Detection for Polish Language*. In Ogrodniczuk and Kobyliński (2019), pp. 121–126.

Cano E., He Y., Liu K. and Zhao J. (2013). *A Weakly Supervised Bayesian Model for Violence Detection in Social Media*. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan.

Ciura M. (2019). *Przetak: Fewer Weeds on the Web*. In Ogrodniczuk and Kobyliński (2019), pp. 127–133.

Czapla P., Gugger S., Howard J. and Kardas M. (2019). *Universal Language Model Fine-Tuning for Polish Hate Speech Detection*. In Ogrodniczuk and Kobyliński (2019), pp. 149–159.

Dinakar K., Jones B., Havasi C., Lieberman H. and Picard R. (2012). *Commonsense Reasoning for Detection, Prevention and Mitigation of Cyberbullying*. „ACM Transactions on Intelligent Interactive Systems", 2(3).

Dooley J. J., Pyżalski J. and D. C. (2009). *Cyberbullying Versus Face-to-Face Bullying: A Theoretical and Conceptual Review*. „Zeitschrift für Psychologie/Journal of Psychology", 217(4), p. 182–188.

Hatakeyama S., Masui F., Ptaszynski M. and Yamamoto K. (2015). *Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information*. In *Proceedings of the Demo Session of The 2015 ACM Symposium on Cloud Computing 2015 (ACM-SoCC 2015)*, Kohala Coast, Hawai'i.

Hatakeyama S., Masui F., Ptaszynski M. and Yamamoto K. (2016a). *Statistical Analysis of Automatic Seed Word Acquisition to Improve Harmful Expression Extraction for Cyberbullying Detection*. „Proceedings of the International Conference on Advanced Technology Innovation 2016 (ICATI2016)".

Hatakeyama S., Masui F., Ptaszynski M. and Yamamoto K. (2016b). *Statistical Analysis of Automatic Seed Word Acquisition to Improve Harmful Expression Extraction in Cyberbullying Detection*. „International Journal of Engineering and Technology Innovation", 6(2), p. 165–172.

Korzeniowski R., Rolczyński R., Sadownik P., Korbak T. and Możejko M. (2019). *Exploiting Unsupervised Pre-training and Automated Feature Engineering for Low-resource Hate Speech Detection in Polish*. In Ogrodniczuk and Kobyliński (2019), pp. 141–148.

Krasnowska-Kieraś K. and Wróblewska A. (2019). *A Simple Neural Network for Cyberbullying Detection*. In Ogrodniczuk and Kobyliński (2019), pp. 161–163.

Lempa P., Ptaszynski M. and Masui F. (2015). *Cyberbullying Blocker Application for Android*. In *Proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions (EDO 2015)*, pp. 408–412, Poznań, Poland.

Nitta T., Masui F., Ptaszynski M., Kimura Y., Rzepka R. and Araki K. (2013a). *Cyberbullying Detection Based on Category Relevance Maximization*. In *Proceedings of the Demo Session of 20th International Conference on Language Processing and Intelligent Information Systems (LP & IIS 2013)*, Warsaw, Poland.

Nitta T., Masui F., Ptaszynski M., Kimura Y., Rzepka R. and Araki K. (2013b). *Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization*. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp. 579–586, Nagoya, Japan.

Ogrodniczuk M. and Kobyliński Ł., editors (2019). *Proceedings of the PolEval 2019 Workshop*, Warsaw. Institute of Computer Science, Polish Academy of Sciences.

Prońko R. (2019). *Simple Bidirectional LSTM Solution for Text Classification*. In Ogrodniczuk and Kobyliński (2019), pp. 111–119.

Ptaszynski M., Dybala P., Matsuba T., Masui F., Rzepka R., Araki K. and Momouchi Y. (2010a). *In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis*. „International Journal of Computational Linguistics Research", 1(3), p. 135–154.

Ptaszynski M., Dybala P., Matsuba T., Masui F., Rzepka R. and Araki K. (2010b). *Machine Learning and Affect Analysis Against Cyber-Bullying*. In *Proceedings of The 36th Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB-10)*, pp. 7–16. De Montfort University, Leicester, UK.

Ptaszynski M., Masui F., Kimura Y., Rzepka R. and Araki K. (2015). *Extracting Patterns of Harmful Expressions for Cyberbullying Detection*. In *Proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*

(LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions (EDO 2015), pp. 370–375, Poznan, Poland.

Ptaszynski M., Masui F., Nakajima Y., Kimura Y., Rzepka R. and Araki K. (2016a). *Detecting Cyberbullying with Morphosemantic Patterns*. In *Proceedings of the Joint 8th International Conference on Soft Computing and Intelligent Systems and 17th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2016)*, pp. 248–255, Sapporo, Japan.

Ptaszynski M., Masui F., Nitta T., Hatakeyama S., Kimura Y., Rzepka R. and Araki K. (2016b). *Sustainable Cyberbullying Detection with Category-Maximized Relevance of Harmful Phrases and Double-Filtered Automatic Optimization*. „International Journal of Child-Computer Interaction (IJCCI)", 8, p. 15–30.

Ptaszynski M., Kalevi J., Eronen K. and Masui F. (2017). *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.

Ptaszynski M., Masui F., Kimura Y., Rzepka R. and Araki K. (2018). *Automatic Extraction of Harmful Sentence Patterns with Application in Cyberbullying Detection*. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pp. 349–362. Springer International Publishing. Lecture Notes in Computer Science (LNCS) vol. 10930.

Ptaszynski M., Masui F., Kimura Y., Rzepka R. and Araki K. (2019). *Brute Force Sentence Pattern Extortion from Harmful Messages for Cyberbullying Detection*. „Journal of the Association for Information Systems (JAIS)".

Ptaszynski M. E. and Masui F. (2018). *Automatic Cyberbullying Detection: Emerging Research and Opportunities*. IGI Global Publishing.

Tversky A. and Kahneman D. (1974). *Judgment under Uncertainty: Heuristics and Biases*. „Science, 185(4157)", pp. 1124–1131.

Wróbel K. (2019). *Approaching Automatic Cyberbullying Detection for Polish Tweets*. In Ogrodniczuk and Kobyliński (2019), pp. 135–140.