NLP - Hate speech classification - Milestone 1

Adrianna Klimczak, Adam Kowalczyk and Marcel Wenka

March 28th, 2022

1 Introduction

The goal of this project is to build a machine learning model for automatic cyberbullying detection and classification of the hate speech. The goal of this first milestone is to present the findings of literature and dataset analysis along with the proposed solution of the problem.

2 Problem overview

While social media offer many great communication opportunities, they also increase the vulnerability of young people to threatening situations online. Recent studies [11] shows that among youngsters cyberbullying constitutes an ever growing problem. Hence, automatic cyberbullying detection is also a task of rapidly growing interest. In recent years, the interest in this topic grew significantly in the Natural Language Processing and Machine Learning communities. It is both a challenging and extremely relevant problem. Especially when one takes into consideration how social networks have become a vital part of our lives and how dire the consequences of cyberbullying can be, especially among adolescents [11].

The goal of this project is to build a machine learning model for automatic cyberbullying detection and hate speech classification for polish language. The data will contain tweets collected from openly available Twitter discussions and forums. Final model should be able to distinguish the three possible classes of tweets:

1. 0 (non-harmful)
2. 1 (cyberbullying)
3. 2 (hate-speech)

Since there are many different definitions of hate-speech and cyberbullying there is a need to define the difference between them. The specific conditions on which we will be basing the annotations have been worked in the paper [9]. The main difference that distinguishes the cyberbullying from the hate-speech is towards whom the tweet is addressed. If it is addressed towards a private person then it will be labelled as cyberbullying and if it is addressed towards a public person/entity/large group it will be considered a hate speech [1].

The four basic phases in the hate speech classification problem are:

1. Preprocessing phase
2. Data representation phase
3. Detection phase
4. Classification phase

The first phase and perhaps one of the most important ones is the preprocessing. Among many things it involves: noise separation, labelling the data and extracting the features from raw data [2]. Twitter is a medium that has a lot of messages that can be meaningless so it it crucial to sort them out.

In the data representation phase the creation of model's tweet-topics occurs. It involves the categorisation of tweets into topics using possible different machine learning algorithms. The data preprocessed in this way is used as the training data in the detection phase's algorithms.

In the detection phase the clustering of tweets into the correct topic clusters takes place. The fourth phase is the classification phase and it can contain many possible methods and machine learning algorithms. This phase involves the analysis of the hate speech topics detected earlier and then the distinction of different classes [2].

3 Literature analysis

There is no doubt that research on hate speech classification is much more narrow compared to the social studies of this phenomenon. However, many important advances have been made in recent years. In this chapter, a brief overview of some of the most important natural language processing approaches to hate speech classification is being presented. For a more detailed overview of this problem we refer to the survey paper [2].

3.1 Convolutional neural network (CNN)

One of the most common methods used in NLP for hate speech classification are convolution neural networks. For example, Kern and Winter [13] have developed such model used for multilingual hate speech detection of tweets. The CNN was deployed in order to detect hate speech in Spanish or in English in the messages from Twitter. For feature extraction the model used word embedding technique. In the first place the model detected whether a tweet was hateful speech or not and then it decided on the severity level and target of the tweet. The CNN described in the paper produced very good accuracy and if compared to other baseline classifiers it performed better.

Another such example is Ribeiro and Silva [10]. In the paper they have proposed also a CNN architecture. The model was used for hate speech detection against immigrants and women on Twitter using pretrained word embeddings (GloVe and FastText). Here the model was also multilingual, also for tweets written in English or in Spanish. The task also consisted of two classification tasks. The first was to predict if a multilingual tweet that was targeted against women or immigrants was hateful or not. The second task was whether the target of hate speech was a group of individuals or a single individual. CNN's architecture turned out to perform better for tweets written in Spanish.

3.2 Kernel method (SVM)

Another possible method used for hate speech classification is SVM. For example, it was described by Vega et al. [12]. In the paper they have implemented a SVM classifier to detect and classify hate speech against immigrants and women in Twitter. It was the same problem that was described in [10]. For feature extraction SVM used features representation model. Then the extracted features served as input into the SVM classifier. Superior performance of SVM have been noticed when compared to similar algorithms. Malmasi et al. [6] also described a system that was ensemble-based and used linear SVM classifiers in parallel. The problem was to distinguish and classify hate speech and profanity generally in social media.

3.3 Deep learning (DL)

Deep learning methods have been widely used in data mining and text classification fields, also for detection, classification and prediction of events like hate speech detection. For example, Komal Florio et al. [3] presented a deep learning approach for hate speech identification in Twitter data. Their approach included CNN and RNN classifier by using biGRU. The results presented in the paper showed that this approach could outperform other classifiers such as logistic regression.

3.4 Embedding and deep learning (EMB-DL)

Another possible approach to hate speech classification in Twitter data is to combine word embeddings and deep learning architectures. Word embedding is a specific technique for language modelling. In natural language processing word embedding performs vector representation of words context. There are many word embeddings techniques, some of which include fastText, BERT, Count Vectorizer, Hashing Vectorizer, TF–IDF Vectorizer, Word2Vec etc. Word embedding is used in order to enhance baseline classifiers' performance in fields such as sentiment classification and data analysis.

3.5 Competition results

Specific methods that were the most successful in the competition regarding this tasks (as mentioned in [8]) were based on:

1. SVM
2. a combination of ensemble of classifiers from spaCy with tpot and BERT
3. fasttext

Worth mentioning is the fact that most of the participants used mainly lexical information represented by words (words, word embeddings,tokens, etc.). More sophisticated methods such as feature engineering or incorporating other features such as named entities, parts-of-speech or semantic features were not being applied [8].

4 Dataset overview

The dataset for this project comes from the PolEval competition, 2019, Task 6-2. It contains tweets collected from openly available Twitter discussions and forums. These tweets are subdivided into 3 categories:

1. 0 (non-harmful)
2. 1 (cyberbullying)
3. 2 (hate-speech)

4.1 Data acquisition

First the data was downloaded from the PolEval website and decompressed by using the following commands:

```
wget "http://2019.poleval.pl/task6/task_6-2.zip"
unzip "/content/task_6-2.zip" -d "/content/data"
```

Then it was loaded in python using the pandas library:

```
original_tags = pd.read_csv(
    '/content/data/training_set_clean_only_tags.txt', header = None)
with open('/content/data/training_set_clean_only_text.txt', 'r') as file:
    lines = file.readlines()
original_text = pd.DataFrame(lines)
df = pd.concat([original_text,original_tags], axis=1)
df.columns =['Text', 'Tag']
```

4.2 Basic dataset information

The data contains 10041 rows of text and assigned classes. The classes distribution presents as follows:

Table 1: Class distribution

Class	Count
0	9190
1	253
2	598

Over 90% of the instances are non-harmful. 2.5% are examples of cyberbullying and the remaining 6% is hate-speech.

4.3 Word counts

The tweets differ in length. Actually, on average the lengths are different in different classes. Figure 1 shows word counts per class (the class is included in the title of the graph).

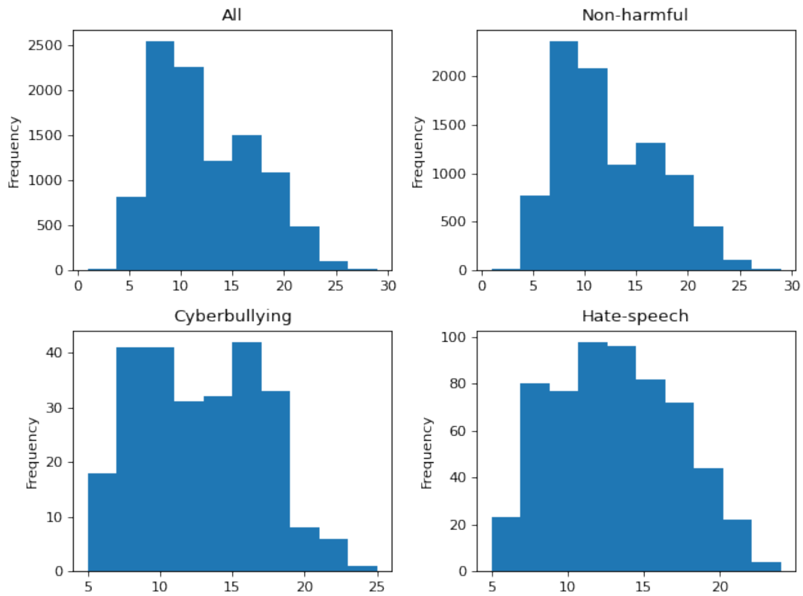

Figure 1: Word counts per class.

On average hate-speech and cyberbullying was composed of more words than the non-harmful tweets. The variation is rather similar. Notably, cyberbullying and hate-speech was never composed of less than 5 words, in contrast to non-harmful tweets, which were shorter on a few occasions.

4.4 Text length

All tweets followed a bell curve (normal distribution) when it comes to the length of text. Figure 2 shows text lengths per class (similarly to before, the class is included in the title of the graph).

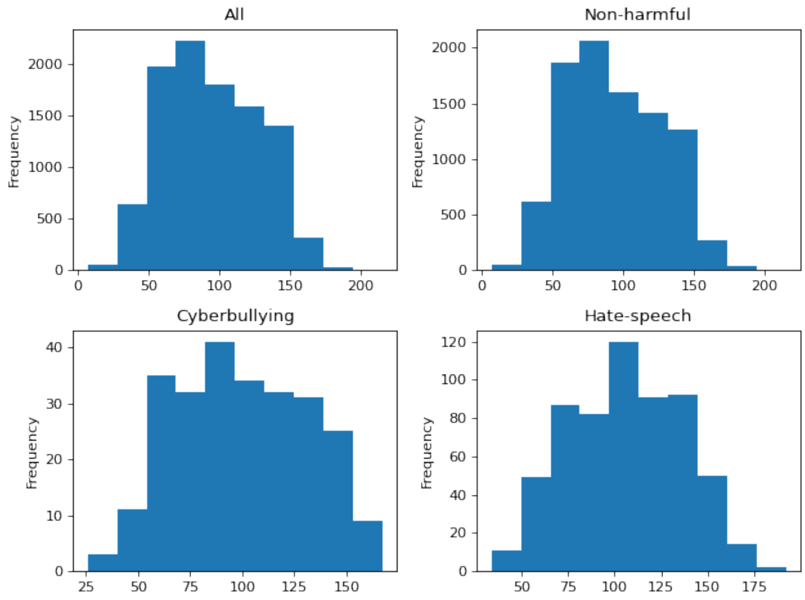

Figure 2: Text lengths per class.

Text lengths of cyberbullying and hate-speech had a bit more variation with means at around 100 characters, whereas non-harmful tweets were usually shorter with the distribution being slightly left-skewed.

4.5 Word lengths

The distributions of word lengths were all left skewed as there were more shorter words. The distributions are shown in Figure 3.

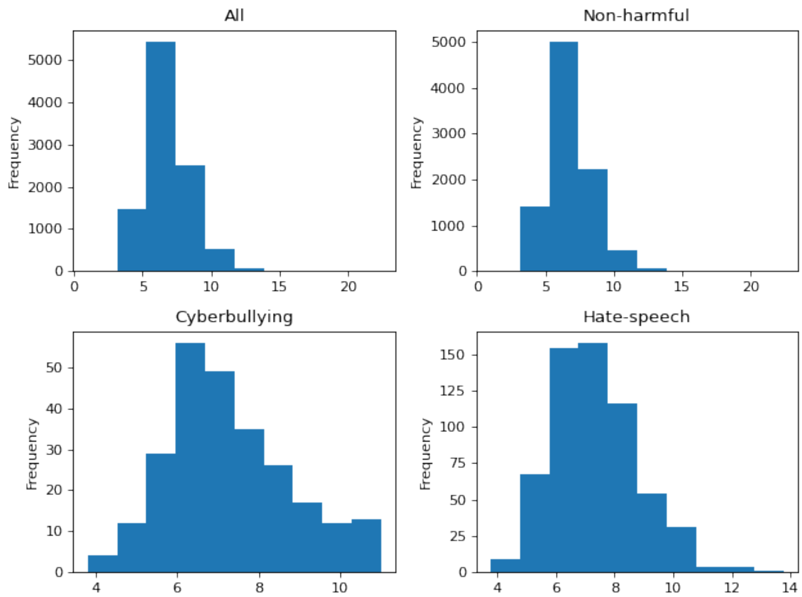

Figure 3: Word lengths per class.

Non-harmful tweets were mostly composed of 5-8 letter word as these are the most common word lengths in polish language. [7] Noticeably, cyberbullying contained more lengthier words. Again, the variation in cyberbullying and hate-speech was greater than in non-harmful tweets.

4.6 Stop words

The words which are generally filtered out before processing a natural language are called stop words. These are actually the most common words in any language (like articles, prepositions, pronouns nd conjunctions) and they do not add much information to the text. By removing these words, the low-level information is removed from text in order to give more focus to the important information. In other words, the removal of such words should not have any negative consequences on the model. [5]

The most common stop words in the dataset where presented in Figure 4. These stop words were extracted using the python library spacy with locale set to polish (as shown in the code snipped below).

```
from spacy.lang.pl import Polish
nlp = Polish()
stopwords_count = []
for index, row in df.iterrows():
  for tok in nlp(row['Text']):
    if tok.is_stop:
      stopwords_count.append(tok.lower_)
```

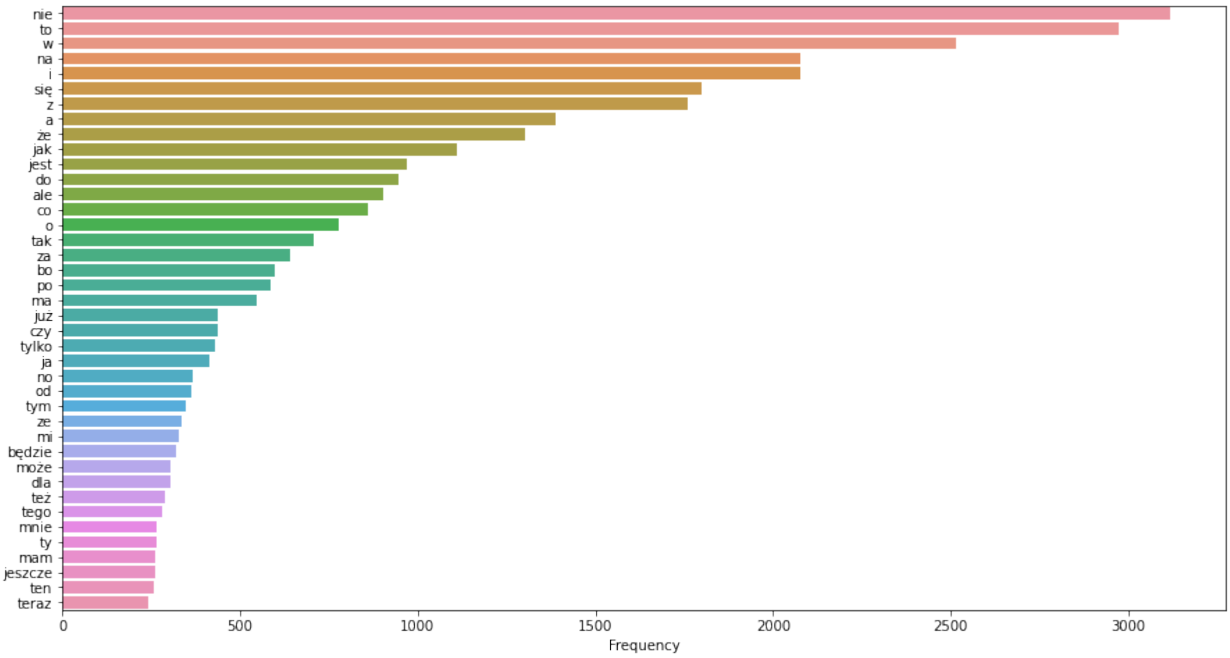

Figure 4: Most common stop words.

4.7 Non stop words

Non stop words are the words that remain after filtering out the stop words and punctuation. As tweets are very concise, because of their limit of 280 characters (which before November 2017 was 140) [4], they contain a lot of emoticons. Figure 5 shows the most common non stop words in all classes.

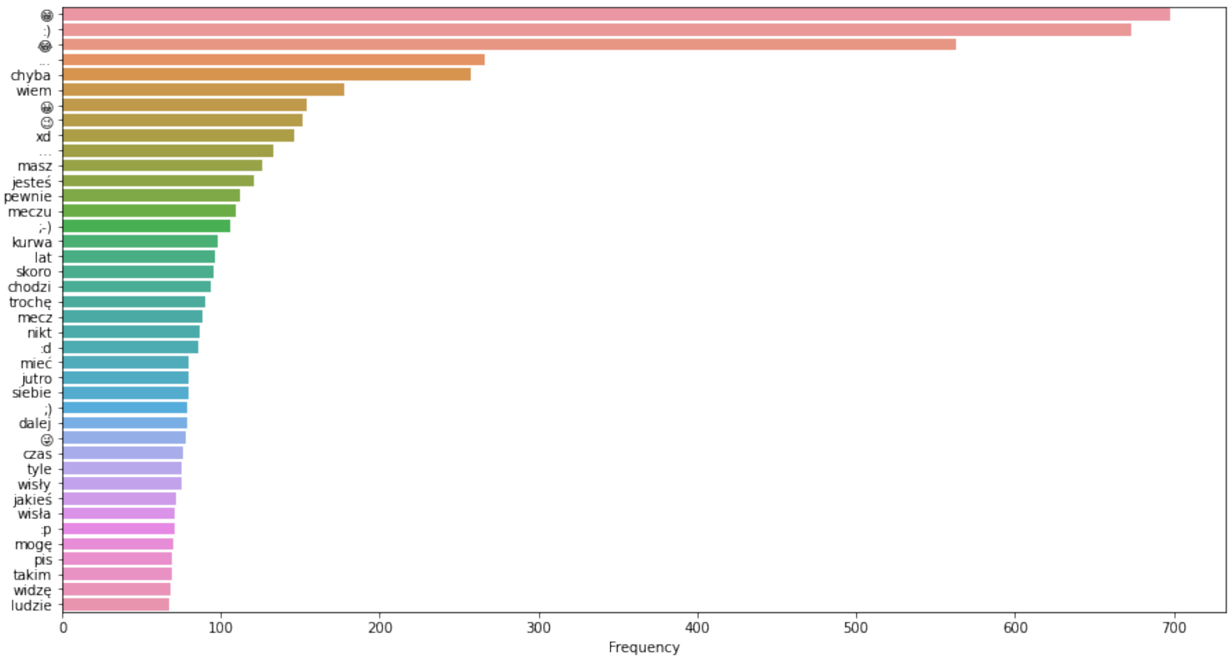

Figure 5: Most common non stop words in all classes.

Notably, some of these most common non stop words are swear words, even though over 90% of the tweets were not harmful. Additionally, it seems like spacy does not treat suspension points as punctuation as they appear as one of the most used tokens.

Figure 6 shows most common non stop words for cyberbullying and hate-speech.

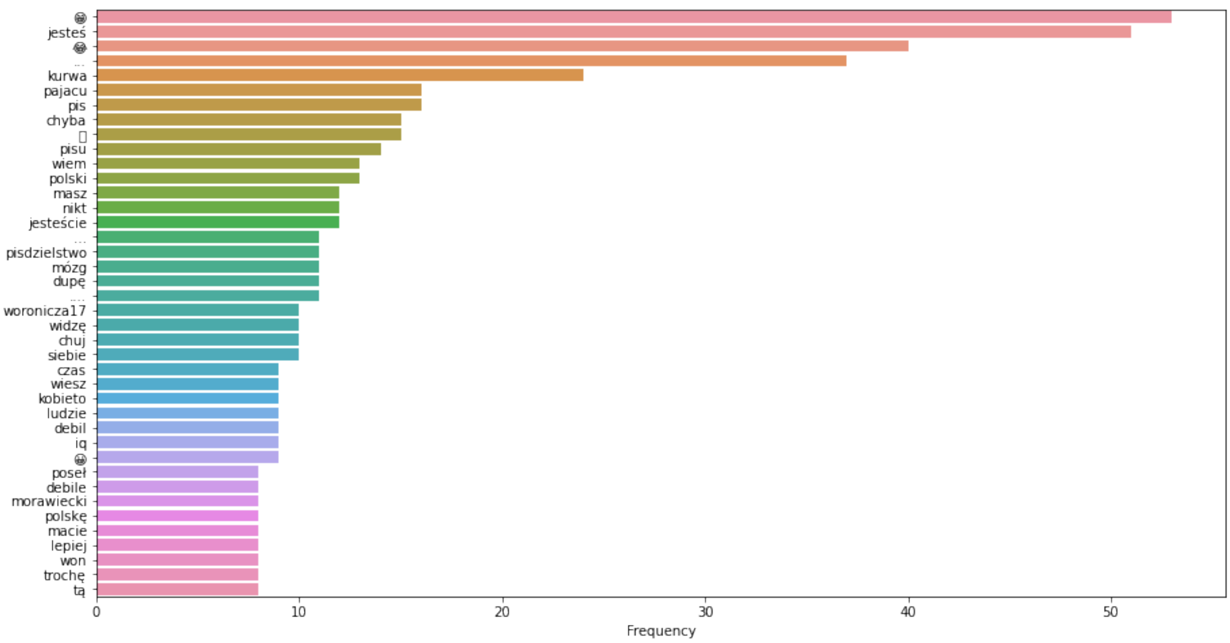

Figure 6: Most common non stop words in cyberbullying and hate-speech.

Obviously, there are much more swear words and they upper way up the list. Additionally, there appear words connected with politics:

- ”PiS” – ruling party in Poland
- ”Woronicza 17” – address of the national TV broadcaster
- ”poseł” – member of parliament
- ”Morwiecki” – polish prime minister

4.8 Preprocessing

Preprocessing included a few steps:

1. removing whitespaces
2. tokenizing the text
3. removing retweets (marked with the ”RT” keyword)
4. removing mentions to other accounts starting or ending the tweet
5. removing punctuation
6. removing stop words
7. converting the text to lowercase
8. removing emoticons

The code for all aforementioned steps is presented below:

```
text = ' '.join(text.split())

doc = [tok for tok in nlp(text)]

if str(doc[0]) == "RT":
    doc.pop(0)

while str(doc[0]) == "@anonymized_account":
    doc.pop(0)
while str(doc[-1]) == "@anonymized_account":
    doc.pop()

doc = [t for t in doc if t.text not in string.punctuation]

doc = [tok for tok in doc if not tok.is_stop]

doc = [tok.lower_ for tok in doc]

doc = [RE_EMOJI.sub(r'', str_text) for str_text in doc]
```

5 Proposed solution

We plan on using word embeddings techniques such as:

- Count Vectorizer
- TF–IDF Vectorizer
- Word2Vec

Then we want to compare several machine learning models. We are certainly going to try support-vector machine and convolutional neural networks. We would like to analyse these techniques and come up with the most accurate combination. We might also try some ensembling methods to increase the model accuracy.

Additionally, we do not rule out modifying the preprocessing steps as possibly some additional tweets might get removed and some tokens might be added (for example the @ mention).

References

[1] PolEval 2019. Task 6: Automatic cyberbullying detection, 2019.

[2] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions.

[3] Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12), 2020.

[4] Kristina Gligoric, Ashton Anderson, and Robert West. Adoption of twitter's new length limit: Is 280 the new 140? *CoRR*, abs/2009.07661, 2020.

[5] Chetna Khanna. Text pre-processing: Stop words removal using different libraries, 2021.

[6] S. Malmasi and M. Zampieri. Challenges in discriminating profanity from hate speech., 2018.

[7] Tomasz Moździerz. Długość przecietnego polskiego wyrazu w tekstach pisanych w świetle analizy korpusowej. *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców*, 27:177–192, December 2020.

[8] Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.

[9] Michal E. Ptaszynski and Fumito Masui. Automatic cyberbullying detection: Emerging research and opportunities, 2018.

[10] A. Ribeiro, N. Silva, and F-HatEval. Semeval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[11] H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, and I. Trancoso. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345, 2019.

[12] E.A. Vega, J.C. Reyes-Magaña, H. Gómez-Adorno, and G. Bel-Enguix and. Mineriaunam semeval-2019 task 5: Detecting hate speech in twitter using multiple features in a combinatorial framework, 2019.

[13] K. Winter, R. Kern, and Know-center. Semeval-2019 task 5: Multilingual hate speech detection on twitter using cnns. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.