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 [10] 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. 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 [10].

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 be 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 [8]. 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 [4]. 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. [4].

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 [4].

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 [6] 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 [2]. 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. [3]. 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 [2]. 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. [7] 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. [5] 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 achieved 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 [9]) 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 [9].

4 Dataset overview

5 Proposed solution

References

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

[2] F-HatEval A. Ribeiro, N. Silva. 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.

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

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

[5] 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.

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

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

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

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

[10] 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.