Analysing targets of hate speech on X in the Netherlands using BERT-CNN

Rick de Vries
University of Twente
Enschede, The Netherlands
r.devries-13@student.utwente.nl

ABSTRACT

Hate speech poses a challenge to respect and inclusivity, impacting both individuals and society as a whole. Social media platforms like X (formerly Twitter) and Facebook have made it much easier to express hate speech (anonymously), making hate speech detection an important goal. Research on hate speech on social media platforms has been performed in other countries, especially the USA. However, research on hate speech on X in the Netherlands is minimal, focusing mainly on the effects rather than the targets. This is an important motivation: to explore this research field and provide recommendations regarding hate speech target classification models. X is a suitable platform for hate speech analysis since it is one of the most popular social media platforms mainly about giving opinions and interacting with others. There are many models which can be used to detect hate speech, but this research uses a BERT-CNN model since current research indicates that is outstanding in understanding the context of text. Hate speech identification and target prediction models have been created for the IMSyPP project, but have not been used to analyse X posts on a large scale. In this research, a model is trained on a labelled dataset from the IMSyPP project. This research analyses the targets of hate speech on X in the Netherlands, to more clearly understand hate speech in the Netherlands. This contributes to society and science since it provides insights into hate speech targets and how to train the classifying models.

KEYWORDS

BERT-CNN, online hate speech, Twitter (X), Netherlands, hate speech detection

1 INTRODUCTION

Hate speech has always been a problem, not only affecting small or minority groups [1]and harming them on an individual level, but also society as a whole [2]. According to [1], exposure to hate speech leads to greater stress expression. Being targeted at individuals or groups based on race, ethnicity, religion, gender, sexual orientation or disability, hate speech poses a serious challenge to online respect and inclusivity. In the age of digital connectivity provided by social media like X (formerly Twitter) and Facebook, it is much easier to communicate with masses of people. Therefore, spreading hate speech has become a larger problem than it already was.

University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.
TScIT 41, July 5, 2024, Enschede, The Netherlands

1.1 Background

It is clear that with the amount of daily tweets posted, manually filtering tweets is out of the question. According to Mohiyaddeen and Siddiqui [3], the manual method of detecting and eliminating hate speech posts or comments is time-consuming and computationally expensive. Because of these issues and the prevalence of hateful content on social media, there is a strong case for automated hate speech identification. Various techniques (BERT [4], SVM [5], VADER [6], GPT [7], to name a few) have already been used in an attempt to automatically detect online hate speech. However, an important part of social media and a valued human right is freedom of speech. If a method of hate speech detection has insufficient precision, it could severely impact online freedom of speech.

One such model to detect hate speech is BERT (Bidirectional Encoder Representations from Transformers), which since its release has become the industry standard in many automated word processing tasks [8]. Transformers are a type of AI model that can understand and process text by focusing on important words and their connections, even if they are far apart in the sentence. BERT is outstanding in understanding the context of text and it can be enhanced with other models for even better performance. The authors of [9] show that combining a CNN (Convolutional Neural Network) with BERT is better than using BERT alone. CNNs excel at extracting features from data. For that reason and its ease of use, this research is also based on a BERT-CNN model.

Earlier hate speech classification attempts either extracted the targets using a dictionary approach [4] or did not focus on targets [10], [11]. The latter two only focus on whether a tweet is considered hate speech or not and do not identify the target. This research uses a CNN-based approach to finding the targets, which to the best of our knowledge, has not been done in the Netherlands. Therefore, the goal of this work is to train a BERT-CNN model to find the targets of hate speech in the Netherlands and provide recommendations regarding this model structure and the datasets which are used.

1.2 Research Questions

This paper investigates online hate speech in the Netherlands since hate speech analysis in this country on social media is very minimal. The goals are to train a BERT-CNN model to classify hate speech in Dutch tweets and to use that model on a dataset of Dutch tweets. Another goal is to evaluate the performance of the BERT-CNN model in comparison to other hate speech detection techniques. To accomplish this, the following research questions (RQ) are the foundation of this research:

- RQ 1: What are the primary targets of hate speech on X in the Netherlands?

- RQ 2: How well-suited is a BERT model for classifying hate speech and how do its metrics compare to the original IMSyPP model?

RQ 1 is answered by training a BERT-CNN model on a labelled dataset of Dutch hate speech [12]. This model is then used on a large, unlabeled dataset to find hate speech. This paper investigates hate speech during the 2018-2020 period, and both these datasets originate from that time frame. This includes the start of the COVID-19 pandemic, which is especially interesting when analysing hate speech since earlier research [4] proves that the amount of hate speech increased during this period. After training the model, the main targets are extracted by another trained model that predicts the target from a tweet marked as hate speech. RQ 2 will be answered by analysing the standard [13] performance characteristics of machine learning models such as accuracy, precision, recall, and F1 score. These metrics are then compared to the same metrics achieved by the IMSyPP model. They are explained in more detail below.

1.3 Performance metrics

To answer RQ 2, the following performance metrics are used: precision, recall, f1-score and accuracy. This is an industry-standard way of analysing the performance of a model [13]. These metrics are calculated using the amount of True Positive (TP, samples that the model predicted correctly as hate speech), True Negative (TN, samples that the model predicted correctly as NOT hate speech), False Positive (FP, samples that the model predicted as hate speech but were normal speech) and False Negative (FN, samples that the model predicted as normal speech but were actually hateful). Using the metrics as explained in [13], we can analyse the model's performance. Precision reflects how well the model is able to classify the correct target class and is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall reflects how well the model is able to find all samples of the target class and is calculated as follows:

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

The f1-score provides a metric that balances both recall and precision. It is defined by the following formula:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

A model's accuracy defines how well a model works in general. It is not suited for providing an understanding in how the model performs on edge cases which is why the aforementioned scores are equally important. It is defined as follows:

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

A confusion matrix shows the number for each of the four categories (TP, FP, TN, FN). An example confusion matrix, extracted from [13], can be found in Table 1. In this example, the accuracy is:

$$\text{Accuracy} = \frac{261 + 193}{261 + 193 + 107 + 39} = \frac{454}{600} \approx 0.76\%$$

Predicted\True class	Pos.	Neg.
Pos.	TP = 261	FP = 107
Neg.	FN = 39	TN = 193

Table 1: Example Confusion Matrix

1.4 Paper outline

The structure of the paper is as follows: In Section 2, background knowledge on BERT and CNNs and the state of the art in detecting hate speech are discussed. After that, in Section 3 the specific techniques used in this research are discussed. In Section 4 both research questions are answered. Section 6 discusses the ethical aspects of using X data and training a hate speech classifier. Lastly, Section 6.1 states for what purposes AI was used during this research.

2 RELATED WORK

This section will introduce and indicate the start of the art for BERT and CNN models. It also presents related work on hate speech detection, specifically using BERT-CNN models.

2.1 BERT and CNNs

NLP, or Natural Language Processing has evolved much over the years. Understanding and processing text with computers has always been an important task, and its research dates back to the 1950's [14]. What started with the goal to represent text mathematically is now a large research field with applications in everyone's daily life, such as question-answering systems and automated text summarisation [15].

A leap in NLP models happened with the introduction of BERT, or Bidirectional Encoder Representation from Transformers. Introduced by Devlin et al. in 2019 [16], it has become the industry standard [8] in many automated word processing tasks. BERT uses a transformer architecture[1] to capture bidirectional context in text. The model is pre-trained on large amounts of text and can be fine-tuned for more specific uses.

Since the introduction of Devlin et al.'s BERT, many variants have been developed, each with its specific use. One such BERT variant is called DistilBERT [17]. It covers one of BERT's major weaknesses: its size. DistilBERT uses knowledge distillation to "reduce BERT's size by 40% while retaining 97% of its language understanding capabilities and being 60% faster." It is also a multilingual model, so it is also capable of tokenising Dutch.

Another BERT variant, called BERTje was developed for the Dutch language [18]. This research shows that BERTje outperforms multilingual BERT models on Dutch NLP tasks.

Convolutional Neural Networks (CNNs) are a type of neural network and are primarily used in the field of pattern recognition within images [19]. CNNs can be used to encode image-specific features more easily due to their convolutional layers. However, CNNs are not only suitable for image classification. They can also be used for greater understanding of context in texts [20].

[1] An AI architecture that can understand and process text by focusing on important words and their connections, even if they are far apart in the sentence.

A combination of the two discussed models results in a BERT-CNN model. Authors in [9] developed such a BERT-based CNN, and proved that the combination is better than using BERT on its own: the combination had a macro f1-score of 0.851 compared to BERT on its own with 0.841.

2.2 Hate speech detection

This section outlines the various attempts and techniques used in analysing hate speech on social media.

In Silva, Mondal, Correa, *et al.* [21] researchers claim to provide the first of a kind systematic large scale measurement study of the main targets of hate speech in online social media. Using a simple grammar structure they identified the targets of hate speech in X and Whisper[2] posts. Whisper is social media platform where users can anonymously post short text messages. The main difference between X and Whisper is the anonimicity: whereas X posts are linked to a user, Whisper posts are anonymous. A limitation of this work is that this approach is very simple and does not take into account the full nuance of language. It only classifies sentences that are structured like "I really hate <target>". Advanced AI models like BERT can achieve a much greater understanding of language and thus find more results. This research also uses the HateBase[3] as a dictionary of hate words. For the English language, it contains 1,565 words (at the moment of writing this paper).

Kupi, Bodnar, Schmidt, *et al.* [11] evaluated a dictionary-enhanced CNN model for detecting hate speech, also using the previously mentioned HateBase. The dictionary-enhanced model increased the CNN model's predictive power by seven percentage points. This paper only focuses on the English language, not on Dutch.

A European project called IMSyPP[4] (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech) has been monitoring hate speech in Europe. They have developed a hate speech detection and target prediction model for several languages, including Dutch. However, these models have not been used on a dataset of Dutch tweets.

In 2020 a BERT-CNN model was developed for identifying offensive speech in social media (specifically X) [9]. Using a multilingual model, they analysed Arabic, Greek and Turkish tweets. In their paper, the researchers compared their model to several others, such as SVM (Support Vector Machine) and BERT on its own. They concluded that their BERT-CNN model was the best of all evaluated models, having an average macro F1 score of 0.851 for all languages. The model was trained for 10 epochs with learning rate of 2e-5.

In an effort to distinguish hateful tweets from regular ones in Dutch, Caselli et al. [10] developed the "Dutch Abusive Language Corpus (DALC)". This corpus consists of manually annotated Dutch Tweets gathered between 2015 and 2019. The authors also evaluated several models (MFC, SVM, BERTje, Dictionary) on the dataset, achieving the highest macro-F1 score of 0.748 with BERTje. The focus of this model however is on whether the tweet is hate speech or not, and whether it is implicit or explicit. There are no labels for the target of hateful tweets.

[2]https://whisper.sh/
[3]https://hatebase.org
[4]http://imsypp.ijs.si/

Besides the creation of the DALC dataset, research on online hate speech in the Netherlands has been sparse, focusing mainly on the effects of hate speech ([22], [23]) and not the targets of hate speech. To the best of our knowledge, no analysis and classification of hate speech targets on X has been conducted in the Netherlands.

3 METHODOLOGY

This section will describe the methodology of the different phases of the research. Firstly, the data collection will be described. Secondly, the model training process is presented, after which the way of classifying targets.

3.1 Data collection

To analyse hate speech on social media, 2 types of datasets are needed. Firstly, a dataset of labelled Dutch hate speech (tweets). 2 of those datasets are the following:

- Labeled hate speech in Dutch [5]
- The HateBase [6]

The first dataset is manually labelled, containing 25,719 training messages and 2,858 evaluation messages, gathered not only from X but also from other platforms, including but not limited to Dumpert and YouTube. The data was gathered between January 2018 and October 2020. It is created for the IMSyPP[7] project [12]. This EU project aims to apply machine learning and a data-driven approach to "hate speech regulation, prevention and awareness-raising." The dataset has been labelled by 15 annotators. The posts are marked with the following labels:

(1) Appropriate
(2) Inappropriate - contains terms that are obscene, vulgar; but the text is not directed at any person specifically
(3) Offensive - including offensive generalisation, contempt, dehumanisation, indirect offensive remarks
(4) Violent author threatens, indulges, desires or calls for physical violence against a target; it also includes calling for, denying or glorifying war crimes and crimes against humanity

The label distribution can be found in Figure 1, as well as in the dataset documentation [12]. When a post is marked as either offensive or violent, a target is assigned. The following target labels are used (as defined by the dataset creators), and their distribution can be found in Figure 2. This distribution was not provided and was analysed using Python.

(1) Racism - intolerance based on nationality, ethnicity, language, towards foreigners; and based on race, skin colour
(2) Migrants - intolerance of refugees or migrants, offensive generalisation, call for their exclusion, restriction of rights, non-acceptance, denial of assistance…
(3) Islamophobia - intolerance towards Muslims
(4) Antisemitism - intolerance of Jews; also includes conspiracy theories, Holocaust denial or glorification, offensive stereotypes…
(5) Religion - other than above

[5]https://github.com/textgain/IMSyPP-DATA
[6]https://hatebase.org
[7]https://imsypp.ijs.si/

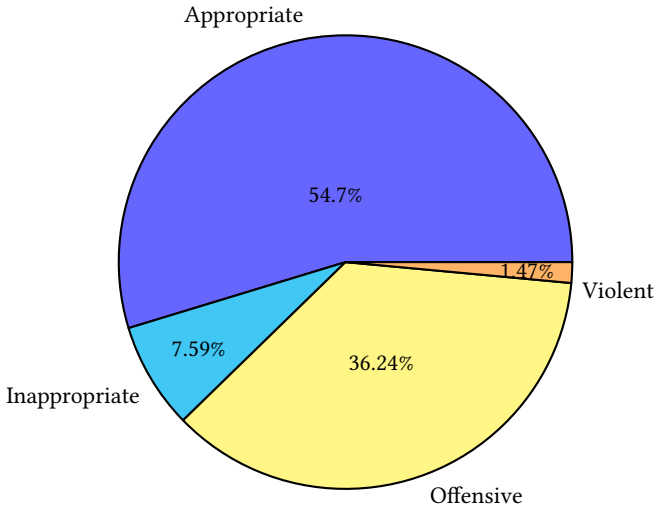

Figure 1: Label distribution of dataset

(6) Homophobia - intolerance based on sexual orientation and/or identity, calls for restrictions on the rights of LGBTQ persons
(7) Sexism - offensive gender-based generalisation, misogynistic insults, unjustified gender discrimination
(8) Ideology - intolerance based on political affiliation, political belief, ideology… e.g. "communists", "leftists", "home defenders", "socialists", "activists for…"
(9) Media - journalists and media, also includes allegations of unprofessional reporting, false news, bias
(10) Politics - intolerance towards individual politicians, authorities, system, political parties
(11) Individual - intolerance toward any other individual due to individual characteristics; like commentator, neighbour, acquaintance
(12) Other - intolerance towards members of other groups due to belonging to this group. A few examples include elderly people or members of subcultures like goth or punk. The category 'Religion' includes all tweets that are targeted at religions other than Islam and Judaism. Examples include but are not limited to Christianity, Buddhism, or Flying Spaghetti Monster[8].

The second dataset is simply an international database of hate speech vocabulary which also includes Dutch hate speech. This database was used in [4] and [21]. Pandey, Garcia-Robledo, and Zangiabady [4] used a scraper script to extract all words marked as hate speech. The HateBase also links a 'hate speech word' to a category. Using this technique, it is possible to extract the hate speech targets from tweets marked as hate speech containing the words from the HateBase.

The second type of dataset needed for this research is a large dataset of Dutch tweets on which to use the trained model to classify hate speech. In the past, it was possible to scrape X for data. However, in recent years X has changed its policies a lot[9]. As a result, scraping X for free is no longer possible. A paid subscription allows for gathering a limited amount of recent tweets per day. Therefore this research will use a pre-existing Dutch tweets dataset[10] (271,342 tweets). This dataset contains Tweets from the Netherlands, gathered between January 2020 and December 2020.

Both the training dataset and the dataset used for analysis are from before and during the start of the Covid era (2018-2020), to ensure that the model is trained optimally for usage on the test set.

[8]https://www.spaghettimonster.org/
[9]https://x.com/XDevelopers/status/1621026986784337922?lang=en

	Accuracy	Precision	Recall	F1
Hate speech	81.2	83.4	81.2	81.1
Target	27.4	22.0	27.3	24.0

Table 2: Performance metrics for IMSyPP Models: hate speech detection and target identification

3.2 Model training

The IMSyPP dataset has been used to train two classifier models: a model for classifying a tweet as hate speech[11] and a model for classifying the target of the hate speech[12]. The model cards included in the model do not reflect on the performance of the model, so an analysis was necessary to evaluate whether this model could be used. Table 2 shows the performance metrics of these models, after running the model on the evaluation set provided in the dataset.

Since the target classifier has a significantly lower performance than the hate speech classifier, this research focused on developing a new BERT-CNN model for the former. Training another model for the latter was out of scope for this project.

The new target classifier has been trained on the same dataset. The non-hate speech tweets from the dataset (labelled 0 or 1) have been removed, so that the model is not concerned with non-hate speech. The model has been trained on a Jupyter Notebook server from the University of Twente, since training the model and running the model on a large dataset requires serious resources. For training and evaluating the PyTorch library[13] was used.

3.3 Classifying targets

Initially, there were two ways classify the targets. The first one was using simple regular expressions [4], to allow the targets of hate speech to be extracted from the tweets marked as hateful by the model. The words extracted from HateBase can then be used to filter through the tweets to gather the main categories. This can be done using HateBase's coupling that matches a hateful word to a category of hate speech. By analysing the words with the highest frequencies in the tweets (apart from articles and adjectives and so forth), the main targets of hate speech can be extracted.

However, after careful examination of the Dutch HateBase, we have drawn the conclusion that this method is not suitable. The Dutch HateBase only contains 126 hateful words, as opposed to the English HateBase containing 1,565. Moreover, the HateBase has officially been retired, meaning its API is deprecated and it is no longer maintained.

[10]https://www.kaggle.com/datasets/skylord/dutch-tweets
[11]https://huggingface.co/IMSyPP/hate_speech_nl
[12]https://huggingface.co/IMSyPP/hate_speech_targets_nl
[13]https://pytorch.org/

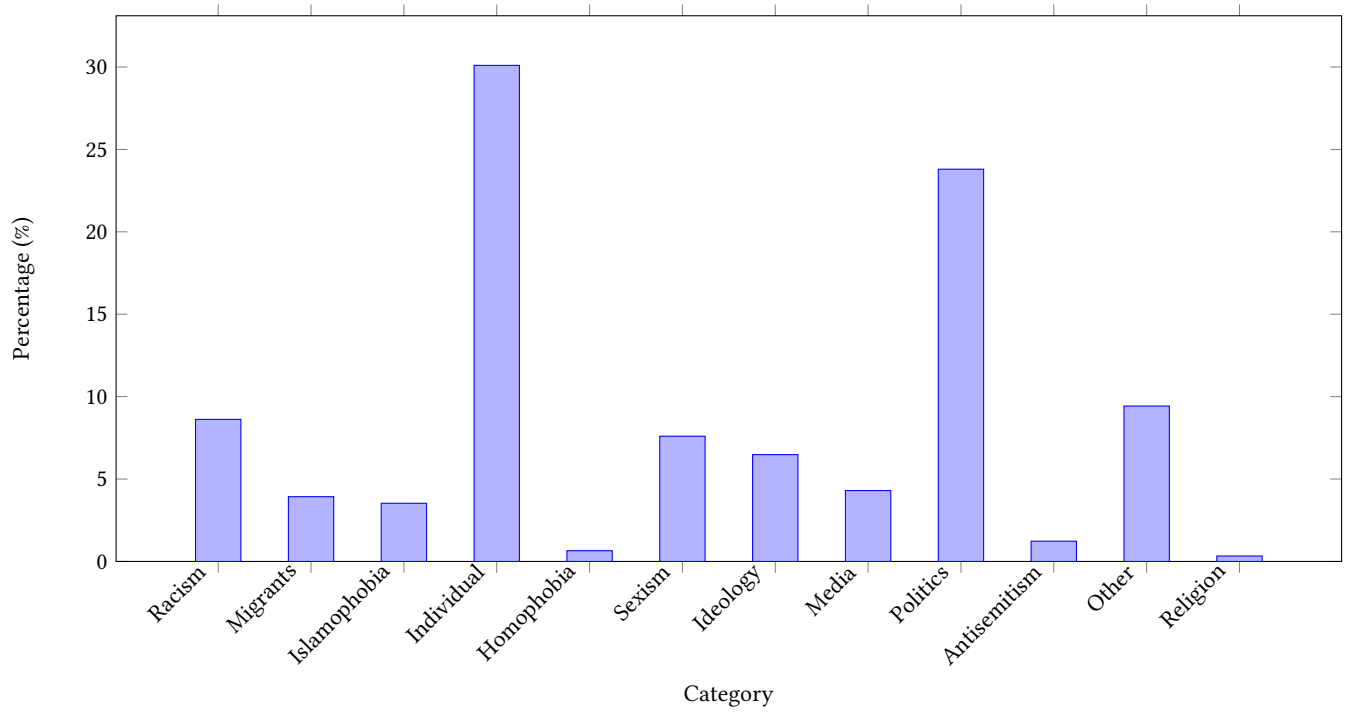

Figure 2: Percentage distribution of target categories

The second way is to use the 'target' label in the labelled IMSyPP dataset to train a new model, since as mentioned above, the model created for the IMSyPP project has a low performance.

The model was trained on top of the BERTje model and tokeniser. These embed the sentences into size 768 embeddings. BERTje's output was then transferred to a CNN classifier based on [24]. We set the maximum amount of words in a tweet to 100 words, truncating longer texts and padding shorter texts. On the word embedding from BERTje [100 x 768], convolutions of different sizes are used: [2 x 768], [3 x 768] and [4 x 768]. This way, the model looks at combinations of 2, 3 or 4 words. A Rectified Linear Unit (ReLU) is then applied as the activation function. 1-max pooling(A process that reduces the size of data by selecting the largest value in each region of the input) is applied to down-sample the input representation. To prevent overfitting, a dropout layer is added. Lastly, a softmax function is applied to distribute the probability between classes. Using this structure, the model was trained during 10 epochs with a learning rate of 0.001. Smaller epochs were tested (3, 5) but this led to underfitting. Larger epochs (15, 20) led to overfitting. Smaller and larger learning rates were also tested for each epoch amount (0.01, 0.1, 0.0001), The Binary Cross-Entropy loss function [25] was used together with the standard Adam optimiser [26].

The model was trained on all posts marked as either offensive or violent. It only took about 1 minute to train the CNN classifier, since the (much larger) BERTje was pretrained and did not have to be altered.

To answer the RQs, this model is applied to the large unlabelled dataset of tweets. After that, it will be compared to the IMSyPP model based on the performance metrics discussed in 1.3.

4 RESULTS

In this section, the results of this research are discussed. Firstly, in Section 4.1 the results of running the model on the dataset are shown. Then, in Section 4.2 the model is evaluated. Lastly, in Section 4.3 the model and training dataset are discussed, and recommendations regarding these are provided.

4.1 Targets of Hate Speech in the Netherlands

To recap, RQ 1 is "What are the primary targets of hate speech on X in the Netherlands?". After running the IMSyPP model on the entire dataset, 31155 (11.5%) tweets were classified as hate speech. On this set of 30K tweets the second (self-trained) model was run, for which the results are shown in Table 3. As can be seen in the table, more than half of the posts are marked as 'homophobia'. 25% of posts are marked as 'other'. The categories 'religion' and 'Islamophobia' are also very present with 9.74% and 7.81% respectively. All other categories are predicted very little, all being predicted for less than 2% of all the posts. An analysis of hate speech over time, as presented in [4] was not performed, since the tweets were gathered from a much smaller time frame.

4.2 Model performance

To answer RQ 2, the performance metrics discussed in Section 1.3 are needed. The performance of the model can be analysed and compared to the IMSyPP model. The scores for the newly trained model can be found in Table 4. Compared to the model trained by [4] and [11], this model has lower overall accuracy. On the other hand, the other models were not able to classify the target, only

Category	Percentage (%)
Homophobia	51.30
Other	25.00
Religion	9.74
Islamophobia	7.81
Individual	1.96
Media	1.64
Sexism	1.26
Antisemitism	0.91
Migrants	0.36
Racism	0.05
Ideology	0.02

Table 3: Distribution of Categories and Their Percentages

Category	Precision	Recall	F1-score
Antisemitism	0.54	0.64	0.58
Homophobia	0.00	0.00	0.00
Ideology	0.48	0.61	0.54
Individual	0.61	0.61	0.61
Islamophobia	0.62	0.44	0.52
Media	0.47	0.51	0.49
Migrants	0.40	0.33	0.36
Other	0.31	0.33	0.32
Politics	0.68	0.68	0.68
Racism	0.55	0.58	0.56
Religion	0.00	0.00	0.00
Sexism	0.51	0.46	0.49
Accuracy			0.56
Macro avg	0.43	0.43	0.43
Weighted avg	0.56	0.56	0.56

Table 4: Model performance

whether the tweet contained hate speech or not. The advantage of this approach is that with a neural network, it is easier to validate the correctness of the approach as opposed to the method used in [4].

The newly trained model (containing a CNN) does perform significantly better (65% accuracy as opposed to 27%) than the original IMSyPP target classifier that only uses BERT.

There are some interesting values in Table 4. Both Homophobia and Religion have '0.00' in all columns. This is the result of the underrepresentation of those labels in the dataset. This is further discussed in the next section.

4.3 Discussion and Recommendations

Since this research provides a novel way of finding targets in hate speech tweets, it comes with its limitations.

Firstly, the reliance on the dataset from the IMSyPP project means that the model's performance depends on the quality and contents of this dataset. Since binary classifiers (like models that simply predict whether a tweet is hate speech or not) need fewer training samples than multilabel classifiers (like the one trained in this paper), the dataset is better suited for binary classification than multilabel classification. To ensure a higher performance, simply more data is needed. This means that the model will have a better understanding of the different hate speech categories and will more accurately predict each category.

This leads directly to a similar limitation: it becomes clear from the label distribution in Figure 2 that the dataset is imbalanced. Whereas the dataset contains 30% individual labels and 23% politics labels, the rest of the labels are all less than 10%. Especially Homophobia and religion are under-represented. Comparing this to the results in Table 3 shows that the trained model could be improved when these categories were more present in the dataset. Currently, it seems that because the model lacks an understanding of Homophobia and Religion, it classifies a significant amount of tweets as said categories. Compared to the results of other studies, like [4], these results seem quite extreme which could be caused by the dataset imbalance.

In general, because of the very different sample sizes in the labelled dataset, the model was not able to generalise every category equally well. This implicates that further research is required to confirm or disprove the results shown in Table 3.

5 CONCLUSION AND FUTURE WORK

In this paper, a BERT-CNN classifier model was developed and used for classifying hate speech targets from tweets.

It becomes clear that adding a CNN classifier on top of the pretrained BERTje model improves its classification capabilities by comparing Table 2 and Table 4. Even though the model does not achieve the highest performance, it is a promising step in this approach to target classification. With a larger and less imbalanced dataset, it seems that a target prediction model could be a viable approach to extracting targets from hate speech tweets.

The results in Table 3 imply that more than half of Dutch hate speech tweets are of the category 'homophobia'. If this were truly be the case, this would be a shocking result and would definitely require more research.

Table 4 explains this rather strange distribution found in Table 3. The model is not able to identify homophobia and religion well enough due to its underrepresentation in the dataset. This causes the model to mark any tweets that it is not sure about as homophobia, religion, or other.

Future work might investigate hate speech in the Netherlands on a larger scale. The data that was analysed in this paper consisted of 270,000 tweets, and a larger dataset is needed to make more accurate claims about Dutch hate speech targets. Moreover, adding a CNN classifier on top of the prepossessing model, which was initially the goal of this research, might result in significantly higher performance.

6 ETHICS

Since this research worked with tweets (which are written by humans), it is important to address the following ethical concerns:

(1) We work with datasets containing personal data
(2) The consequences of training a model

The datasets used in this research are publicly available. There is ongoing development in data science ethics related to (re)publishing publicly available data. For that reason, the GitHub repository where this project is stored does not contain the datasets.

The second ethical issue, regarding the consequences of AI, is very pressing these days [27]. We are aware that the datasets could contain demographic biases, resulting in a biased model. We did not analyse the dataset for demographic biases since that was out of scope for this project. Because of this, we recommend that the model should only be used for further research, not in practical applications yet.

6.1 AI use

Conforming to the University's regulations on the usage of ChatGPT, this is a brief overview of the usage of ChatGPT or other generative AI tools in this project.

ChatGPT was used for two tasks: rephrasing some sentences for a more academic style and asking questions about error messages when running code or building the project.

REFERENCES

[1] K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and Psychological Effects of Hateful Speech in Online College Communities," *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference*, vol. 2019, pp. 255–264, Jun. 2019. DOI: 10.1145/3292522.3326032. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7500692/ (visited on 05/05/2024).

[2] C. Calvert, "Hate speech and its harms: A communication theory perspective," en, *Journal of Communication*, vol. 47, no. 1, pp. 4–19, 1997, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1997.tb02690.x, ISSN: 1460-2466. DOI: 10.1111/j.1460-2466.1997.tb02690.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1997.tb02690.x (visited on 05/04/2024).

[3] M. Mohiyaddeen and D. Siddiqui, "Automatic Hate Speech Detection: A Literature Review," *International Journal of Engineering and Management Research*, vol. 11, pp. 116–121, Apr. 2021. DOI: 10.31033/ijemr.11.2.17.

[4] S. S. Pandey, A. Garcia-Robledo, and M. Zangiabady, "Decoding Online Hate in the United States: A BERT-CNN Analysis of 36 Million Tweets from 2020 to 2022," en, in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA: IEEE, 2023, pp. 329–334, ISBN: 9798350385359. DOI: 10.1109/ICSC59802.2024.00059. [Online]. Available: https://ieeexplore.ieee.org/document/10475658/ (visited on 04/25/2024).

[5] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, *Hate Speech Classification Using SVM and Naive BAYES*, arXiv:2204.07057 [cs], Mar. 2022. DOI: 10.9790/0050-09012734. [Online]. Available: http://arxiv.org/abs/2204.07057 (visited on 06/25/2024).

[6] S. M. A. Shah and S. Singh, "Hate Speech and Offensive Language Detection in Twitter Data Using Machine Learning Classifiers," en, in *Innovations in Computer Science and Engineering*, H. S. Saini, R. Sayal, A. Govardhan, and R. Buyya, Eds., Singapore: Springer Nature, 2023, pp. 221–237, ISBN: 978-981-19745-5-7. DOI: 10.1007/978-981-19-7455-7_17.

[7] K.-L. Chiu, A. Collins, and R. Alexander, *Detecting Hate Speech with GPT-3*, arXiv:2103.12407 [cs], Mar. 2022. DOI: 10.48550/arXiv.2103.12407. [Online]. Available: http://arxiv.org/abs/2103.12407 (visited on 06/25/2024).

[8] M. V. Koroteev, *BERT: A Review of Applications in Natural Language Processing and Understanding*, arXiv:2103.11943 [cs], Mar. 2021. DOI: 10.48550/arXiv.2103.11943. [Online]. Available: http://arxiv.org/abs/2103.11943 (visited on 04/25/2024).

[9] A. Safaya, M. Abdullatif, and D. Yuret, *KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media*, arXiv:2007.13184 [cs], Jul. 2020. [Online]. Available: http://arxiv.org/abs/2007.13184 (visited on 04/25/2024).

[10] T. Caselli, A. Schelhaas, M. Weultjes, *et al.*, "DALC: The Dutch Abusive Language Corpus," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 54–66. DOI: 10.18653/v1/2021.woah-1.6. [Online]. Available: https://aclanthology.org/2021.woah-1.6 (visited on 06/20/2024).

[11] M. Kupi, M. Bodnar, N. Schmidt, and C. E. Posada, *dictNN: A Dictionary-Enhanced CNN Approach for Classifying Hate Speech on Twitter*, arXiv:2103.08780 [cs], Mar. 2021. DOI: 10.48550/arXiv.2103.08780. [Online]. Available: http://arxiv.org/abs/2103.08780 (visited on 04/25/2024).

[12] P. K. Novak, I. Mozetič, G. de Pauw, and M. Cinelli, *Multilingual Hate Speech Database*, English, Feb. 2021. [Online]. Available: http://imsypp.ijs.si/wp-content/uploads/2021/12/IMSyPP_D2.2_multilingual-dataset.pdf (visited on 06/20/2024).

[13] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," en, *Scientific Reports*, vol. 14, no. 1, p. 6086, Mar. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-56706-x. [Online]. Available: https://www.nature.com/articles/s41598-024-56706-x (visited on 06/23/2024).

[14] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, "Natural Language Processing: History, Evolution, Application, and Future Work," en, in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, A. Abraham, O. Castillo, and D. Virmani, Eds., Singapore: Springer, 2021, pp. 365–375, ISBN: 9789811597121. DOI: 10.1007/978-981-15-9712-1_31.

[15] S. Jusoh, "A STUDY ON NLP APPLICATIONS AND AMBIGUITY PROBLEMS," en, . *Vol.*, no. 6, 2005.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 [cs], May 2019. DOI: 10.48550/arXiv.1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805 (visited on 06/20/2024).

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*, arXiv:1910.01108 [cs], Feb. 2020. DOI: 10.48550/arXiv.1910.01108. [Online]. Available: http://arxiv.org/abs/1910.01108 (visited on 06/17/2024).

[18] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, *BERTje: A Dutch BERT Model*, en, arXiv:1912.09582 [cs], Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.09582 (visited on 05/21/2024).

[19] K. O'Shea and R. Nash, *An Introduction to Convolutional Neural Networks*, en, arXiv:1511.08458 [cs], Dec. 2015. [Online]. Available: http://arxiv.org/abs/1511.08458 (visited on 06/20/2024).

[20] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, Conference Name: IEEE Transactions on Neural Networks and Learning Systems, ISSN: 2162-2388. DOI: 10.1109/TNNLS.2021.3084827. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9451544?casa_token=mYjBZ2vKKgEAAAAA:XaQ9kyUJZQ6CpTrMZiQfaVMX-iovfhRCFidbqTCHpj6ZsU_AnmSiL1Z1Se_u-bD82YU9bL8OIA (visited on 06/11/2024).

[21] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the Targets of Hate in Online Social Media," en, *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, pp. 687–690, 2016, Number: 1, ISSN: 2334-0770. DOI: 10.1609/icwsm.v10i1.14811. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14811 (visited on 05/04/2024).

[22] A. Braeckman, "Hate speech on the internet. Case study of Belgium and the Netherlands," en, Paris: Ghent University, Jul. 2007.

[23] L. Jacobs and J. van Spanje, "A Time-Series Analysis of Contextual-Level Effects on Hate Crime in The Netherlands," *Social Forces*, vol. 100, no. 1, pp. 169–193, Sep. 2021, ISSN: 0037-7732. DOI: 10.1093/sf/soaa102. [Online]. Available: https://doi.org/10.1093/sf/soaa102 (visited on 05/07/2024).

[24] Y. Kim, *Convolutional Neural Networks for Sentence Classification*, en, arXiv:1408.5882 [cs], Sep. 2014. [Online]. Available: http://arxiv.org/abs/1408.5882 (visited on 06/23/2024).

[25] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2962617. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8943952 (visited on 06/28/2024).

[26] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980 [cs], Jan. 2017. DOI: 10.48550/arXiv.1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980 (visited on 06/28/2024).

[27] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," en, *Sci*, vol. 6, no. 1, p. 3, Mar. 2024, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2413-4155. DOI: 10.3390/sci6010003. [Online]. Available: https://www.mdpi.com/2413-4155/6/1/3 (visited on 06/17/2024).