university of groningen / faculty of science and engineering

Using Machine Translated Lexicons for Hate Speech Classification on Dutch Covid-19 Twitter Data

Bachelor's Project Thesis

Amber Chen, s3333302, a.chen.1@student.rug.nl,
Supervisors: J. Doornkamp and Dr J. Spenader

Abstract: Increased user population of social platforms, and increased time spent online per user during the COVID-19 pandemic, have resulted in the occurrence of more online hate speech. Efforts to create automated hate speech classification systems have mainly made use of English resources. For low-resource languages like Dutch, machine translation of existing lexicons is a possible workaround. A COVID-19 related Twitter data set was filtered using these lexicons. The data was annotated and classified using an SVM, which was trained on emotional models and anger intensity. TF-IDF showed that (translated) lexical entries were among the 20 highest scored n-grams. Inter-rater agreement was calculated after annotation and was found to be low. Of the calculated features, only anger intensity seemed useful in classification: the other features were scored several magnitudes smaller in terms of information gain. The low agreement and low information gain returned by the majority of the features showed in the results of the classification, where chance level accuracy scores were found for all systems.

Keywords: COVID-19; Hate Speech; Twitter; Machine Translation; Emotion Intensity

1 Introduction

Online hate speech has been a topic for research since the early days of the internet (Weintraub-Reiter, 1998), and is therefore not limited to times of the COVID-19 pandemic. However, the growing number of users of the internet, combined with increased time spent online per user (Van der Veer, Boekee, & Hoekstra, 2020, 2021), do go hand in hand with the emergence of more online hate speech. Efforts to combat online hate, especially against minorities, have proven ineffective, as observed cases have been increasing instead of decreasing (Shields, 2020; United Nations Human Rights Office of the High Commissioner, 2021).

Twitter is often used as a source in online hate speech research and is especially interesting during a pandemic since it expresses traits from both social networks and news media (Kwak, Lee, Park, & Moon, 2010). Its 2.9 million Dutch user population partly consists of government officials and official organizations (Twitter, 2021), who use the platform to broadcast important (news) messages. COVID-19 press conferences by the Dutch government are usually prefaced with speculations of measures that are to be announced, originating from official organizations such as newspapers. In turn, this can incite reactions from the general public.

The mix of news and reactions from the general public contribute to the emergence of discussions about the consequences of the pandemic and (rumored) measures to prevent the spread of the virus. Twitter can serve as a soundboard for the people to display their opinions and thoughts (Younus et al., 2011), and in more extreme cases as a place for broadcasting hate speech.

To maintain a safe space online, ways to systemically detect and remove instances of hate speech need to be found. A machine learning algorithm trained on hate speech tweets could help identify these instances, such that platforms don't have to solely depend on dedicated employees or users that

flag offensive/abusive content. Machine learning is therefore proposed as the automated solution to the online hate speech problem.

Automated hate speech classification systems do come with challenges. Despite the many attempts, there seems to be no unified solution as of yet. Waseem, Davidson, Warmsley, and Weber (2017) discovered the cause to be two-fold: lack of universal definitions, and lack of universal subtasks.

Many countries, including the Netherlands (Wetboek van Strafrecht, 2020), have documented a legal definition of hate speech, but these definitions often stay ambiguous and furthermore, they are different depending on the country. There currently is no international legal definition of hate speech (United Nations, 2019), and the definitions used in existing research are inconsistent as identified by Waseem et al. (2017).

Besides definitions, the subtasks of a hate speech classification task often differ across studies. Even paired with an extensive definition, hate speech is a general term that leaves room for interpretation. Therefore, subsets of hate speech are often identified for classification. While these subsets (e.g. racist/sexist remarks) are often consistent across studies, the classes the subsets are ascribed to are often not. To illustrate, discriminatory messages such as sexist and racist remarks are identified by Van Hee et al. (2015) as 'insults', while they would be classified as 'hate speech' or 'derogatory language' by Nobata, Tetreault, Thomas, Mehdad, and Chang (2016). While both identified the same subset of messages, they are classified as less abusive in one study (Van Hee et al., 2015) compared to the other (Nobata et al., 2016).

As such, the criteria (i.e. the questions a human annotator needs to ask in order to assign labels to the data) of the hate speech classification task become inconsistent.

While hate speech classification tasks are no easy feat of their own, an added challenge occurs when using low-resource languages such as Dutch. Around 60% of all web pages are written in English (W3Tech, 2021), making most online hate speech resources (such as annotated data sets and lexicons) only available in English.

Workarounds for low-resource languages are possible by crowdsourcing. It should be noted that the collected offensive terms will be dependent on the platform and its demographic, and the time of collection. A term's offensive meaning in one context is not guaranteed in another, and terms may be missing since the users of the platform used for crowdsourcing do not make use of them. Crowdsourcing is therefore likely to create resources that are 'neither exhaustive nor conclusive' (Sigurbergsson & Derczynski, 2020).

Another possibility is translating existing resources in languages that are more commonly used for hate speech research. Since translation by humans can be costly in terms of time and financial resources and requires an expert who is trained and skilled in translation, this task lends itself well to machine translation, where texts are translated by software. Aluru, Mathew, Saha, and Mukherjee (2021) found machine translation to work well in hate speech classification tasks for several different languages. However, machine translation of *large* data sets takes us back to the problem of translation by humans: unless the paid version of machine translation software is used, daily input maximums are often encountered. This means that large data sets will have to be broken down into smaller sets, which will have to be passed through the system over a long period of time.

In this study, the resources for the translation of the full data set were simply not available. It is therefore desirable to use machine translation as little as possible, while still utilizing the extensive collection of English hate speech resources out there. Instead of using English lexicons on machine translated tweets, Dutch lexicons and supplementary machine translated lexicons collected by Davidson, Warmsley, Macy, and Weber (2017) will be used to filter Dutch tweets.

Furthermore, to move towards a universal definition of hate speech, definitions from previous works will be compiled, and the typology of abusive language created by Waseem et al. (2017) will be used in this hate speech classification task on COVID-19 related tweets.

In the following sections, related work will first be discussed, after which the data set, annotation process, constructed features, and classifier will be described and evaluated. Conclusions and discussion of the study as a whole are included as well.

2 Related Work

Marinov, Spenader, and Caselli (2020) used topic modeling and emotional analysis on COVID-19 related tweets. The data used by Marinov et al. (2020) consist of tweets of the first three months of the data set used in our study. Anger was found among the top five emotions across user groups and months. Since anger is detected in the data set, and hate speech is primarily based on the emotion anger (Alorainy, Burnap, Liu, Javed, & Williams, 2018), it is suggested that hate speech occurs in the tweets used by Marinov et al. (2020). Hate speech is therefore also suggested to occur in the tweets used in this study.

Davidson et al. (2017) conducted a study on Twitter hate speech, distinguishing between three different classes: hate speech, offensive language, and neither. By using this three-class system, they identified the correlation between certain offensive n-grams (series of consecutive words from a text, of length n) and the occurrence of hate speech. These n-grams were collected in a lexicon, that was used for filtering the data in this study. Furthermore, the three-class system and part of the definition of hate speech by Davidson et al. were used in our study.

Martins, Gomes, Almeida, Novais, and Henriques (2018) reused the data from Davidson et al. (2017), and investigated the use of emotional analysis to train a model. More specifically, a word-emotion association lexicon was used to calculate emotion scores for each word in a tweet. Furthermore, the anger intensity score was calculated using an emotion intensity lexicon. A significant increase in precision and recall could be observed compared to the study by Davidson et al., suggesting that the use of emotions and intensity scoring can help distinguish between hate speech and offensive language. Based on these findings, emotion and anger intensity scores will be used as features in our study.

The study by Vidgen et al. (2020) looked into online discourse about East-Asia during the COVID-19 pandemic. Their error analysis showed that the majority of the classification errors (i.e. the predicted class differed from the true class of the data) were caused by the machine learning algorithm. More precisely, the larger part of those machine learning errors consisted of misclassification of tweets in closely related categories. The rest of the classification errors were caused by annotator error, in which case the machine learning algorithm actually correctly classified the data. The labels given by the annotators however were evaluated to be incorrect. The error analysis by Vidgen et al. gives an overview of the difficulties that can be encountered during the annotation process and classification by the system. This in turn helped prepare for the work done in our study.

Aluru et al. (2021) compared the performance of different deep learning methods and machine translation, using data sets in multiple languages (Arabic, English, German, Indonesian, Italian, Polish, Portuguese, Spanish, and French). For the Polish language, which like Dutch has an occurrence of about 0.5% on the internet (W3Tech, 2021), machine translating texts into English before classification often resulted in the best or second-best performance compared to word embedding methods (Aluru et al., 2021). This suggests that machine translating texts from low-resource languages into English is a viable method for hate speech classification.

3 Data

These sections will elaborate on the creation of the data set, how it was preprocessed, and show the results of preliminary data analysis.

3.1 Constructing the Data Set

The data set `40twene_nl`[1] constructed by Caselli and Basile (2020) is a set of Dutch Tweets, written between February 2020 and December 2020. The data set was created by filtering an ongoing collection of Dutch tweets (Sang, 2011; Bouma, 2015) on a set of hashtags related to COVID-19, such as `#coronavirusNederland`, `#thuisblijven` (staying home), and `#anderhalvemeter` (1.5 meters). It consists of around 10 million tweets.

The data was selected based on the occurrence of key words originating from two different lexicons. The first lexicon consisted of Dutch terms found on Hatebase.org, that were manually filtered for relevance, leaving 113 terms (e.g. 'tokkie', 'slet', 'mongool') out of the original 126. Terms that are more often used in an unoffensive context rather than an

[1] `https://osf.io/pfnur/`

offensive context were excluded, such as 'banaan' (banana) and 'poot' (paw).

The second lexicon consisted of a machine translated Dutch version (using Google Translate) of the refined n-gram lexicon provided by Davidson et al. (2017), as well as the original lexicon to account for Dutch-English code switching, a phenomenon often encountered on social media (Broersma, 2009; Das & Gambäck, 2015). The original and machine translated n-gram lexicon total 335 hate speech terms (e.g. 'dumb monkey', 'domme aap').

After filtering the data set using all lexicons combined (totaling 448 entries), 10850 tweets remained, which corresponds to roughly 10% of the original dataset. This is the subset used in the experiments and will be referred to as `HSD` (Hate Speech Data) from this point forward. The remaining 90% does not contain any hate speech keywords. All tweets need to be labeled prior to classification, and filtering cuts down the large size of the data set, while including hate speech per the working definition (explained in Section 4) in the subset of filtered data. The remaining 90% will therefore not be considered in this hate speech classification task.

It should be noted that filtering data based on a collection of keywords can introduce bias. Tweets that do contain hateful messages but do not mention any of the keywords from our lexicons are not included in the final data set. In the opposite direction, tweets that contain hate speech words but do not intend harm are included in the final data set (e.g. when a user tweets a quote, or when a user refutes the use of the term). While lexicons are widely used for filtering data sets, it is not the most accurate method: 5% of filtered tweets were labeled as Hate Speech by Davidson et al. (2017) and 11.6% by Burnap and Williams (2015). Possible alternatives to filtering using lexicons will be explored in Section 7.

3.2 Text Pre-processing

The tweets were presented in their original form to the annotators and pre-processed to remove noise and distill essential information for classification by the model.

The steps that were taken in the pre-processing of the data are as follows:

- Removal of hyperlinks
- Converting text to lowercase
- Removal of user mentions (always begin with an at sign, e.g. '@rivm')
- Removal of non-alphabetic characters, including numerals, emoji and multiple whitespaces
- Tokenization of text
- Removal of stopwords

An example of a tweet before and after preprocessing can be found in Table 3.1.

It is worth noting that stemming and lemmatization as a pre-processing step for data in a classification task has been a separate topic of research and has produced varying results, as shown by Bao, Quan, Wang, and Ren (2014), Magliani, Fontanini, Fornacciari, Manicardi, and Iotti (2016), and Pradana and Hayaty (2019). It is certain that stemming and lemmatization of texts help reduce the feature space of data, but there is inconclusive evidence that it also improves the accuracy of a system. It was not included in the general preprocessing pipeline, but rather performed whenever a word could not be found in one of the lexicons (EmoLex by Mohammad and Turney (2013) and NRC-EIL by Mohammad (2018)) used to construct the features, see Section 5. Whenever possible, the word in its original form would be used, otherwise, the stemmed version of the word would be used (using NLTK SnowballStemmer[2]) in order to preserve as much information as possible.

A collection of two sets of stop words were used in the pre-processing of the data: the set of stop words provided by the NLTK toolkit (Loper & Bird, 2002) consisting of 101 words, and the set of stop words by Balucha (2014) consisting of 191 words, totaling a collection of 292 stop words[3].

3.3 Preliminary Data Analysis

After pre-processing the data, a preliminary analysis was conducted. In Table 3.2, the 20 most *common* hate speech terms found in HSD are listed, with their counts and meaning in English (adapted

[2] `https://www.nltk.org/_modules/nltk/stem/snowball.html`

[3] Full list can be found on the repository `https://github.com/Amber-ch/BSc_thesis_twitter`

Table 3.1: Example of tweet before and after preprocessing

Unprocessed tweet	Preprocessed tweet
'@KimBoon94 Ik hoop dat die gl trut een schop onder haar reet krijgt netjes gezegd en nergens meer aan de bak komt en van een bijstandsuitkering kan gaan leven wat een achterlijke trut is dit wijf zoveel mensen hebben zich opgeofferd voor de coronacrisis en nog steeds en zij fucking 0'	'hoop gl trut schop reet krijgt netjes gezegd nergens bak komt bijstandsuitkering gaan leven achterlijke trut wijf zoveel mensen opgeofferd coronacrisis fucking'

from Hatebase.org). Some words in this list are ambiguous, in the sense that they have both a hateful and non-hateful meaning: 'doos', means either 'stupid woman' or 'box'; 'nicht', can either refer to a homosexual male in a derogative manner, or one's niece/cousin; 'muts', means either 'stupid woman' or 'hat'. Still, most of the words in this list carry an unambiguous offensive or hateful meaning. Of the 20 terms in this list, nine are included in the translated lexicon by Davidson et al. (2017), and eleven are included in the Hatebase.org lexicon.

Additionally, the 20 most *relevant* terms according to TF-IDF can be seen in Table 3.3. TF-IDF consists of the two metrics Term Frequency (TF) and Inverse Document Frequency (IDF) and is used to measure the relevance of a particular term in relation to a set of documents.

Terms that occur more frequently across different documents (terms with high Document Frequency), are found to be less informative than those with low Document Frequency. Term Frequency refers to the raw count of that particular term over the full collection of tweets. Therefore, terms that occur often but in a small number of documents are given a higher TF-IDF score, indicating higher relevance with regards to the collection of documents. In this case, the goal of using TF-IDF scoring is to gain insight into the relevance of the hate speech keywords used in the creation of the data set.

Eight out of 20 terms found in Table 3.3 can also be found in Table 3.2, indicating that these Hate Speech terms are not only frequent within the filtered data set, but also relevant. Most other words in Table 3.3 are related to COVID-19, such as 'corona', 'protesten' (referring to heavily critiqued Black Lives Matter protests during the pandemic in Amsterdam (de Waard, 2020)), and 'RIVM' (Dutch National Institute for Public Health and Environment). Of these eight terms, three are included in the translated lexicon by Davidson et al. (2017), and five are included in the Hatebase.org lexicon.

The full lists of n-grams with their respective counts, and n-grams with their respective TF-IDF scores can be found on the repository[4].

4 Annotation

Since the hate speech classification task in this study is based on supervised learning, the data needs to be labeled by human annotators before the machine learning algorithm can be trained.

The following sections will introduce the working definitions of hate speech and offensive language, elaborate on the annotation process, and discuss the results of the annotation.

4.1 Definitions

Consistent with work by Davidson et al. (2017); Martins et al. (2018), three classes will be used: hate speech, offensive language, and neither (i.e. neither hate speech nor offensive language).

4.1.1 Hate Speech

The working definition of hate speech was constructed based on the papers by Davidson et al. (2017) and Martins et al. (2018). Although the definition by Davidson et al. (2017) captures the intent of hate speech messages, it fails to acknowledge that hate speech is not always targeted towards groups of people. Especially on the internet, (public) individuals such as politicians often fall victim to hate speech (Pelzer, Kaati, & Akrami, 2018). The definition by Martins et al. (2018) mentions the relevant aspect of subjectivity in messages of hate speech but does not explicitly mention groups or individuals. Components of both definitions have there-

[4] https://github.com/Amber-ch/BSc_thesis_twitter

Table 3.2: The 20 most common n-grams found in HSD with counts and meaning in English

Count	N-gram	Meaning
1716	achterlijke	mentally disabled person
1106	blanke	white person
1094	achterlijk	backward, retarded
655	doos	stupid woman (literally: box)
576	de schuld geven	blame
492	nicht	homosexual male
488	blanken	white people
468	debiel	moronic
455	flikker	homosexual male
352	zwarten	black people
309	tokkies	white trash
298	slaaf	slave
225	homo	homosexual person
218	hoeren	whores
218	muts	stupid woman (literally: hat)
160	mongolen	Mongoloid
149	allochtoon	Non-Dutch person
139	een blanke	a white person
138	trut	unpleasant woman
130	tokkie	white trash

Table 3.3: The 20 most relevant words according to TF-IDF, with meaning in English

TF-IDF	N-gram	Meaning
0.6146	corona	corona
0.2773	achterlijke	mentally disabled person
0.2658	mensen	people
0.1746	achterlijk	backward, retarded
0.1726	blanke	white person
0.1351	gaan	to go
0.1150	geven	to give
0.1122	covid	covid
0.1077	doos	stupid woman (literally: box)
0.1064	schuld	blame
0.0995	gaat	goes
0.0951	coronavirus	corona virus
0.0949	protesten	protests
0.0900	schuld geven	blame
0.0887	echt	for real
0.0863	rivm	Dutch National Institute for Public Health and Environment
0.0819	nicht	homosexual male
0.0797	testen	to test
0.0751	debiel	moronic
0.0750	flikker	homosexual male

fore been incorporated to compose a more complete definition of hate speech: *"A subjective statement meant to negatively target a group or individual. The tweet expresses hatred, or is intended to be derogatory, to humiliate, or to insult the members of the group/the individual. Hate speech tweets do not necessarily contain profanity."*

4.1.2 Offensive Language

Zampieri et al. (2019) define offensive language as profanity, but also insults and threats. According to the working definition of hate speech in Section 4.1.1, insults and threats fall under the category of hate speech. The working definition for offensive language is therefore as follows: *"A message that does not have discriminatory or hateful intent towards a group or individual, but does include profanity/swear words."*

A user might for example be using offensive language when quoting song lyrics or expressing strong emotions.

4.2 Procedure

For classification of the data, `HSD` is divided into `VHSD` (Validated Hate Speech Data) and `AHSD` (Annotated Hate Speech Data). `VHSD` is split into `VHSD1` and `VHSD2`, that each consist of 500 randomly selected tweets (about 5% of `HSD`) with no overlap between the sets. `VHSD1` and `VHSD2` are labelled by three annotators each. `AHSD` consists of 9850 tweets (about 90% of `HSD`), and is labelled by one annotator only.

The data was split into multiple subsets to investigate and account for the annotators' interpretations of the class definitions in this task. The data sets that have been labeled by multiple annotators (`VHSD`) will therefore be compared to the full data set (HSD), which includes data that has been labeled by only one annotator (`AHSD`). A visualization of the different subsets can be found in Figure 4.1.

The annotators are Bachelor's students who are native Dutch speakers and fluent English speakers. All received an annotation guide with some background information, the class definitions, typology of abusive language as defined by Waseem et al. (2017), and instructions for Universal Data Tool, which was used for the actual annotation. Since the labeling happened before any text reduction of the tweets as to not influence the meaning of the texts, filtering of duplicate tweets happened manually by the annotators.

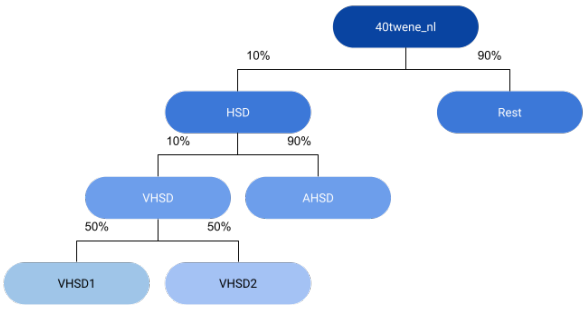

Figure 4.1: Division of 40twene_nl data set into subsets

4.3 Agreement

Krippendorff's alpha was calculated to investigate inter-coder agreement: the percentage of instances where all 3 annotators assigned the same label to a particular tweet. In `VHSD1` the agreement was 29.2%, and in `VHSD2` the agreement was 7.9%. These numbers are low compared to Sigurbergsson and Derczynski (2020), with the agreement being around 40%, and especially low compared to Davidson et al. (2017), with the agreement being 92%. Findings by Waseem (2016) suggest that inter-rater agreement among amateur annotators is likely to be lower than agreement among expert annotators. The same findings confirmed the claim by Ross et al. (2017) that hate speech is hard to classify unless the annotators have extensive knowledge on the topic. The low scores could also be the result of unclear definitions or subtasks. Further evaluation of the annotators and the annotation process overall can be found in Section 7.

Since Hate Speech and Offensive Language are the two most similar classes, the expectation is that labels often differed across annotators for these classes specifically. The similarity between classes and low agreement raised the question if merging these classes, transforming the multiclass classification task into a binary classification task, would result in a higher percentage of inter-rater agreement.

Merging the classes increased the agreement slightly, with VHSD1 rising to 35.5%, and VHSD2 to 10.9%. However, since it is now a binary classification task, the chance of a tweet being classified to either class is now $\frac{1}{2}$ instead of $\frac{1}{3}$. Therefore, it can be argued that the increased agreement is not actually an improvement.

This indicates that in many cases, the difference between a hateful/offensive tweet and a benevolent tweet proved difficult to distinguish. To further investigate the consequences of merging classes, both the binary classification task and the multiclass classification task will be considered in Section 6.

4.4 Results

Consistent with previous work, a majority vote was performed on all the tweets belonging to VHSD. While the agreement was low, a total of 703 tweets from VHSD remained for which the majority vote could be calculated. The tweets that did not have a majority were excluded.

The data set used in the multiclass VHSD classification task, VHSDm, consists of 703 tweets, of which 20.8% was categorized as Hate Speech, 16.1% as Offensive, and 63.3% as Neither. For the binary VHSD classification task, VHSDb, the Hate Speech/Offensive Language class comprised around $\frac{1}{3}$ (36.8%) of the data.

Combined with the tweets from AHSD, the full data set used in the multiclass HSD classification task, HSDm, consists of 9099 tweets, of which 27.2% was categorized as Hate Speech, 39.7% as Offensive, and 33.1% as Neither. For the binary HSD classification task, HSDb, the Hate Speech/Offensive Language class comprised around $\frac{2}{3}$ (66.9%) of the tweets. For both sets, the remaining 751 tweets were labeled as duplicates and therefore removed from the dataset.

The annotation results of the binary classification tasks can be found in Table 4.1. For the multiclass classification tasks, results can be found in Table 4.2.

The low agreement scores for the VHSD tweets are expected to also translate into the AHSD tweets, if there had been multiple annotators. However, since AHSD was annotated by one person only, there is no way to validate the tweets by calculating the majority vote. The annotations of this data set are therefore likely more subject to the annotators' personal beliefs and opinions, which can influence the performance of the classification system. The classification results of AHSD should therefore be interpreted with caution.

5 Features

In order to train the model, a set of features was constructed. The following sections will discuss the lexicons that were used to create the features, and evaluate the importance of those features in the classification task.

5.1 Lexicons

The features were constructed using the NRC Word-Emotion Association Lexicon, or EmoLex for short (Mohammad & Turney, 2013), and the NRC Emotion Intensity Lexicon, or NRC-EIL (Mohammad, 2018). Both EmoLex and NRC-EIL were modeled after the eight basic emotions as described by Plutchik (1980): anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The lexicons were originally published in English, and later machine translated by the original authors into a multitude of languages, including Dutch.

For each word that occurs in EmoLex, scores are given for each of the eight emotions, as well as a flag for the polarity of the word (positive or negative). The higher the score of an emotion for a word, the more accurate the word conveys that particular emotion.

For each word in NRC-EIL, intensity scores for each of the eight emotions are given. A high anger intensity score indicates that a word is associated with a high degree (or amount) of anger.

First, each word was looked up in EmoLex to determine the scores of all of the above-mentioned eight emotions, as well as the flags for negative and positive polarity. These scores and flags were vectorized to construct the overall emotion model of the tweet. Then, each word was searched for in NRC-EIL, to determine the anger intensity of that word, consistent with the method of Martins et al. (2018). These scores were summed to obtain the overall anger intensity score of the tweet, which was added to the feature vector. The emotion scores, positive and negative word counts, and anger intensity scores made up a total of 11 features per

Table 4.1: Annotation results for binary classification tasks

	VHSDb		HSDb	
Class	**# Tweets**	**Percentage**	**# Tweets**	**Percentage**
Hate Speech/Offensive Language	259	36.8%	6089	66.9%
Neither	444	63.3%	3010	33.1%

Table 4.2: Annotation results for multiclass classification tasks

	VHSDm		HSDm	
Class	**# Tweets**	**Percentage**	**# Tweets**	**Percentage**
Hate Speech	146	20.8%	2479	27.2%
Offensive Language	113	16.1%	3610	39.7%
Neither	444	63.3%	3010	33.1%

tweet.

5.2 Importance of Features

To evaluate the relevance of the features that were constructed, they were ranked according to information gain (or decrease in entropy) in relation to the different classes[5].

To understand the concept of entropy, the concept of information content (or surprisal) must be introduced first. The information content of an event is closely related to the probability of that event happening. As the probability of event E increases, the surprisal decreases. To illustrate, for a hypothetical data set only containing tweets that are labeled as hate speech, the chances of a tweet belonging to the hate speech class (event E) are 100%. This event therefore yields no information/surprise. When more classes are added, the chances of a tweet belonging to the hate speech class decrease, and the information content/surprisal increases.

The entropy of a variable is the average information content given by event E, when taking into account all of the possible outcomes (i.e. the different classes). As such, features that make the task more *predictable* lower the overall entropy of the classification task, and are ranked and scored higher.

The performed evaluation determines to what extent a feature helps distinguish between classes, by calculating the contribution of each feature to the decrease of overall entropy. The more a feature contributes to the decrease of overall entropy, the more useful it is deemed in the classification task.

[5]Using Weka InfoGainAttributeEval `https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html`

5.3 Results

The feature rankings can be seen in Table 5.1. Consistent with the findings of Martins et al. (2018), intensity is the highest-ranking feature across all experiments. Where anger was expected to be the second-highest-ranking feature, this was only the case in one of four experiments: in the other three, it ranked third, fourth, and sixth. The positive emotions anticipation and joy ranked higher than expected.

After closer inspection of the scores, it can be seen that the features in second place or lower are scored (several) powers smaller than intensity: 0.166 versus 0.021 for anticipation in VHSDb, 0.199 versus 0.0012 for anger in HSDb, 0.27 versus 0.027 for anticipation in VHSDm, and 0.216 versus 0.0022 for anticipation in HSDm.

6 Classifier

The following sections will introduce the model used for the classification of the tweets, explain the choice of hyperparameters, and discuss the results of the model.

6.1 Support Vector Machine

As Davidson et al. (2017) and Martins et al. (2018) have found the SVM (Support Vector Machine) to

Table 5.1: Importance of features in tweet classification

	VHSDb		HSDb		VHSDm		HSDm	
Feature	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**
Anger	0.0064	6	0.0012	2	0.019	4	0.0021	3
Anticipation	0.021	2	0.0009	4	0.027	2	0.0022	2
Disgust	0.0055	8	0.0007	6	0.015	6	0.0017	4
Fear	0.0012	11	0.0009	3	0.011	9	0.0013	8
Intensity	0.166	1	0.199	1	0.270	1	0.216	1
Joy	0.0066	5	0.0008	5	0.013	7	0.0016	5
Negative	0.0049	9	0.0007	7	0.0156	5	0.0013	7
Positive	0.0062	7	0.0006	9	0.0067	10	0.0012	9
Sadness	0.0099	3	0.0005	10	0.0211	3	0.0010	10
Surprise	0.0081	4	0.0004	11	0.011	8	0.0006	11
Trust	0.0030	10	0.0006	8	0.0056	11	0.0013	6

perform well in hate speech classification tasks, it was also selected for these experiments. The SVM is a supervised machine learning method used for the classification of data belonging to two or more classes, by choosing the decision boundary (hyperplane that separates the data) with the largest possible margin between classes.

Consistent with the method by Davidson et al. (2017), testing sets corresponding to 10% of each data set were held out. The remaining 90% of the data sets were used to evaluate the performance of the models, using 5-fold cross-validation.

The SVM was used with balanced class weights, which adjusts the weights of the classes to be inversely proportional to the frequencies of the classes, preventing bias towards the prevalent class. Another hyperparameter that was considered is the kernel shape used in the SVM, which is based on the shape of the data points. The kernel defines the set of mathematical formulae used to calculate the decision boundary. For example, a linear kernel returns a decision boundary that is a straight line (in two-dimensional space).

To determine the shape of the data, the data was plotted using Principal Component Analysis (PCA) since the data is comprised of more than three dimensions. PCA allows multidimensional data to be visualized in a two-dimensional space, where the distance between data points is maximized, therefore facilitating the identification of clusters.

Figures 6.1, 6.2, 6.3, and 6.4 show the results of PCA on the different data sets. Since the majority of data points seem to be overlapping in all figures, a non-linear kernel shape should be considered. The default non-linear kernel option is the radial basis function (RBF) kernel, which was selected for all classification tasks.

6.2 Binary Classification

The mean overall evaluation metrics of the binary classification tasks can be found in Table 6.1.

The observed mean accuracy ($\mu = 0.5420, \sigma = 0.0378$) suggests that the system in VHSDb performs slightly better than chance ($\frac{1}{2}$). A one-sample t-test however did not find the system to be performing significantly different from chance, $t(4) = 2.4851, p = 0.678$. The system in HSDb seems to be performing approximately according to chance ($\frac{1}{2}$). A one-sample t-test showed that the accuracy of this model ($\mu = 0.4780, \sigma = 0.0146$) is actually significantly less than $\frac{1}{2}$, $t(4) = 3.3641, p = 0.0282$. For all statistical tests, an alpha level of 0.05 was used.

Figure 6.5 and 6.6 show the normalized confusion matrices of predictions given by the binary classification models. Figure 6.5 shows a bottom-left to upper-right diagonal, meaning that the system misclassified the majority of the data. Figure 6.6 shows that the majority of tweets belonging to the merged Hate Speech/Offensive Language class were incorrectly classified as Neither.

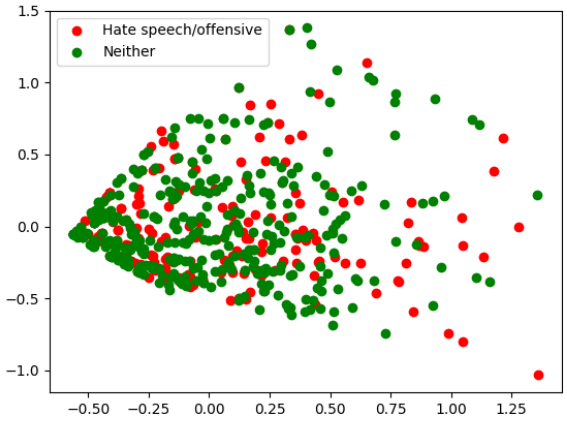

Figure 6.1: PCA of VHSDb

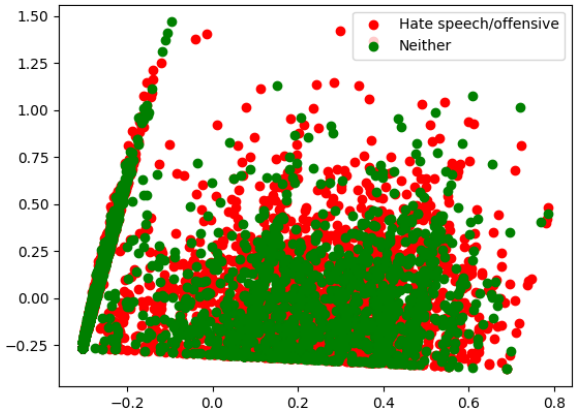

Figure 6.2: PCA of HSDb

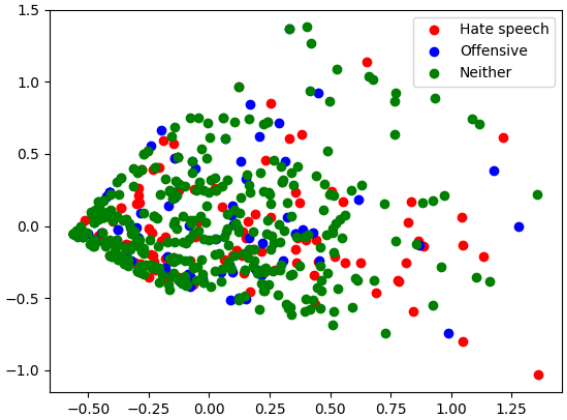

Figure 6.3: PCA of VHSDm

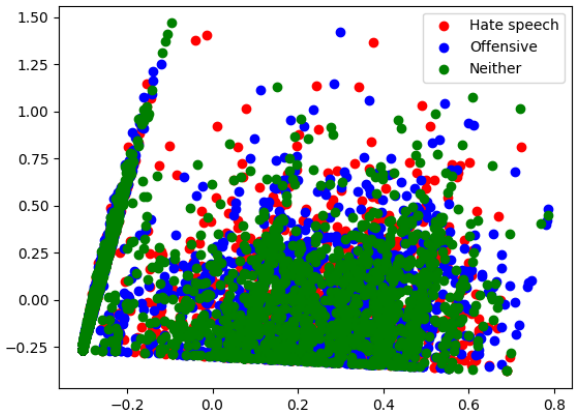

Figure 6.4: PCA of HSDm

Table 6.1: Mean overall evaluation metrics of the SVM with 5-fold Cross Validation for all Hate Speech Classification experiments

	VHSDb	**HSDb**	**VHSDm**	**HSDm**
Accuracy	0.5420	0.4780	0.4280	0.3557
Precision	0.5263	0.4965	0.35235	0.3256
Recall	0.5286	0.4958	0.3586	0.3348
F1	0.5217	0.4699	0.3437	0.3133

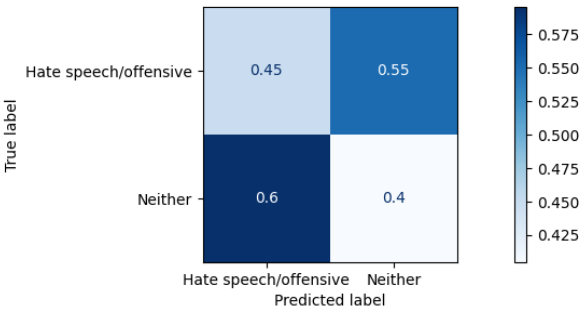

Figure 6.5: Normalized confusion matrix of VHSDb

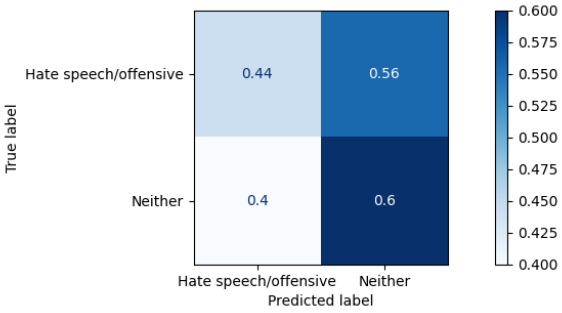

Figure 6.6: Normalized confusion matrix of HSDb

Figure 6.7: Normalized confusion matrix of VHSDm

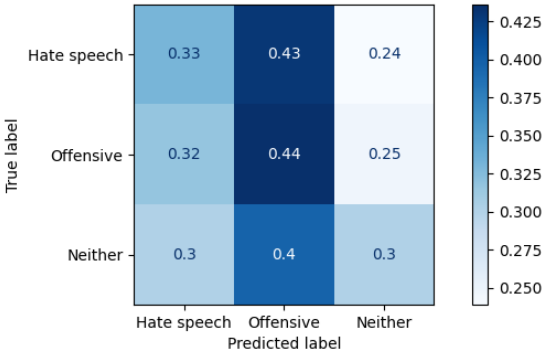

Figure 6.8: Normalized confusion matrix of HSDm

6.3 Multiclass Classification

Since an SVM in its usual form is used for binary classification problems, a one-versus-rest setup is used. This way, the SVM can still be used in a multiclass setting. For each class, a separate binary model is trained to distinguish between class A and the rest of the data.

The mean overall evaluation metrics of the multiclass classification tasks can be found in Table 6.1.

Since this experiment involved a three-class classification task, a system that performs on chance level will have a mean accuracy score of $\frac{1}{3}$. The accuracy score of the system used in VHSDm seems to be slightly higher than chance. To get a conclusive answer, a one-sample t-test was performed on the data. The mean accuracy of the model in VHSDm ($\mu = 0.4280, \sigma = 0.0325$) was found to be significantly greater than $\frac{1}{3}$, $t(4) = 6.5122, p = 0.00287$. The accuracy of the model in HSDm ($\mu = 0.3557, \sigma = 0.009$) was found to be significantly greater than $\frac{1}{3}$, $t(4) = 5.5589, p = 0.00512$.

Figure 6.7 and 6.8 show the normalized confusion matrices of predictions given by the multiclass classification models. The system in VHSDm most often misclassified tweets belonging to the Offensive class as Neither. Figure 6.8 shows that the system in HSDm most often classified tweets as Offensive.

7 Discussion

In the following sections, the results of this study will be evaluated. Improvements on this work and suggestions for future work will also be discussed.

7.1 Machine Translation

Since the lexicons were translated instead of the Twitter data, the computational expense and time spent on translation were kept to a minimum. Accuracy is the other side of the trade-off. While machine translating lexicons is an efficient solution, the contexts in which the lexical entries are used are missing.

The missing context can express itself in the incorrect translation of words. Some words may have multiple possible translations for which the most fitting option is picked based on contextual information. Context is especially important for slang words, where often a secondary (offensive) meaning will be attached to existing phrases (e.g. the term 'coon', which can refer to the animal raccoon but is also used as a racial slur).

Given the informal nature of these secondary meanings, they are often not included in the databases of machine translation software.

Aside from missing contextual information when only machine translating the lexicons, information is missing with regards to secondary meanings of words (slang). It is therefore recommended to further investigate the use of machine translation on tweets, provided that enough time and resources are available. Furthermore, since the information from secondary meanings is often missing, it would be recommended to either consult an expert or make use of additional slang lexicons to verify translations of lexicons and/or tweets.

7.2 Features

Consistent with work by Martins et al. (2018), anger was the only emotion of which the intensity was calculated. Since anger intensity was evaluated as the most important feature both in this study and by Martins et al., it would be interesting to investigate the use of intensity scores of the full available range of emotions as features for hate speech classification.

The scores of the emotion features indicated that they were irrelevant in the classification process. Since the lexicons used to construct the features were first published in English and later machine translated into different languages, it does not always fully capture the nuances of those languages, as explained in Section 7.1. While an English word may have multiple Dutch translations, only one is included in the lexicon. As such, many words that occurred in the data could not be found in the lexicon and could therefore not be scored.

The words that did find an exact match with the translations in the lexicon were not always scored either, since a large part of words in the lexicon simply are not assigned any scores. For neutral terms, e.g. 'molecular', this makes sense. However, words that are often used in offensive contexts, e.g. 'monkey', are also lacking scores, showing that the lexicons need to be expanded. The incompleteness of the lexicons, and as a consequence, the underwhelming effect of the features, partly explain the

worse than chance performance of the classification systems.

Other possible additions to the feature space include the TF-IDF scores, as seen in work by Davidson et al. (2017), as well as the count of hate speech terms in a tweet. The TF-IDF scores of our data set were already calculated and compiled into a lexicon for preliminary data analysis, but unused as features in the end.

Martins et al. (2018) included a flag feature for the occurrence of hate speech words in a tweet. Since all tweets in our data set contained hate speech words per definition, this feature was left out. It might however be interesting to investigate the role of the *number* of hate speech words in a tweet on classification.

7.3 Classification

As with any supervised machine learning experiment, labeled data was needed to evaluate the performance of the SVM. Principal Component Analysis showed that many of the data points overlap each other, which suggested that it would be difficult to distinguish the data clusters based on the constructed features. This was confirmed by the classification results of the SVM. The results of the annotation and classification process are believed to be influenced by two main factors, the definitions and annotators, which will be discussed in the following sections.

7.3.1 Definitions

By compiling definitions from various previous works, the working definitions in this experiment are made more complete and nuanced than the ones found in previous work. Still, the definitions might not have been specific enough. Davidson et al. (2017) refrained from limiting the definition to threats and/or messages that incite violence, since a large part of hate speech would then be excluded. Given the high amount of hate speech found after annotation in the data set of this study (roughly 30%) compared to the amount found by Davidson et al. (roughly 5%), enough instances of hate speech should remain after limiting the definition to more extreme cases.

Narrowing the scope of the definition may result in better-defined clusters of data, compared to the majorly overlapping data points encountered at this stage. Without proper definitions of hate speech and offensive language, it is difficult to research possible methods to combat online hate speech. It is therefore advised to reevaluate hate speech and offensive language in a multidisciplinary setting where law and social science are also taken into consideration.

7.3.2 Annotators

Given the limited time and resources, the annotators involved were Bachelor's students instead of expert/professional annotators. Waseem (2016) found that inter-rater agreement among amateur annotators is likely to be lower than among expert annotators, and that amateur annotators are more likely to label data as hate speech than expert annotators. While the annotations have not been compared to those of experts, inter-rater agreement was found to be low. It would therefore be interesting to compare the annotations of experts, to see if the inter-rater agreement increases, and the amount of data labeled as hate speech decreases.

Furthermore, only three different annotators were involved per data set, since this was the minimum required to calculate a majority vote. It would however be beneficial to have more annotators review the data. While it may bring down agreement scores, since the chances of having a unanimous vote decrease when more annotators are involved, there are numerous benefits to increasing the number of annotators.

With three classes and three annotators, a tie is obtained when every annotator assigns a different class to a tweet. The chances of getting a tied vote decrease as more annotators are involved in the annotation process. Moreover, the individual choices of the annotators are less likely to skew the data towards one of the classes, as the number of annotators increases. This results in a more even distribution of votes and a more representative view of classification with regards to the majority vote, which is preferred over a unanimous vote by fewer annotators.

For future experiments, it is therefore recommended to not only involve expert annotators that have experience with similar annotation tasks, but also have the data reviewed by more annotators.

8 Conclusion

This study aimed to create an automated hate speech classifier, using data filtered on a Dutch and a machine translated lexicon. TF-IDF scoring showed that (translated) lexical entries were among the 20 most relevant n-grams of the new data set. The preliminary results were promising, and although the classes of the data proved difficult to distinguish by the annotators, the results of the machine classification were significantly above chance level.

Part of the data set was labeled by multiple annotators, where inter-rater agreement scores were found to be low. The low agreement scores could be indicative of definitions or subtasks that are not clear enough, or the need for expert annotators (Waseem, 2016). Since the human annotators seemed to struggle with distinguishing the different classes, it is expected that the system also had difficulty with this task. While the performance of the system was found to be significantly above chance level, there is a possibility that the data used for this task does not fully encompass the true scope of hate speech.

The other part of the data was labeled by one annotator, leaving it unvalidated. It is therefore not possible to draw any conclusions from this part of the data at this stage.

Of the constructed features, anger intensity was ranked the highest in terms of information gain, confirming its relevance in the classification of hate speech and offensive language found by Martins et al. (2018). The other features were scored several magnitudes smaller compared to the intensity feature, suggesting that emotional analysis might not be as useful for distinguishing between hate speech and offensive language as first thought.

Statistical tests showed that the accuracy of the binary classification systems was equal to or less than chance. While the accuracy of the multiclass classification systems was significantly higher than chance, it cannot be concluded that the systems were actually *successful* in the classification of hate speech. None of the obtained results exceeded the ones from previously created binary and three-class classification systems (Zampieri et al., 2019; Davidson et al., 2017; Martins et al., 2018). This is consistent with the findings by Waseem (2016), who found that systems trained on expert-annotated data outperformed systems on amateur-annotated data.

Training models on data labeled by multiple annotators instead of data labeled by one annotator did improve performance, suggesting that the majority vote conducted on the validated data decreases the chances of annotator error and that it helps the system in distinguishing between hate speech and offensive language.

Overall, the results of this study show a step in the right direction of automated hate speech classification systems for low-resource languages like Dutch. However, the steps taken in this process can still be improved upon and require critical review.

References

Alorainy, W., Burnap, P., Liu, H., Javed, A., & Williams, M. L. (2018). Suspended accounts: A source of tweets with disgust and anger emotions for augmenting hate speech data sample. In *2018 international conference on machine learning and cybernetics (ICMLC)* (Vol. 2, pp. 581–586).

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2021). A deep dive into multilingual hate speech classification. *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, 423–439.

Balucha. (2014). *Stop-words.* Google Code. Retrieved from `https://code.google.com/archive/p/stop-words/`

Bao, Y., Quan, C., Wang, L., & Ren, F. (2014). The role of pre-processing in Twitter sentiment analysis. In *International conference on intelligent computing* (pp. 615–624).

Bouma, G. (2015). N-gram frequencies for Dutch Twitter data. *Computational Linguistics in the Netherlands Journal*, *5*, 25–36.

Broersma, M. (2009). Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, *12*(4), 447–462.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, *7*(2), 223–242.

Caselli, T., & Basile, V. (2020, Sep). *40twene_nl.* OSF. Retrieved from `https://osf.io/pfnur`

Das, A., & Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier? *Association pour le Traitement Automatique des Langues (ATALA).*

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (Vol. 11).

de Waard, P. (2020, June 1). Kritiek op niet in-grijpen van burgemeester bij grote Black Lives Matter-betoging in Amsterdam. *De Volkskrant.*

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In *In proceedings of the ACL workshop on effective tools and methodologies for teaching natural language processing and computational linguistics. Philadelphia: Association for computational linguistics.*

Magliani, F., Fontanini, T., Fornacciari, P., Manicardi, S., & Iotti, E. (2016, Dec). A comparison between preprocessing techniques for sentiment analysis in Twitter. In *2nd international workshop on knowledge discovery on the web.*

Marinov, B., Spenader, J., & Caselli, T. (2020). Topic and emotion development among Dutch COVID-19 Twitter communities in the early pandemic. In *Proceedings of the third workshop on computational modeling of people's opinions, personality, and emotion's in social media* (pp. 87–98).

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. In *2018 7th brazilian conference on intelligent systems (BRACIS)* (pp. 61–66).

Mohammad, S. M. (2018). Word affect intensities. In *Proceedings of the 11th edition of the language resources and evaluation conference (lrec-2018).*

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. , *29*(3), 436–465.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).

Pelzer, B., Kaati, L., & Akrami, N. (2018). Directed digital hate. In *2018 IEEE international conference on intelligence and security informatics (ISI)* (pp. 205–210).

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3–33). Elsevier.

Pradana, A. W., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on Indonesian-language texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 375–380.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *3rd Workshop on Natural Language Processing for Computer-Mediated Communication Social Media*.

Sang, E. T. K. (2011). Het gebruik van Twitter voor taalkundig onderzoek. *TABU: Bulletin voor Taalwetenschap*, *39*(1/2), 62–72.

Shields, M. (2020, Feb 27). U.N. asks world to fight virus-spawned discrimination. *Reuters*.

Sigurbergsson, G. I., & Derczynski, L. (2020, May). Offensive language and hate speech detection for Danish. In *Proceedings of the 12th language resources and evaluation conference* (pp. 3498–3508).

Twitter. (2021). *About government and state-affiliated media account labels on Twitter.* Retrieved 01-06-2021, from `https://help.twitter.com/en/rules-and-policies/state-affiliated`

United Nations. (2019, May). *Strategy and plan of action on hate speech* (Tech. Rep.).

United Nations Human Rights Office of the High Commissioner. (2021, March). *Report: Online hate increasing against minorities, says expert* (Tech. Rep.).

Van der Veer, N., Boekee, S., & Hoekstra, H. (2020). *Nationale social media onderzoek 2020.* Newcom Research & Consultancy.

Van der Veer, N., Boekee, S., & Hoekstra, H. (2021). *Nationale social media onderzoek 2021.* Newcom Research & Consultancy.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., ... Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *International conference recent advances in natural language processing (RANLP)* (pp. 672–680).

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., ... Tromble, R. (2020). Detecting East Asian prejudice on social media. In (p. 162-172). Association for computational linguistics.

W3Tech. (2021). *Usage statistics of content languages for websites.* Retrieved 04-05-2021, from `https://w3techs.com/technologies/overview/content_language`

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In (pp. 78–84).

Weintraub-Reiter, R. (1998). Hate speech over the internet: A traditional constitutional analysis or a new cyber constitution. *Boston University Public Interest Law Journal*, *8*, 145.

Wetboek van Strafrecht. (2020, Jan). *Artikel 137d.* Retrieved July 2021, from `http://www.wetboek-online.nl/wet/Wetboek%20van%20Strafrecht/137d.html`

Younus, A., Qureshi, M. A., Asar, F. F., Azam, M., Saeed, M., & Touheed, N. (2011). What do the average Twitterers say: A Twitter model for public opinion analysis in the face of major political events. In *2011 international conference on advances in social networks analysis and mining* (pp. 618–623).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019, June). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 1415–1420).

A Precision and Recall Scores per Experiment, per Class

Table A.1: Precision and recall scores per class in VHSDb

Class	Precision	Recall
Hate Speech/Offensive	0.448	0.342
Neither	0.405	0.472

Table A.2: Precision and recall scores per class in HSDb

Class	Precision	Recall
Hate Speech/Offensive	0.444	0.677
Neither	0.6	0.363

Table A.3: Precision and recall scores per class in VHSDm

Class	Precision	Recall
Hate Speech	0.154	0.095
Offensive	0.250	0.222
Neither	0.381	0.5

Table A.4: Precision and recall scores per class in HSDm

Class	Precision	Recall
Hate Speech	0.331	0.289
Offensive	0.436	0.392
Neither	0.302	0.396