Gorrotoa sarean automatikoki detektatzeko teknikak ez dira oraindik perfektuak

Ziberjazarpena eta gorrotozko diskurtsoa handitzen ari dira, eta horrek jazarpenaren aurkako politikak eskatzen ditu. Hala ere, zaila da haiek detektatzea eta ikertzea, Facebook, Twitter eta gainerako sare sozialetan eta blogetan edukiak azkar ugaritzen ari direlako. Gainera, gorrotozko diskurtsoa identifikatzea konplexua izan daiteke, hiztunak kaltea eragiteko asmoa ote duen argitu behar delako. Gorrotozko diskurtsoa automatikoki atzemateko adimen artifizialeko teknikak baliatu dituzte. Ziberjazarpenaren garrantzia gorakada ikusita, baliabide gehiago behar dira detekzio-teknikak fintzeko.

Gorrotozko diskurtsoa da norbaiti buruz gaizki esaka aritzea arrazari edo generoari lotutako ezaugarriengatik. *Stormfront* foroak, gorrotozko diskurtsoaren datu-multzo berri bat argitaratu du, ikerketari laguntzeko asmoz. Eta lan horretarako adimen artifizialeko teknikak erabili dituzte, *GitHub* erreminta ezagunari esker eskuragarri daudenak.

«Gorrotozko diskurtsoan etniari eta generoari lotutako gorrotoa dira kategoriarik ohikoenak»

10.578 esaldi aztertu dituzte. Sistemak banaka sailkatzen du esaldi bakoitza: gorrotozko diskurtsoa ote den, ez den, edo berariazko harreman-kategoria bat, non gorrotozko diskurtsoa inplizitua baitago beste esaldi batzuekin konbinatzean. Sailkapen horren gidalerroak kontu handiz prestatu ziren, idazleen arteko koherentzia bermatzeko. Gero, esaldi laburregiak edo luzeegiak kendu zituzten, datu «garbiak» sortzeko.

Gorrotoa bilatzeko bidea

Datu-multzoa desorekatuta dago: gorrotorik gabeko esaldiak ugariagoak dira gorrotoa dutenak baino. Gorroto-indize bat kalkulatu zen, gorrotozko diskurtsoari lotutako hitzak identifikatzeko eta, beraz, gorrotoaren hiztegi bat osatzeko. Gorrotozko diskurtsoaren datu-basearekin gainjarrita, etnia eta generoa dira kategoriarik ohikoenak.

Artikuluak oinarrizko esperimentuak aurkezten ditu, gorrotozko testuen datu-multzo batean eginak. Datu-multzoko esaldiak etiketatuta daude —gorrotodunak edo gorrotorik ez dutenak—, esperimentuan egin ziren oharpenen baliozkotasuna frogatzeko eta etorkizuneko ikerketetarako erreferentzia ezartzeko.

Erroreak ere aztertu egin ziren. Sistemak «gorrotorik gabe» etiketaz sailkatzen zituen zenbait esaldi, lehenago eskuz «gorrotozkoa» etiketaz sailkatutakoak. Horren arrazoia izaten zen, oro har, sistemak testuingurua falta zuela. Eta kontrako akats-mota ere izaten zen; sistemak «gorrotozkoa» sailkatzen zituen zenbait esaldi, lehenago eskuz «gorrotorik gabe» etiketaz sailkatutakoak. Arrazoia izaten zen esaldiak ohiko hiztegi iraingarria erabiltzen zuela, kalterik egiteko asmorik gabe.

Esperimentuek gorroto-adierazpenak sailkatzemetodoen erronkak nabarmendu zituzten, batez ere testuingurua eta ezagutza funtsezkoak direnean emaitza zehatzak lortzeko. Gai garrantzitsua izanik, baliabide gehiago jarri beharko dira sarean gorrotoa detektatzeko teknikak hobeak izan daitezen.

InpAkta 00 (2023) 285-297
http://dx.doi.org/10.18653/v1/W18-5102

Hate speech dataset from a white supremacy forum

Ona de Gibert, Naiara Perez, Aitor García-Pablos, Montse Cuadros
HSLT Group at Vicomtech

ABSTRACT: Hate speech is commonly defined as any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic. Due to the massive rise of user-generated web content on social media, the amount of hate speech is also steadily increasing. Over the past years, interest in online hate speech detection and, particularly, the automation of this task has continuously grown, along with the societal impact of the phenomenon. This paper describes a hate speech dataset composed of thousands of sentences manually labelled as containing hate speech or not. The sentences have been extracted from Stormfront, a white supremacist forum. A custom annotation tool has been developed to carry out the manual labelling task which, among other things, allows the annotators to choose whether to read the context of a sentence before labelling it. The paper also provides a thoughtful qualitative and quantitative study of the resulting dataset and several baseline experiments with different classification models. The dataset is publicly available.

1. Introduction

The rapid growth of content in social networks such as Facebook, Twitter and blogs, makes it impossible to monitor what is being said. The increase of cyberbullying and cyberterrorism, and the use of hate on the Internet, make the identification of hate in the web an essential ingredient for anti-bullying policies of social media, as Facebook's CEO Mark Zuckerberg recently acknowledged[1]. This paper releases a new dataset of hate speech to further investigate the problem.

Although there is no universal definition for *hate speech*, the most accepted definition is provided by

[1] https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/

* **Corresponding author / Harremanetan jartzeko:** Ona de Gibert. HSLT Group at Vicomtech, Donostia/San Sebastián, Spain. e-mail: odegibert@vicomtech.org. **Contributed by** Arantza Del Pozo Echezarreta **editoreak egindako ekarpena**.
How to cite / Nola aipatu: de Gibert, Ona *et al.* (2023). «Hate Speech Dataset from a White Supremacy Forum», *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, 2018, pp. 11-20. (http://dx.doi.org/10.18653/v1/W18-5102)

Nockleby (2000): "any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic". Consider the following[2]:

(1) "God bless them all, to hell with the blacks"

This sentence clearly contains hate speech against a target group because of their skin colour. However, the identification of hate speech is often not so straightforward. Besides defining hate speech as a verbal abuse directed to a group of people because of specific characteristics, other definitions of hate speech in previous studies care to include the speaker's determination to inflect harm (Davidson *et al.*, 2017).

In all, there seems to be a pattern shared by most of the literature consulted (Nockleby, 2000; Djuric *et al.*, 2015; Gitari *et al.*, 2015; Nobata *et al.*, 2016; Silva *et al.*, 2016; Davidson *et al.*, 2017), which would define hate speech as *a)* a deliberate attack, *b)* directed towards a specific group of people, and *c)* motivated by actual or perceived aspects that form the group's identity.

This paper presents the first public dataset of hate speech annotated on Internet forum posts in English at sentence-level. The dataset is publicly available in GitHub[3]. The source forum is Stormfront[4], the largest online community of white nationalists, characterised by pseudo-rational discussions of race (Meddaugh and Kay, 2009), which include different degrees of offensiveness. Stormfront is known as the first hate website (Schafer, 2002).

The rest of the paper is structured as follows: Section 2 describes the related work and contextualises the work presented in the paper; Section 3 introduces the task of generating a manually labelled hate speech dataset; this includes the design of the annotation guidelines, the resulting criteria, the inter-annotator agreement and a quantitative description of the resulting dataset; next, Section 4 presents several baseline experiments with different classification models using the labelled data; finally, Section 5 provides a brief discussion about the difficulties and nuances of hate speech detection, and Section 6 summarises the conclusions and future work.

2. Related Work

Research on hate speech has increased in the last years. The conducted studies are diverse and work on different datasets; there is no official corpus for the task, so usually authors collect and label their own data. For this reason, there exist few publicly available resources for hate speech detection.

Hatebase[5] is the an online repository of structured, multilingual, usage-based hate speech. Its vocabulary is classified into eight categories: archaic, class, disability, ethnicity, gender, nationality, religion, and sexual orientation. Some studies make use of Hatebase to build a classifier for hate speech (Davidson *et al.*, 2017; Serra *et al.*, 2017; Nobata *et al.*, 2016). However, Saleem *et al.* (2016) prove that keyword-based approaches succeed at identifying the topic but fail to distinguish hateful sentences from clean ones, as the same vocabulary is shared by the hateful and target community, although with different intentions.

Kaggle's Toxic Comment Classification Challenge dataset[6] consists of 150k Wikipedia comments annotated for toxic behaviour. Waseem and Hovy (2016) published a collection of 16k tweets classified into racist, sexist or neither. Sharma *et al.* (2018) collected a set of 9k tweets containing harmful speech and they manually annotated them based on their degree of hateful intent. They describe three different classes of hate speech. The definition on which this paper is based overlaps mostly with their Class I, described as speech *a)* that incites violent actions, *b)* directed at a particular group, and *c)* with the intention of conveying hurting sentiments.

[2] The examples in this work may contain offensive language. They have been taken from actual web data and by no means reflect the authors' opinion.

[3] https://github.com/aitor-garcia-p/hate-speech-dataset

[4] www.stormfront.org

[5] https://www.hatebase.org/

[6] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

Google and Jigsaw developed a tool called Perspective[7] that measures the "toxicity" of comments. The tool is published as an API and gives a toxicity score between 0 and 100 using a machine learning model. Such model has been trained on thousands of comments manually labelled by a team of people[8]; to our knowledge, the resulting dataset is not publicly available.

The detection of hate speech has been tackled in three main different ways. Some studies focus on subtypes of hate speech. This is the case of Warner and Hirschberg (2012), who focus on the identification of anti-Semitic posts versus any other form of hate speech. Also in this line, Kwok and Wang (2013) target anti-black hate speech. Badjatiya *et al.* (2017); Gambäck and Sikdar (2017) study the detection of racist and sexist tweets using deep learning.

Other proposals focus on the annotation of hate speech as opposed to texts containing derogatory or offensive language (Davidson *et al.*, 2017; Malmasi and Zampieri, 2017, 2018; Watanabe *et al.*, 2018). They build multi-class classifiers with the categories "hate", "offensive", and "clean".

Finally, some studies focus on the annotation of hate speech versus clean comments that do not contain hate speech (Nobata *et al.*, 2016; Burnap and Williams, 2015; Djuric *et al.*, 2015). Gitari *et al.* (2015) follow this approach but further classify the hateful comments into two categories: "weak" and "strong" hate. Del Vigna *et al.* (2017) conduct a similar study for Italian.

In all, experts conclude that annotation of hate speech is a difficult task, mainly because of the data annotation process. Waseem (2016) conducted a study on the influence of annotator knowledge of hate speech on classifiers for hate speech. Ross *et al.* (2016) also studied the reliability of hate speech annotations and acknowledge the importance of having detailed instructions for the annotation of hate speech available.

This paper aims to tackle the inherent subjectivity and difficulty of labelling hate speech by following strict guidelines. The approach presented in this paper follows (Nobata *et al.*, 2016; Burnap and Williams, 2015; Djuric *et al.*, 2015) (i.e., "hateful" versus "clean"). Furthermore, the annotation has been performed at sentence level as opposed to full-comment annotation, with the possibility to access the original complete post for each sentence. To our knowledge, this is the first work that releases a manually labelled hate speech dataset annotated at sentence level in English posts from a white supremacy forum.

3. Hate Speech Dataset

This paper presents the first dataset of textual hate speech annotated at sentence-level. Sentence-level annotation allows to work with the minimum unit containing hate speech and reduce noise introduced by other sentences that are clean.

A total number of 10,568 sentences have been extracted from Stormfront and classified as conveying hate speech or not, and into two other auxiliary classes, as per the guidelines described in Section 3.2. In addition, the following information is also given for each sentence: a post identifier and the sentence's position in the post, a user identifier, a sub-forum identifier[9]. This information makes it possible re-build the conversations these sentences belong to. Furthermore, the number of previous posts the annotator had to read before making a decision over the category of the sentence is also given.

3.1. Data extraction and processing

The content was extracted from Stormfront using web-scraping techniques and was dumped into a database arranged by sub-forums and conversation threads (Figea *et al.*, 2016). The extracted forum

[7] https://www.perspectiveapi.com

[8] https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html

[9] All the identifiers provided are fake placeholders that facilitate understanding relations between sentences, Stormfront users, etc., but do not point back to the original source.

content was published between 2002 and 2017. The process of preparing the candidate content to be annotated was the following:

1. A subset of 22 sub-forums covering diverse topics and nationalities was random-sampled to gather individual posts uniformly distributed among sub-forums and users.
2. The sampled posts were filtered using an automatic language detector[10] to discard non-English texts.
3. The resulting posts were segmented into sentences with ixa-pipes (Agerri *et al.*, 2014).
4. The sentences were grouped forming batches of 500 complete posts (~ 1,000 sentences per batch).

The manual annotation task was divided into batches to control the process. During the annotation of the first two batches, the annotation procedure and guidelines were progressively refined and adapted. In total, 10,568 sentences contained in 10 batches have been manually annotated.

A post-processing step was performed to filter excessively long or short sentences. The cleansing process removed sentences shorter than 3 words or longer than 50 words. In total, 652 sentences were discarded, which represent 6.17% of the original dataset. The resulting dataset (henceforth referred to as the "clean" dataset) is the one described in Section 3.4 and used for experimentation in Section 4. The release of the dataset will contain both the raw sentences without any post-processing and the clean version, both annotated.

3.2. *Annotation guidelines*

Schmidt and Wiegand (2017) acknowledge that the procedure for hate speech annotation is fairly vague in previous studies, which translates into low agreement scores. In this study, all the annotators together created and discussed the guidelines to ensure all participants had the same understanding of hate speech. The final guidelines consider 4 types of sentences:

3.2.1. HATE

Sentences in this category contain hate speech. Hate speech is a

a) deliberate attack
b) directed towards a specific group of people
c) motivated by aspects of the group's identity.

The three premises must be true for a sentence to be categorized as HATE. Consider the following examples:

(2) "Poor white kids being forced to treat apes and parasites as their equals."
(3) "Islam is a false religion however unlike some other false religions it is crude and appeals to crude people such as arabs."

In (2), the speaker uses "apes" and "parasites" to refer to children of dark skin and implies they are not equal to "white kids". That is, it is an attack to the group composed of children of dark skin based on an identifying characteristic, namely, their skin colour. Thus, all the premises are true and (2) is a valid example of HATE. Example (3) brands all people of Arab origin as crude. That is, it attacks the group composed of Arab people based on their origin. Thus, all the premises are true and (3) is a valid example of HATE.

3.2.2. NOHATE

This label is used to categorise sentences that do not convey hate speech per the established definition. Consider the following examples:

(4) "Where can I find NS speeches and music, also historical, in mp3 format for free download on the net."
(5) "I know of Chris Rock and subsequently have hated him for a long time."

Example (4) mentions National Socialism ("NS"), but the user is just interested in documentation about it. Therefore, the sentence itself is not an attack, i.e., premise a) is not true, despite the sound assumption that the speaker forms part of a hating community. Thus, (4) is not a valid instance of HATE. Example (5) is directed towards an individu-

[10] https://github.com/shuyo/language-detection/blob/wiki/ProjectHome.md

al; thus, premise b) is false and the sentence is not a valid example of HATE, despite the sound assumption that the attack towards the individual is based on his skin colour.

Finally, it must be emphasized that the presence of pejorative language in a sentence cannot systematically be considered sufficient evidence to confirm the existence of hate speech. The use of "fag" in the following sentence:

(6) "Two black fag's holding hands."

cannot be said to be a deliberate attack, taken without any more context, despite it likely being offensive. Therefore, it cannot be considered H AT E.

3.2.3. Relation

When (6) (repeated as (7.1)) is read in context:

(7.1) "Two black fag's holding hands."
(7.2) "That's Great!"
(7.3) "That's 2 blacks won't be having kids."

it clearly conveys hate speech. The author is celebrating that two people belonging to the black minority will not be having children, which is a deliberate attack on a group of people based on an identifying characteristic. The annotation at sentence-level fails to discern that there exists hate speech in this example. The label RELATION is for specific cases such as this, where the sentences in a post do not contain hate speech on their own, but the combination of several sentences does. Consider another example:

(8.1) "Probably the most disgusting thing I've seen in the last year."
(8.2) "She looks like she has some African blood in her, or maybe it's just the makeup."
(8.3) "This is just so wrong."

Each sentence in isolation does not convey hate speech: in (8.1) and (8.3), a negative attitude is perceived, but it is unknown whether it is targeted towards a group of people; in (8.2), there is no hint of an attack, not even of a negative attitude. However, the three sentences together suggest that having "African blood" makes a situation (whatever "this" refers to) disgusting, which constitutes hate speech according to the definition proposed.

The label RELATION is given separately to all the sentences that need each other to be understood as hate speech. That is, consecutive sentences with this label convey hate speech but depend on each other to be correctly interpreted.

3.2.4. Skip

Sentences that are not written in English or that do not contain information as to be classified into HATE or NOHATE are given this label.

(9) "Myndighetene vurderer n om de skal f permanent oppholdstillatelse."
(10) "YouTube - Broadcast Yourself."

Example (9) is in Norwegian and (10) is irrelevant both for HATE and NOHATE.

3.3. Annotation procedure

In order to develop the annotation guidelines, a draft was first written based on previous similar work. Three of the authors annotated a 1,144-sentence batch of the dataset following the draft, containing only the categories HATE, NOHATE and SKIP. Then, they discussed the annotations and modified the draft accordingly, which resulted in the guidelines presented in the previous section, including the RELATION category. Finally, a different batch of 1,018 sentences was annotated by the same three authors adhering to the new guidelines in order to calculate the inter-annotator agreement.

Table 1 shows the agreements obtained in terms of the average percent agreement (*avg %*), average Cohen's kappa coefficient (Cohen, 1960) (*avg k*), and Fleiss' kappa coefficient (Fleiss, 1971) (*fleiss*). The number of annotated sentences (# sent) and the number of categories to label (# cat) are also given for each batch. The results are in line with similar works (Nobata *et al.*, 2016; Warner and Hirschberg, 2012).

Table 1
Inter-annotator agreements on batches 1 and 2

	# sent	# cat	*avg %*	*avg k*	*fleiss*
1	1,144	3	91.03	0.614	0.607
2	1,018	4	90.97	0.627	0.632

All the annotation work was carried out using a web-based tool developed by the authors for this purpose. The tool displays all the sentences belonging to the same post at the same time, giving the annotator a better understanding of the post's author's intention. If the complete post is deemed insufficient by the annotator to categorize a sentence, the tool can show previous posts to which the problematic post is answering, on demand, up to the first post in the thread and its title. This consumption of context is registered automatically by the tool for further treatment of the collected data.

As stated by other studies, context appears to be of great importance when annotating hate speech (Watanabe *et al.*, 2018). Schmidt and Wiegand (2017) acknowledge that whether a message contains hate speech or not can depend solely on the context, and thus encourage the inclusion of extra-linguistic features for annotation of hate speech. Moreover, Sharma *et al.* (2018) claim that context is essential to understand the speaker's intention.

3.4. Dataset statistics

This section provides a quantitative description and statistical analysis of the clean dataset published. Table 2 shows the distribution of the sentences over categories. The dataset is unbalanced as there exist many more sentences not conveying hate speech than "hateful" ones.

Table 3 refers to the subset of sentences that have required reading additional context (i.e. previous comments to the one being annotated) to make an informed decision by the human annotators. The category HATE is the one that requires more context, usually due to the use of slang unknown to the annotator or because the annotator needed to find out the actual target of an offensive mention.

Table 2
Distribution of sentences over categories in the clean dataset

Assigned label	# sent	%
HATE	1,119	11.29
NOHATE	8,537	86.09
RELATION	168	1.69
SKIP	92	0.93
total	9,916	100.00

Table 3
Percentage of sentences for which the human annotators asked for additional context

	Context used	No context used
HATE	22.70	77.30
NOHATE	8.00	92.00

The remaining of the section focuses only on the subset of the dataset composed of the categories HATE and NOHATE, which are the core of this work. Table 4 shows the size of said subset, along with the average sentence length for each class, their word counts and their vocabulary sizes.

Table 4
Size of the categories HATE and NOHATE in the clean dataset

	Hate	noHate
sentences	1,119	8,537
sentence length	20.39 ±9.46	15.15 ±9.16
word count	24,867	144,353
vocabulary	4,148	13,154

Regarding the distribution of sentences over Stormfront accounts, the dataset is balanced as

there is no account that contributes notably more than any other: the average percentage of sentences is of 0.50 ± 0.42 per account, the total amount of accounts in the dataset being 2,723. The sub-forums that contain more HATE belong to the category of news, discussion of views, politics, philosophy, as well as to specific countries (i.e., Ireland, Britain, and Canada). In contrast, the sub-forums that contain more NOHATE sentences are about education and homeschooling, gatherings, and youth issues.

In order to obtain a more qualitative insight of the dataset, a HATE score (H S) has been calculated based on the Pointwise Mutual Information (PMI) value for each word towards the categories HATE and NOHATE. PMI allows calculating the correlation of each word with respect to each category. The difference of the PMI value of a word *w* and the category H AT E and the PMI of the same word *w* and the category NOHATE results in the HATE score of *w*, as shown in Formula 1.

$$\text{H S}(w) = PMI\,(w, \text{H AT E}) - PMI\,(w, \text{N O HAT E}) \quad (1)$$

Intuitively, this score is a simple way of capturing whether the presence of a word in a HATE context occurs significantly more often than in a NOHATE context. Table 5 shows the 15 most and least hateful words: the more positive a HATE score, the more hateful a word, and vice versa.

The results show that the most hateful words are derogatory or refer to targeted groups of hate speech. On the other hand, the least hateful words are neutral in this regard and belong to the semantic fields of Internet, or temporal expressions, among others. This shows that the vocabulary is discernible by category, which in turn suggests that the annotation and guidelines are sound.

Performing the same calculation with bi-grams yields expressions such as "gene pool", "race traitor", and "white guilt" for the most hateful category, which appear to be concepts related to race issues. The less hateful terms are expressions such as "white power", "white nationalism" and "pro white", which clearly state the right-wing extremist politics of the forum users.

Table 5
Most (positive HS) and least (negative HS) hateful words

	H S		H S
ape	6.81	pm	−3.38
ape	6.81	pm	−3.38
scum	6.25	group	−3.34
savages	5.73	week	−3.13
filthy	5.73	idea	−2.70
mud	5.31	thread	−2.68
homosexuals	5.31	german	−2.67
filth	5.19	videos	−2.67
apes	5.05	night	−2.63
beasts	5.05	happy	−2.63
homosexual	5.05	join	−2.63
threat	5.05	pictures	−2.60
monkey	5.05	eyes	−2.54
libtard	5.05	french	−2.52
coon	5.05	information	−2.44
niglet	4.73	band	−2.44

Finally, the dataset has been contrasted against the English vocabulary in Hatebase. 9.28% of HATE vocabulary overlaps with Hatebase, a higher percentage than for NOHATE vocabulary, of which 6.57% of the words can be found in Hatebase. In Table 6, the distribution of HATE vocabulary is shown over Hatebase's 8 categories. Although some percentages are not high, all 8 categories are present in the corpus. Most of the HATE words from the dataset belong to ethnicity, followed by gender. This is in agreement with Silva *et al.* (2016), who conducted a study to analyse the targets of hate in social networks and showed that hate based on race was the most common.

Table 6
Distribution of HATE vocabulary over Hate-base categories

category	%	examples
archaic	2.46	div, wigger
ethnicity	41.63	coon, paki
nationality	7.03	guinea, leprechaun
religion	1.34	holohoax, prod
gender	36.05	bird, dyke
sexual orientation	2.34	fag, queer
disability	2.01	mongol, retarded
social class	7.14	slag, trash
total	100.00	

4. Experiments

In order to further inspect the resulting dataset whether the two annotated classes are separable based solely on the text of the labelled instances) a set of baseline experiments have been conducted. These experiments do not exploit any external resource such as lexicons, heuristics or rules. The experiments just use the provided dataset and well-known approaches from the literature to provide a baseline for further research and improvement in the future.

4.1. Experimental setting

The experiments are based on a balanced subset of labelled sentences. All the sentences labelled as HATE have been collected, and an equivalent number of NOHATE sentences have been randomly sampled, summing up 2k labelled sentences. From this amount, the 80% has been used for training and the remaining 20% for testing.

The evaluated algorithms are the following:

- Support Vector Machines (SVM) (Hearst *et al.*, 1998) over Bag-of-Words vectors. Word-count-based vectors have been computed and fed into a Python Scikit-learn LinearSVM[11] classifier to separate HATE and NOHATE instances.
- Convolutional Neural Networks (CNN), as described in (Kim, 2014). The implementation is a simplified version using a single input channel of randomly initialized word embeddings[12].
- Recurrent Neural Networks with Long Short-term Memories (LSTM) (Hochreiter and Schmidhuber, 1997). A LSTM layer of size 128 over word embeddings of size 300.

All the hyperparameters are left to the usual values reported in the literature (Greff *et al.*, 2017). No hyperparameter tuning has been performed. A more comprehensive experimentation and research has been left for future work.

4.2. Results

The baseline experiments include a majority class baseline showing the balance between the two classes in the test set. The results are given in terms of accuracy for HATE and NOHATE individually, and the overall accuracy, calculated according to the equations 2, 3 and 4, where *TP* are the true positives and *FP* are the false positives.

$$Acc_{\text{HATE}} = \frac{TP_{\text{HATE}}}{TP_{\text{HATE}} + FP_{\text{HATE}}} \quad (2)$$

$$Acc_{\text{NOHATE}} = \frac{TP_{\text{NOHATE}}}{TP_{\text{NOHATE}} + FP_{\text{NOHATE}}} \quad (3)$$

$$Acc_{\text{ALL}} = \frac{TP_{\text{ALL}}}{TP_{\text{ALL}} + FP_{\text{ALL}}} \quad (4)$$

We show the accuracy for the both complementary classes instead of the precision-recall of a single class to highlight the performance of the classifiers for the both classes individually. Table 7 shows the results of using only sentences that did *not* require additional context to be labelled, while Table 8 shows the results of including those sentences that required additional context. Not surprisingly, the results are lower when including sentences that required additional context. If a human annotator required additional information to make

[11] http://scikit-learn.org/stable/modules/svm.html

[12] https://github.com/dennybritz/cnn-text-classification-tf

a decision, it is to expect that an automatic classifier would not have enough information or would have a harder time making a correct prediction. The results also show that NOHATE sentences are more accurately classified than HATE sentences. Overall, the LSTM-based classifier obtains better results, but even the simple SVM using bag-of-words vectors is capable of discriminating the classes reasonably well.

Table 7
Results excluding sentences that required additional context for manual annotation

	Acc_{HATE}	Acc_{NOHATE}	Acc_{ALL}
Majority	n/a	n/a	0.50
SVM	0.72	0.76	0.74
CNN	0.54	0.86	0.70
LSTM	0.76	0.80	0.78

Table 8
Results including sentences that required additional context for manual annotation

	Acc_{HATE}	Acc_{NOHATE}	Acc_{ALL}
Majority	n/a	n/a	0.50
SVM	0.69	0.73	0.71
CNN	0.55	0.79	0.66
LSTM	0.71	0.75	0.73

4.3. Error Analysis

In order to get a deeper understanding of the performance of the classifiers trained, a manual inspection has been performed on a set of erroneously classified sentences. Two main types of errors have been identified:

Type I errors Sentences manually annotated as HATE but classified as NOHATE by the system, usually due to a lack of context or to a lack of the necessary world knowledge to understand the meaning of the sentence:

(11) "Indeed, now they just need to feed themselves, educate themselves, police themselves ad nauseum…"
(12) "If you search around you can probably find 'hoax of the 20th century' for free on the net."

In (11), it is not clear without additional context who "themselves" are. It actually refers to people of African origin. In its original context, the author was implying that they are not able to feed, police nor educate themselves. This would make the sentence an example of hate speech, but it could also be a harmless comment given the appropriate context. In (12), the lack of world knowledge about what the Holocaust is, or what naming it "hoax" implies —i.e., denying its existence—, would make it difficult to understand the sentence as an act of hate speech.

Type II errors Sentences manually labelled as NOHATE and automatically classified as HATE, usually due to the use of common offensive vocabulary with non-hateful intent:

(13) "I dont like reporting people but the last thing I will do is tolerate some stupid pig who claims Hungarians are Tartars."
(14) "More black-on-white crime: YouTube - Black Students Attack White Man For Eating Dinner With Black"

In (13), the user accuses and insults a particular individual. Example (14) is providing information on a reported crime. Although vocabulary of targeted groups is used in both cases (i.e., "Hungarians", "Tartars", "black"), the sentences do not contain HATE.

5. Discussion

There are several aspects of the introduced dataset, and hate speech annotation in general, that deserve a special remark and discussion.

First, the source of the content used to obtain the resulting dataset is on its own a source of offensive language. Being Stormfront a white supremacists' forum, almost every single comment contains

some sort of intrinsic racism and other hints of hate. However, not every expression that contains a racist cue can be directly taken as hate speech. This is a truly subjective debate related to topics such as free speech, tolerance and civics. That is one of the main reasons why this paper carefully describes the annotation criteria for what here counts as hate speech and what not. In any case, despite the efforts to make the annotation guidelines as clear, rational and comprehensive as possible, the annotation process has been admittedly demanding and far from straightforward.

In fact, the annotation guidelines were crafted in several steps, first paying attention to what the literature points about hate speech annotation. After a first round of manual labelling, inconsistencies among the human annotators were discussed and the guidelines and examples were adapted. From those debates we extract some conclusions and pose several open questions. The first annotation criteria (hate speech being a *deliberate attack*) still lacks robustness and a proper definition, becoming ambiguous and subject to different interpretations. A more precise definition of what an *attack* is and what it is not would be necessary: Can an objective fact that however undermines the honour of a group of people be considered an attack? Is the mere use of certain vocabulary (e.g. "nigger") automatically considered an attack? With regard to the second annotation criteria (hate speech being *directed towards a specific group of people*), it was controversial among the human annotators as well. Sentences were found that attacked individuals and mentioned the targets' skin colour or religion, political trends, and so on. Some annotators interpreted these as indirect attacks towards the collectivity of people that share the mentioned characteristics.

Another relevant point is the fact that the annotation granularity is sentence level. Most, if not all, of the existing datasets label full comments. A comment might be part of a more elaborated discourse, and not every part may convey hate. It is arguable whether a comment containing a single hate-sentence can be considered "hateful" or not. The dataset released provides the full set of sentences per comment with their annotations, so each can decide how to work with it.

In addition, and related to the last point, one of the labels included for the manual labelling is RELATION. This label is meant to be used when two or more sentences need each other to be understood as hate speech, usually because one is a premise and the following is the (hateful) conclusion. This label has been seldom used.

Finally, a very important issue to consider is the need of additional context to label a sentence (i.e., the rest of the conversation or the title of the forum-thread). It can happen to human annotators and, of course, to automatic classifiers, as confirmed in the error analysis (Section 4.3). Studying context dependency to perform the labelling, it has been observed that annotators learn to distinguish hate speech more easily over time, requiring less and less context to make the annotations (see Figure 1).

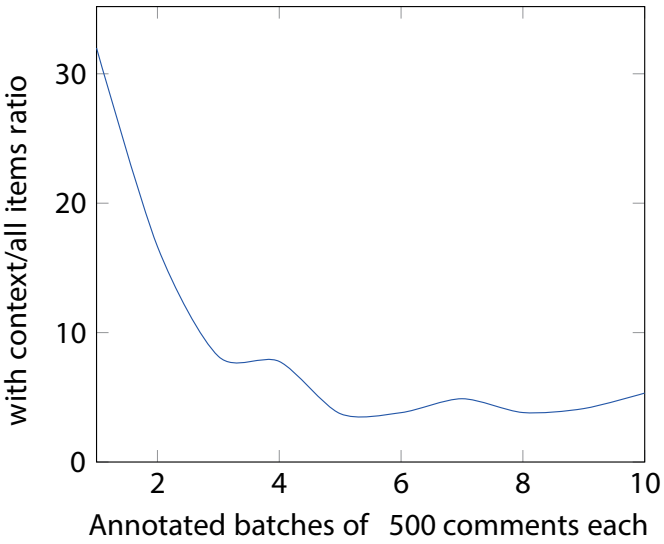

Figure 1. Percentage of comments per batch that required additional context to be manually labelled. The amount of context needed by a human annotator decreases over time

6. Conclusions and Future Work

This paper describes a manually labelled hate speech dataset obtained from Stormfront, a white supremacist online forum.

The resulting dataset consists of ~10k sentences labelled as conveying hate speech or not. Since the definition of hate speech has many subtleties, this work includes a detailed explanation of the manual annotation criteria and guidelines. Furthermore, several aspects of the resulting dataset have been studied, such as the necessity of additional context by the annotators to make a decision, or the distribution of the vocabulary used in the examples labelled as hate speech. In addition, several baseline experiments have been conducted using automatic classifiers, with a focus on examples that are difficult for automatic classifiers, such as those that required additional context or world knowledge. The resulting dataset is publicly available.

This dataset provides a good starting point for discussion and further research. As future work, it would be interesting to study how to include world knowledge and/or the context of an online conversation (i.e. previous and following messages, forum thread title, and so on) in order to obtain more robust hate speech automatic classifiers. Future studies could also explore how sentences labelled as RELATION affect classification, as this sentences have not been included in the experiments presented. In addition, more studies should be performed to characterize the content of the dataset in depth, regarding timelines, user behaviour and hate speech targets, for instance. Finally, since the proportion of HATE/NO-HATE examples tends to be unbalanced, a more sophisticated manually labelling system with active learning paradigms would greatly benefit future labelling efforts.

7. Acknowledgements

This work has been supported by the European Commission under the project ASGARD (700381, H2020-FCT-2015). We thank the Hate-base team, in particular Hatebase developer Timothy Quinn, for providing Hatebase's English vocabulary dump to conduct this study. Finally, we would like to thank the reviewers of the paper for their thorough work and valuable suggestions.

8. References

R. Agerri, J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3823-3828.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*, pages 759-760.

P. Burnap and M. L. Williams. 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.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37-46.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, pages 512-515.

F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86-95.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings, pages 29-30.

L. Figea, L. Kaati, and R. Scrivens. 2016. Measuring online affects in a white supremacy forum. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, pages 85-90.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

B. Gamba¨ck and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online (ALW1)*, pages 85-90.

N. D. Gitari, Z. Zuping, H. Damien, and J. Long. 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215-230.

K. Greff, R. K Srivastava, J. Koutn´ık, B. R. Steunebrink, and J. Schmidhuber. 2017. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222-2232.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18-28.

S. Hochreiter and J. Schmidhuber. 1997. Long Short-term Memory. *Neural Computation*, 9(8):1735-1780.

Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1746-1751.

I. Kwok and Y. Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, pages 1621-1622.

S. Malmasi and M. Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, pages 467-472.

S. Malmasi and M. Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187-202.

P. M. Meddaugh and J. Kay. 2009. Hate Speech or "Reasonable Racism?" The Other in Stormfront. *Journal of Mass Media Ethics*, 24(4):251-268.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*, pages 145-153.

J. T. Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution*, 3:1277-79.

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17, pages 6-9.

H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths. 2016. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016)*.

Joseph A Schafer. 2002. Spinning the web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture*, 9(2):69-88.

A. Schmidt and M. Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (SocialNLP 2017)*, pages 1-10.

J. Serra, I. Leontiadis, D. Spathis, J. Blackburn, G. Stringhini, and A. Vakali. 2017. ¡Class-based prediction errors to detect hate speech with out-of- vocabulary words. In *Abusive Language Workshop*, volume 1. Abusive Language Workshop.

S. Sharma, S. Agrawal, and M. Shrivastava. 2018. Degree based Classification of Harmful Speech using Twitter Data. *arXiv:1806.04197*.

L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pages 687-690.

W. Warner and J. Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media (LSM 2012)*, pages 19-26.

Z. Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. pages 138-142.

Z. Waseem and D. Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop (NAACL SRW 2018)*, pages 88-93.

H. Watanabe, M. Bouazizi, and T. Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access*, 6:13825-13835.