Directed Digital Hate

Björn Pelzer, Lisa Kaati
Swedish Defence Research Agency
FOI
Stockholm, Sweden
Email: firstname.lastname@foi.se

Nazar Akrami
Uppsala University
Uppsala, Sweden
Email: nazar.akrami@psyk.uu.se

***Abstract*—Politicians, journalists and public figures are often exposed to hate speech in various digital environments, especially in different discussion forums and in comment fields. Detecting hate automatically is a difficult task since hate can be expressed in many different ways. In this paper we have developed a method to measure hate directed at politicians using a combination of natural language processing and automated reasoning. Our method is adapted to work on Swedish, although in general it is language independent. We have tested our method in a study where we analyze hate directed at six Swedish politicians. The results shows that our method has a fairly high precision but a low recall compared to a manual assessment.**

I. Introduction

A fundamental element of democracy is the citizens' right to express their views and discuss issues relating to their political interests. This right also follows a responsibility that rests on respect for the opinions of others and the right to comment. In traditional forums where political discussions have been held, social norms have served as the basis for this respect. With the increased use of social media this has changed. Flexibility, promptness and anonymity are just a few examples of factors that have contributed to this change. Today, anyone, at any time, can comment on anything they want and in any way they want, completely anonymously. It has been shown that this not only contributes to greater civil participation in the political debate but also increases the vulnerability for, among others, elected political representatives. Several studies show that in many cases elected politicians receive hateful comments and in some cases even threats [6][1].

In this work we describe a method to measure hate directed at politicians using a combination of natural language processing and automated reasoning. Our method is adapted to work on Swedish, but its basic principles are language independent. We have tested our method in a study where we analyze hate speech directed at six Swedish politicians. In our work we focus on comments in discussion forums and unmoderated commentary fields. We study hateful comments that are directed at an individual, but which are not necessarily meant for the individual to read. We chose to categorize a comment as hate speech if the comment would not be allowed to be expressed in a debate, a direct conversation, or public setting like a television program.

The planned use of our technique is hate speech detection in relation to research on prevention strategies of hate speech. The findings are also implemented in our other project on the antecedents of prejudice and hate - in this context we study the relation between personality and social psychological variables explaining prejudice and hate (or simple disliking). The ultimate research goal is to identify why people express prejudice and hate and provide means prevention. After further improvements to our tool, a practical application can be envisioned in threat assessment by law enforcement agencies, for example in the context of personal security.

This paper is outlined as follows. In Section II we describe our interpretation of hate speech, and explain how we distinguish between three categories of hate comments. In Section III we present our method for measuring directed hate in digital environments. In Section IV we evaluate our method by comparing it to two other methods for measuring directed hate in digital environments. A discussion of the results is presented in Section V and finally, some directions for future research are presented in Section VI

II. Hate

In the psychological literature, hate is regarded as an emotional state. Hate is considered to be consisting of one or more emotions such as anger, fear, contempt or disgust. The intensity of these emotions and the individual's degree of dedication determine the outcome, or the behavior of the individual, which is the direct result of the emotions. The outcome behavior varies between avoiding, punishing or in the most serious form trying to destroy the target of hate [15]. An individual's level of hate is considered to be relatively stable over time, and it is maintained by expressing hate towards one or multiple target individuals or groups.

The expression of hate varies across situations, and hate speech can be a result of prejudices against groups or individuals who (seem to) share a specific trait. Thus, hate and prejudice share some common component, but hate consists entirely of emotions, while prejudice is only partly based on them. In this work we have chosen to study hate speech as expressed in three different categories of offensive comments: *directed at an individual*, *directed at women*, and *directed at minorities*. The latter two were chosen as they represent hate speech based on a trait (gender/minority), to give us a better understanding of the overall composition of digital hate.

978-1-5386-7848-0/18/$31.00 ©2018 IEEE

TABLE I
DIFFERENT CATEGORIES OF HATE WITH SOME EXAMPLE WORDS.

Category	**Size (# terms)**	**Sample terms (ENG)**	**Sample terms (SWE)**
Offensive to individual	356	stupid, ugly, idiot	cepe, ful, idiot
Offensive to women	41	whore, bitch	hora, subban, kärring
Offensive to minority	40	jew, nigger, assboy	jude, neger, stjärtpojke

Detecting and measuring digital hate automatically is a difficult task. Several approaches have been developed and tested. Hate speech can be directed at an individual (called targeted or directed) or it can be directed at a group or class of people (generalized). In [4] generalized hate and hate directed at individuals or entities is studied. Directed hate is defined as hate language towards a specific individual or entity while generalized hate is defined as hate language towards a general group of individuals who share a common protected characteristic, e.g., ethnicity or sexual orientation.

There have been several approaches to detect generalized hate speech. In [10] machine learning is used to separate between hate speech, profanity, and other texts. The dataset that is used is an annotated dataset of tweets.

Another example is Google's Perspective API, which is built for detecting toxic comments online. However, research shows that detecting hate speech is a difficult task and that Perspective can easily be deceived [8]. The majority of studies on hate speech have used English data.

In our work, each of the three different categories of hate is represented by a dictionary of terms, a so-called hate category dictionary. Examples of the terms are listed in Table I. Our study focuses on Swedish data, but to ease understanding we have translated some of the words into English.

There are many limits using dictionary-based approaches. For example, it is not possible to detect linguistic nuances such as irony or sarcasm. Another drawback of using dictionaries is that words are context dependent. One word can have very different meanings, depending on how and where it is used. Each dictionary is constructed by human experts (psychologists and computer scientists). In order to improve the coverage of the dictionaries, a word embedding trained on the environments we study is used to suggest complementary terms to the experts. This is done by computing the 15 nearest neighbors in the embedding space to each term in the dictionaries. For each term suggestion, the expert has the choice to either include or reject the term suggestion.

III. DETECTING AND MEASURING DIGITAL HATE

When detecting expressions of hate a simple and straight forward approach is to search texts for expressions from the aforementioned hate category dictionaries occurring in conjunction with the names of politicians. In this approach a sliding window registers hate expressions in a text only when they occur within a certain distance from the name of a target person, i.e. at most *n* words before or after the name. This approach was used in [9]. A simple variation on this is a sentence-based sliding window, i.e. the window always covers one whole sentence, sliding over the sentences one by one, and registers hate if the current sentence contains both a hate expression and a target person name. However, there are several limitations to the described approaches, such as:

- No accounting for coreference: the target person name has to match exactly, and pronouns or titles (*"the prime minister"*) are insufficient even when the context makes clear that the text is about the target person.
- No handling of morphology: words from the hate categories have to match exactly, and given that Swedish is more highly inflected than English this means the dictionaries only cover a fraction of the actually occurring word-form variants.
- Distance is too simplistic: hate expressions can be directed against a name despite considerable word distance in between, but large windows with good coverage result in poor precision. Furthermore, even closer hate expressions may not actually express hate against the target person, for example because they are negated by the context, such as in: *"No way president Smith is stupid!"*

To overcome some of these limitations, we utilize a more elaborate processing where we combine methods from natural language processing (NLP) and automated reasoning (AR). Since Swedish does not yet have large annotated corpora on a scale comparable to English, we opted for more traditional AI techniques rather than deep learning techniques, as the latter would require extensive training data. However, we acknowledge that deep learning techniques have proven to work well in similar contexts [2].

The implementation was done in *Python* 3.5[1], utilizing several third-party modules. It operates on texts retrieved from the web and stored in a *MongoDB*[2] database. The input is:

- the hate category dictionaries, where optionally each verb expression can be annotated with a restriction indicating that the hate or threat is directed only against the agent or only against the patient of the verb,
- specifications of the text sources to analyze, i.e. the specific database collections and optionally some time interval, for example to investigate only comments and articles published within a certain date range,
- a list of target persons, where for each target person one can specify multiple names, alternative spellings, common misspellings, titles and nicknames, and optionally the natural gender.

[1] http://www.python.org
[2] https://www.mongodb.com

Texts from the specified database that contain the names and nicknames are retrieved. Titles are not used as retrieval criteria, as offices may be held by different people over time.

A Swedish Part-of-Speech tagger from the *efselab* toolset[3] prepares the text for the subsequent dependency-parsing with the efselab-configured version of *MaltParser* [11], which in turn results in parse trees in the *CoNLL-U* Universal Dependencies format,[4] with one CoNLL-U node per word or punctuation symbol.

The parse trees are then processed in a deductive analysis that aims at identifying which hate expressions are actually directed against the target persons. For this analysis the CoNLL-U representation of a text is asserted into a specialized reasoner that we implemented in *PyKE*[5], a Prolog-inspired [3] logic programming language that integrates with Python. Generally, each CoNLL-U node is asserted as a fact, resulting in one fact per word, with the following modifications:

- Multiple successive nodes representing a single person name (e.g. one node for the first name followed by one for the surname) are merged into a single node, so that each mentioning of a person is represented by one node.
- Any node representing a name, nickname or title of a target person is marked as a target person node.
- Any person name node not corresponding to a name or nickname of a target person is marked as a non-target person node, to aid in disambiguation.
- All person nodes are marked with their natural gender, when possible, using the gender information from the target input specification (when available), or a heuristic based on based on a list of 350,000 first names.
- Any node representing a word-form of a lemma from the hate category dictionaries is marked as a hate expression node, and it is annotated with the lemma and its category. A node may have multiple hate categories. If a node corresponds to a word that is part of a multi-word phrase in a hate category dictionary, it will be marked as representing that phrase only if the parse tree of the current sentence also contains nodes for the remaining words of the phrase and all those nodes are dependent on each other (i.e. indicating that the sentence actually contains the phrase, not just the individual words of the phrase in some other arrangement without the meaning of the phrase).

After these assertions, the deductive analysis consists of two phases, each making use both of the structure of the parse tree and the sequential order of words:

1) A forward-chaining phase focuses on coreference resolution. The reasoner exhaustively attempts to identify each personal pronoun node as a person node (target or non-target) by associating it with some person node of matching gender already identified in the text, i.e. either a node representing a person name, or a personal pronoun node whose coreference has already been resolved in an earlier inference.
2) A backward-chaining phase then tries to find non-negated hate expression nodes directed against target person nodes, based on dependencies (such as a hate adjective node depending on a target person node), and taking into account any agent/patient restrictions of hateful verb expressions.

The results are presented as tables suitable for CSV-files.

Using methods of AR in text analysis can face performance issues, as identified in [7]. In the case of the our work the situation is benign: Given the focus on only identifying small aspects of the meaning of the texts, rather than trying to get a full understanding, no attempt is made to represent the full semantics in logic, nor do we use any extensive background knowledge. Instead we operate directly on the parse trees, with a decidable, effectively propositional logic fragment, i.e. Bernays-Schnfinkel Horn with ground units. Thus in practice the deductive analysis has negligible impact on the processing time. More effort needed to be spent on optimizing the earlier stages in the processing, such as parsing multiple texts combined into one in order to reduce the MaltParser's multi-second startup overhead that would make individual text parsing prohibitive for large document collections.

IV. Evaluation

A. Hate towards Swedish politicians

In Sweden, each of the 290 municipalities has a municipal council, which is composed of representatives from different parties and chaired by an elected chairman. In this section we present the results of a study of the presence of hateful comments against six such chairmen in digital environments. The purpose of this is to test our method for detecting hate but also to gain a better understanding of how hatred directed at municipal politicians looks and to what extent it occurs.

B. Data

In this study we have analyzed a number of different Swedish digital environments, both discussion forums and alternative media as well as associated commentary fields. The choice of these environments is motivated by the fact that they differ from traditional media and forums in their free language and that in most cases they have more or less unmoderated comment fields - that is, those who comment are themselves legally responsible for the content of the comments and comments are not considered to be editorial material.

The study contains texts published beginning in 2014, to cover one term of office (4 years). In total, the analyzed text volume consists of approximately 41,000 articles and 4.8 million comments and posts. The digital environments included in our study are as follows:

- *Avpixlat* (www.avpixlat.info) - 13,200 articles, 3.2 million comments (closed since August 2017).
- *Samhällsnytt* (www.samnytt.se) - 1,800 articles, 554,000 comments (successor to Avpixlat, online since September 2017).

[3]https://github.com/robertostling/efselab
[4]http://universaldependencies.org
[5]*Python Knowledge Engine*, http://pyke.sourceforge.net/

TABLE II
DATASET COMMENTS MENTIONING THE POLITICIANS IN OUR STUDY

Individual	**Number of mentions**
Female 1	269
Female 2	255
Male 1	153
Male 2	59
Female 3	48
Male 3	25
Total number of comments	809

- Nordic Resistance Movement (www.nordfront.se) - 4,400 articles, 113,000 comments.
- *Fria tider* (www.friatider.se) - 16,600 articles.
- Motgift (www.motgift.nu) - 630 articles, 1,400 comments, 14,000 forum posts.
- Nordic Youth (www.nordiskungdom.com) - 600 articles, 200 comments.
- *Nyheter idag* (www.nyheteridag.se) - 3,900 articles, 16,100 comments.
- *Flashback*: Politics / Domestic (www.flashback.se/f77) - 840,000 posts.

We only study open forums where it is assumed that those who choose to publish text agree that the text will be made available to the public. No data has been retrieved from password-protected pages, private pages on Facebook, or other types of web pages or social media where the user has taken actions to keep posted material within a closed circle.

Table II shows how often the six elected municipal council chairmen that we have studied are mentioned in the dataset. In total, the municipal council chairmen are mentioned 809 times. Due to privacy reasons, we have chosen to not reveal the names of the politicians.

C. Manual assessment of directed hate

To have something to compare our method with we have conducted a manual analysis of the posts that we will use as ground truth. In the manual analysis we used two experts that analyzed all comments where the name of the target person was mentioned. The comments where sorted into the three different categories of hate that was previously defined. One thing that we noticed when manually studying the comments was the that variation of expressions of hate was enormous. The majority of the comments to both men and women were offensive to individuals. This includes mean comments about appearance and intelligence, for example sayings such as:

"do not have all the horses in the stable"

"incompetent hypocrite"

"stupid clown"

It is also apparent that only women receive offensive comments directed at minority groups. These comments can be, for example, to call someone a "witch" or to say sexist comments, as well as comments that reflect on the ethnicity of the individual or the individual's relatives. None of the hateful comments were about the political message conveyed or conveyed a political stance.

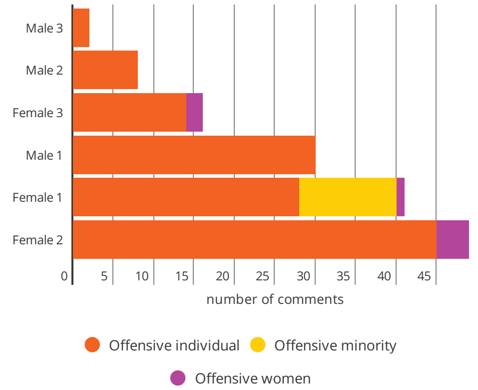

Fig. 1. Hateful comments directed at six politicians using manual assessment.

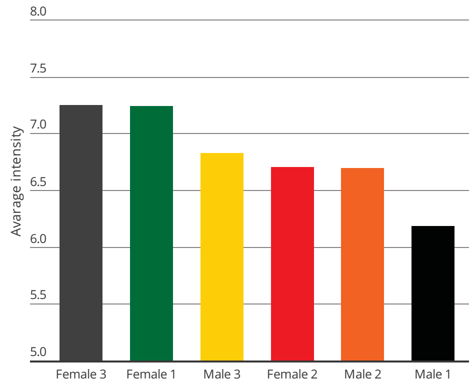

Fig. 2. Average intensity of hateful comments directed at the six politicians.

Figure 1 shows the results of the manual analysis of the comments directed at the six politicians that we studied. The results shows that the share of hateful comments varied between 8 and 35 percent for the politicians in our study. Looking at the comments directed at the males, these vary between 8 and 20 percent, while the female rate varies between 15 and 35 percent of the total comments. In summary, these results show that the percentage of hate comments is around 20 percent for politicians and that women receive a higher proportion of hate comments than men. This means that on average, about 1 in 5 comments that is written about a politician in our study is an offensive comment.

To get an understanding of the intensity of the hateful comments in our study, we assessed how hateful each individual comment is. We did this by asking two reviewers who independently assessed the intensity of the comments (i.e. how mean the comments were) on a scale of 0 (not that mean) to 10

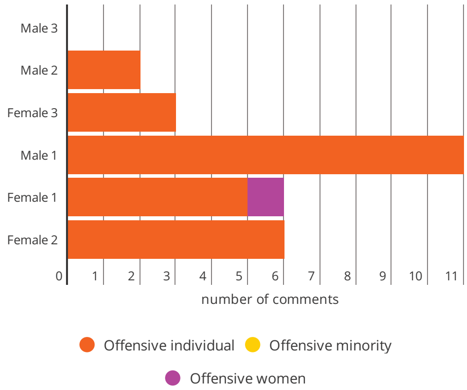

Fig. 3. Hateful comments directed at the six politicians using NLP+AR.

(very mean).[6] The results showed that the reviewers had a high level of consistency as their estimates overlap to 82 percent. We then summarized these assessments by taking the mean of each comment's assessment, which became our measure of the intensity. The average intensity of all comments is 6.8 (on a scale 0-10). The intensity for the comments directed at each politician in our study is shown in Figure 2.

D. Using NLP and AR to Measure Directed Hate

To test our method to analyze directed hate we ran our implementation on the dataset consisting of 4.8 million posts. The result can be seen in Figure 3. As can be seen, the results using our method differs significantly from the manual assessment shown in Figure 1.

TABLE III
HATEFUL COMMENTS ABOUT THE POLITICIANS

Method	Number of hateful comments	Incorrectly classified comments
Our method (NLP + AR)	28	2 (7%)
Sentence-based method (sliding window)	71	14 (20%)
Manual assessment	146	-

E. Comparison with other methods

To get an understanding of how well our technique works we compared our NLP+AR method with the earlier sentence-based sliding-window method that looks for name and an offensive word in the same sentence. The results of these two methods and the manual assessment are shown in Table III. Using our NLP+AR method, 28 unique hateful comments directed at the six politicians were detected. Out of these 28 comments 2 were false positives. This means that around 7 percent of the comments where incorrectly detected as hate directed at any of the politicians.

Using the sentence-based sliding-window method we detected 71 unique comments. Out of these 71 comments, 14 comments where incorrectly classified as directed hate. This indicates that almost 20 percent of the comments where incorrectly classified. The manual assessment detected 146 hateful comments directed at the politicians in our study.

[6]We chose this more fine-tuned scale over the common 5-level scale to allow more variation, and to be able to make a better differentiation between different levels of hate.

V. Discussion

Automatic detection only finds a small part of the hateful comments directed at the six politicians in our study. One of the reasons for this is that the hate category dictionaries do not capture all hate expressions. Hate can be expressed in so many ways - something that our human experts noticed during the manual assessment.

The evaluation of the three methods provided different outcomes. The automatic methods (sentence-based and our NLP+AR method) can be seen as two different instruments - one with high coverage and low precision and one with high precision and low coverage. Comparing these approaches to the human interpretation underlines that automatic hate speech detection is a challenging task.

Both automatic approaches are based on dictionaries. If a hate expression is not included in the dictionary, it will not be detected. Unfortunately our manual assessment indicates that authors of hateful comments like to be creative in coining new slurs, requiring ever growing dictionaries to keep up with a large number of terms that each only occur very rarely. Moreover, in practice hate is also often expressed via negation of positive terms (e.g. *"he is not exactly smart"*). Covering this would require extending the dictionaries with negated phrases, or an additional dictionary of positive terms, whose negated occurrence then could be identified as hate.

To gain a better understanding of the impact of dictionary limitations, we extracted 17,176 comments containing named entities (persons) from our sources above, regardless of whether these were politicians, and manually annotated every occurrence of hate directed against a named person. 27 percent of the comments were manually flagged as hateful. We then tested our NLP+AR method on this set, always using the flagged person as a target. The system identified only 3 percent of the total set as hateful, or 11 percent of the manually flagged comments. However, 71 percent of the manually flagged comments do not contain any words from the hate category dictionaries, so these are out of reach. Conversely, our system did correctly identify 39 percent of those hate comments that are covered by the dictionaries. Thus, optimal dictionaries could in theory almost quadruple the detection rate. Our system also identified 1.5 percent of the non-flagged comments as hateful. This is a low rate of false positives, and many of these cases could even be regarded as ambiguous rather than clearly false. This rate would likely increase with larger dictionaries, but it is good starting point.

A different way to overcome the problems might consist of annotating data and using machine learning (ML). However, this requires human experts to annotate large amounts of data - our annotation set mentioned above is far from sufficient. Also, the resulting model will be language dependent, requiring a new annotation for each new language, whereas our NLP+AR approach is more easily adapted to other languages by modifying the underlying dependency recognition rules. Furthermore, while detecting the general sentiment of a text using ML is fairly well-understood, correctly identifying whether a sentiment, such as hate, is or is not directed at a specific target - i.e. a variant of aspect-based sentiment analysis - remains more elusive [13][16]. Yet this is a critical problem when dealing with user comments on political issues: Unlike in the traditional experimentation data of sentiment analysis, such as product and hotel reviews, the users commenting on politics are likely to juxtapose their favored politicians with their opponents within very limited amounts of text, expressing strongly positive and negative feelings about each respectively. Not mixing up who is hated and who is praised is essential to detecting directed hate against a specific target.

VI. Conclusions and Future Work

In this paper we have studied hate directed at six selected municipal council chairmen in Sweden. These individuals are not national-level politicians and therefore not well-known to the public in Sweden. Regardless of this, hateful comments about these individuals were rather common in Swedish digital environments. Studying hate in digital environments is an important task, as such hate may have damaging impacts on democracy. Studies show that frequent exposure to hateful comments leads to increased prejudice [14] and also to an increased dehumanization of minority groups [5]. In turn, dehumanization of minority groups (or individuals) legitimizes and increases the likelihood of violence [12], which is why law enforcement needs to include a digital component in their threat assessment of public figures, politicians and journalists.

For future work we intend to combine a number of different methods for detecting directed hate. We will use the two methods described in this paper (sentence-based and NLP+AR) and complement them with machine learning approaches:

One ML track will train a model for stand-alone detection of directed hate, working directly on the texts, and then compare its results with those of the other approaches.

Another ML track will be a hybrid of machine learning and our NLP+AR approach. While the parsing phase and the coreference analysis are robust and provide valuable data about the inspected text, the identification of hate expressions and the final dependency analysis require the manual creation of extensive dictionaries and rule sets. ML might supplant these steps, and by operating on the parser results after the coreference resolution it would have more semantics data to work with than when dealing with only the raw text.

We are also investigating means of automatically measuring the intensity of hate in the comments, similar to the manual assessment that was done in this work.

Finally, we also consider the ability to include include human assessments of hateful comments. Since detecting and analyzing hate directed at individuals is a challenging task, different methods that can be seen as different instruments can be combined to provide a better assessment of hate directed at individuals. The different methods have their pros and cons in terms of accuracy, recall and precision, and hopefully by combining them a more reliable assessment can obtained.

We believe anything approaching optimal precision and recall to be well out of reach at the moment. But depending on the application, a useful tool may already be feasible. When the objective is to observe hate trends over time, the absolute amount of false positives and negatives is less relevant, as long as their ratio does not change. On the other hand, in situations where a human analyst studies every comment flagged by the tool, minimizing false positives is essential.

Acknowledgment

This work was funded by a commission from the Swedish government and by a project financed by the Swedish Association of Local Authorities and Regions (SKL).

References

[1] *Sexism, Harassment and Violence Against Women Parliamentarians*. Inter-Parliamentary Union, 2016.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, 2017.

[3] A. Colmerauer and P. Roussel. History of programming languages—ii. chapter The Birth of Prolog, pages 331–367. ACM, New York, NY, USA, 1996.

[4] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. M. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. *CoRR*, abs/1804.04257, 2018.

[5] F. Fasoli, M. Paladino, A. Carnaghi, J. Jetten, B. Bastian, and P. Bain. Not "just words": Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology*, 46:237–248, 2016.

[6] A. Frenzel. *The Politicians Safety Survey 2017*. The Swedish National Council for Crime Prevention, 2017.

[7] U. Furbach, I. Glöckner, and B. Pelzer. An application of automated reasoning in natural language question answering. *AI Commun.*, 23(2-3):241–265, 2010.

[8] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective API built for detecting toxic comments. *CoRR*, abs/1702.08138, 2017.

[9] T. Isbister, M. Sahlgren, L. Kaati, M. Obaidi, and N. Akrami. Monitoring Targeted Hate in Online Environments. *Second workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*, 2018.

[10] S. Malmasi and M. Zampieri. Detecting hate speech in social media. *CoRR*, abs/1712.06427, 2017.

[11] J. Nivre, J. Hall, and J. Nilsson. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, pages 2216–2219, 2006.

[12] T. Rai, P. Valdesolo, and J. Graham. Dehumanization increases instrumental violence, but not moral violence. *Pnas*, 114(32):8511–8516, 2017.

[13] K. Schouten and F. Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge & Data Engineering*, 28(3):813–830, March 2016.

[14] W. Soral, M. Bilewicz, and M. Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive Behaviour*, 2017.

[15] R. J. Sternberg and K. Sternberg. *The Nature of Hate*. Cambridge University Press, 2008.

[16] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(4), 2018.