Detecting Hate Speech on the World Wide Web

William Warner and Julia Hirschberg
Columbia University
Department of Computer Science
New York, NY 10027
whw2108@columbia.edu, julia@cs.columbia.edu

Abstract

We present an approach to detecting *hate speech* in online text, where hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation. While hate speech against any group may exhibit some common characteristics, we have observed that hatred against each different group is typically characterized by the use of a small set of high frequency stereotypical words; however, such words may be used in either a positive or a negative sense, making our task similar to that of words sense disambiguation. In this paper we describe our definition of hate speech, the collection and annotation of our hate speech corpus, and a mechanism for detecting some commonly used methods of evading common "dirty word" filters. We describe pilot classification experiments in which we classify anti-semitic speech reaching an accuracy 94%, precision of 68% and recall at 60%, for an F1 measure of .6375.

1 Introduction

Hate speech is a particular form of offensive language that makes use of stereotypes to express an ideology of hate. Nockleby (Nockleby, 2000) defines hate speech as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic." In the United States, most hate speech is protected by the First Amendment of the U. S. Constitution, which, except for obscenity, "fighting words" and incitement, guarantees the right to free speech, and internet commentators exercise this right in online forums such as blogs, newsgroups, Twitter and Facebook. However, *terms of service* for such hosted services typically prohibit hate speech. Yahoo! Terms Of Service [1] prohibits posting "Content that is unlawful, harmful, threatening, abusive, harassing, tortuous, defamatory, vulgar, obscene, libelous, invasive of another's privacy, hateful, or racially, ethnically or otherwise objectionable." Facebook's terms [2] are similar, forbidding "content that: is hateful, threatening, or pornographic; incites violence." While user submissions are typically filtered for a fixed list of offensive words, no publicly available automatic classifier currently exists to identify hate speech itself.

In this paper we describe the small amount of existing literature relevant to our topic in Section 2. In Section 3 we motivate our working definition of hate speech. In Section 4 we describe the resources and corpora of hate and non-hate speech we have used in our experiments. In Section 5 we describe the annotation scheme we have developed and interlabeler reliability of the labeling process. In Section 6 we describe our approach to the classification problem and the features we used. We present preliminary results in Section 7, follow with an analysis of classification errors in 8 and conclude in Section 9 with an outline of further work.

[1] Yahoo TOS, paragraph 9a http://info.yahoo.com/legal/us/yahoo/utos/utos-173.html

[2] Facebook TOS, paragraph 3.7 https://www.facebook.com/legal/terms

Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), pages 19–26,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

2 Previous Literature

There is little previous literature on identifying hate speech.

In (A Razavi, Diana Inkpen, Sasha Uritsky, Stan Matwin, 2010), the authors look for Internet "flames" in newsgroup messages using a three-stage classifier. The language of flames is significantly different from hate speech, but their method could inform our work. Their primary contribution is a dictionary of 2700 hand-labeled words and phrases.

In (Xu and Zhu, 2010), the authors look for offensive language in YouTube comments and replaces all but the first letter of each word with asterisks. Again, while the language and the goal is different, the method may have some value for detecting hate speech. Their detection method parses the text and arranges it into a hierarchy of clauses, phrases and individual words. Both the annotation and the classification strategies found in this paper are based on the sentiment analysis work found in (Pang and Lee, 2008) and (Pang, Lee and Vaithyanathan, 2002).

3 Defining Hate Speech

There are numerous issues involved in defining what constitutes hate speech, which need to be resolved in order to annotate a corpus and develop a consistent language model. First, merely mentioning, or even praising, an organization associated with hate crimes does not by itself constitute hate speech. The name "Ku Klux Klan" by itself is not hateful, as it may appear in historical articles, legal documents, or other legitimate communication. Even an endorsement of the organization does not constitute a verbal attack on another group. While one may hypothesize that such endorsements are made by authors who would also be comfortable with hateful language, by themselves, we do not consider these statements to be hate speech.

For the same reason, an author's excessive pride in his own race or group doesn't constitute hate speech. While such boasting may seem offensive and likely to co-occur with hateful language, a disparagement of others is required to satisfy the definition.

For example, the following sentence does not constitute hate speech, even though it uses the word "Aryan".

> *And then Aryan pride will be true because humility will come easily to Aryans who will all by then have tasted death.*

On the other hand, we believe that unnecessary labeling of an individual as belonging to a group often should be categorized as hate speech. In the following example, hate is conveyed when the author unnecessarily modifies bankers and workers with "jew" and "white."

> *The next new item is a bumper sticker that reads: "Jew Bankers Get Bailouts, White Workers Get Jewed!" These are only 10 cents each and require a minimum of a $5.00 order*

Unnecessarily calling attention to the race or ethnicity of an individual appears to be a way for an author to invoke a well known, disparaging stereotype.

While disparaging terms and racial epithets when used with the intent to harm always constitute hateful language, there are some contexts in which such terms are acceptable. For example, such words might be acceptable in a discussion of the words themselves. For example:

> *Kike is a word often used when trying to offend a jew.*

Sometimes such words are used by a speaker who belongs to the targeted group, and these may be hard to classify without that knowledge. For example:

> *Shit still happenin and no one is hearin about it, but niggas livin it everyday.*

African American authors appear to use the "N" word with a particular variant spelling, replacing "er" with "a", to indicate group solidarity (Stephens-Davidowitz, 2011). Such uses must be distinguished from hate speech mentions. For our purposes, if the identity of the speaker cannot be ascertained, and if no orthographic or other contextual cues are present, such terms are categorized as hateful.

4 Resources and Corpora

We received data from Yahoo! and the American Jewish Congress (AJC) to conduct our research on hate speech. Yahoo! provided data from its news group posts that readers had found offensive. The AJC provided pointers to websites identified as offensive.

Through our partnership with the American Jewish Congress, we received a list of 452 URLs previously obtained from Josh Attenberg (Attenberg and Provost, 2010) which were originally collected to classify websites that advertisers might find unsuitable. After downloading and examining the text from these sites, we found a significant number that contained hate speech according to our working definition; in particular, a significant number were anti-semitic. We noted, however, that sites which which appeared to be anti-semitic rarely contained explicitly pejorative terms. Instead, they presented scientifically worded essays presenting extremely anti-semitic ideologies and conclusions. Some texts contained frequent references to a well known hate group, but did not themselves constitute examples of hate speech. There were also examples containing only defensive statements or declarations of pride, rather than attacks directed toward a specific group.

In addition to the data we collected from these URLs, Yahoo! provided us with several thousand comments from Yahoo! groups that had been flagged by readers as offensive, and subsequently purged by administrators. These comments are short, with an average of length of 31 words, and lacked the contextual setting in which they were originally found. Often, these purged comments contained one or more offensive words, but obscured with an intentional misspelling, presumably to evade a filter employed by the site. For common racial epithets, often a single character substitution was used, as in “nagger”, or a homophone was employed, such as “joo.” Often an expanded spelling was employed, in which each character was separated by a space or punctuation mark, so that “jew” would become “j@e@w@.”

The two sources of data were quite different, but complementary.

The Yahoo! Comment data contained many examples of offensive language that was sometimes hateful and sometimes not, leading to our hypothesis that hate speech resembles a word sense disambiguation task, since, a single word may appear quite frequently in hate and non-speech texts. An example is the word “jew”. In addition, it provided useful examples of techniques used to evade simple lexical filters (in case such exist for a particular forum). Such evasive behavior generally constitutes a positive indicator of offensive speech.

Web data captured from Attenberg’s URLs tended to include longer texts, giving us more context, and contained additional lower frequency offensive terms. After examining this corpus, we decided to attempt our first classification experiments at the paragraph level, to make use of contextual features.

The data sets we received were considered offensive, but neither was labeled for hate speech per se. So we developed a labeling manual for annotating hate speech and asked annotators to label a corpus drawn from the web data set.

5 Corpus Collection and Annotation

We hypothesize that hate speech often employs well known stereotypes to disparage an individual or group. With that assumption, we may be further subdivide such speech by stereotype, and we can distinguish one form of hate speech from another by identifying the stereotype in the text. Each stereotype has a language all its own, with one-word epithets, phrases, concepts, metaphors and juxtapositions that convey hateful intent. Anti-hispanic speech might make reference to border crossing or legal identification. Anti-African American speech often references unemployment or single parent upbringing. And anti-semitic language often refers to money, banking and media.

Given this, we find that creating a language model for each stereotype is a necessary prerequisite for building a model for all hate speech. We decided to begin by building a classifier for anti-semitic speech, which is rich with references to well known stereotypes.

The use of stereotypes also means that some language may be regarded as hateful even though no single word in the passage is hateful by itself. Often there is a relationship between two or more sentences that show the hateful intent of the author.

Using the website data, we captured paragraphs that matched a general regular expression of words relating to Judaism and Israel [3]. This resulted in about 9,000 paragraphs. Of those, we rejected those that did not contain a complete sentence, contained more than two unicode characters in a row, were only one word long or longer than 64 words.

Next we identified seven categories to which labelers would assign each paragraph. Annotators could label a paragraph as anti-semitic, anti-black, anti-asian, anti-woman, anti-muslim, anti-immigrant or other-hate. These categories were designed for annotation along the anti-semitic/not anti-semitic axis, with the identification of other stereotypes capturing mutual information between anti-semitism and other hate speech. We were interested in the correlation of anti-semitism with other stereotypes. The categories we chose reflect the content we encountered in the paragraphs that matched the regular expression.

We created a simple interface to allow labelers to assign one or more of the seven labels to each paragraph. We instructed the labelers to lump together South Asia, Southeast Asia, China and the rest of Asia into the category of anti-asian. The anti-immigrant category was used to label xenophobic speech in Europe and the United States. Other-hate was most often used for anti-gay and anti-white speech, whose frequency did not warrant categories of their own.

5.1 Interlabeler Agreement and Labeling Quality

We examined interlabeler agreement only for the anti-semitic vs. other distinction. We had a set of 1000 paragraphs labeled by three different annotators. The Fleiss kappa interlabeler agreement for anti-semitic paragraphs vs. other was 0.63. We created two corpora from this same set of 1000 paragraphs. First, the *majority* corpus was generated from the three labeled sets by selecting the label with on which the majority agreed. Upon examining this corpus with the annotators, we found some cases in which annotators had agreed upon labels that seemed inconsistent with their other annotations – often they had missed instances of hate speech which they subsequently felt were clear cases. One of the authors checked and corrected these apparent "errors" in annotator labeling to create a *gold* corpus. Results for both the original majority class annotations and the "gold" annotations are presented in Section 7.

As a way of gauging the performance of human annotators, we compared two of the annotators' labels to the gold corpus by treating their labeled paragraphs as input to a two fold cross validation of the classifier constructed from the gold corpus. We computed a precision of 59% and recall of 68% for the two annotators. This sets an upper bound on the performance we should expect from a classifier.

6 Classification Approach

We used the template-based strategy presented in (Yarowsky, 1994) to generate features from the corpus. Each template was centered around a single word as shown in Table 1. Literal words in an ordered two word window on either side of a given word were used exactly as described in (Yarowsky, 1994). In addition, a part-of-speech tagging of each sentence provided the similar part-of-speech windows as features. Brown clusters as described in (Koo, Carreras and Collins, 2008) were also utilized in the same window. We also used the occurrence of words in a ten word window. Finally, we associated each word with the other labels that might have been applied to the paragraph, so that if a paragraph containing the word "god" were labeled "other-hate", a feature would be generated associating "god" with other-hate: "RES:other-hate W+0:god".

We adapted the hate-speech problem to the problem of word sense disambiguation. We say that words have a *stereotype sense*, in that they either anti-semitic or not, and we can learn the sense of all words in the corpus from the paragraph labels. We used a process similar to the one Yarowsky described when he constructed his decisions lists, but we expand the feature set. What is termed log-likelihood in (Yarowsky, 1994) we will call log-odds, and it is calculated in the following way. All templates were generated for every paragraph in the corpus, and a count of positive and negative occurrences for each template was maintained. The ab-

[3] `jewish|jew|zionist|holocaust|denier|rabbi|israel|semitic|semite`

solute value of the ratio of positive to negative occurrences yielded the log-odds. Because log-odds is based on a ratio, templates that do not occur at least once as both positive and negative are discarded. A feature is comprised of the template, its log-odds, and its sense. This process produced 4379 features.

Next, we fed these features to an SVM classifier. In this model, each feature is dimension in a feature vector. We treated the sense as a sign, 1 for anti-semitic and -1 otherwise, and the weight of each feature was the log-odds times the sense. The task of classification is sensitive to weights that are large relative to other weights in the feature space. To address this, we eliminated the features whose log-odds fell below a threshold of 1.5. The resulting values passed to the SVM ranged from -3.99 to -1.5 and from +1.5 to +3.2. To find the threshold, we generated 40 models over an evenly distributed range of thresholds and selected the value that optimized the model's f-measure using leave-1-out validation. We conducted this procedure for two sets of independent data and in both cases ended up with a log-odds threshold of 1.5. After the elimination process, we were left with 3537 features.

The most significant negative feature was the unigram literal "black,", with log-odds 3.99.

The most significant positive feature was the part-of-speech trigram "DT jewish NN", or a determiner followed by jewish followed by a noun. It was assigned a log-odds of 3.22.

In an attempt to avoid setting a threshold, we also experimented with binary features, assigning -1 to negative feature weights and +1 to positive feature weights, but this had little effect, and are not recorded in this paper. Similarly, adjusting the SVM soft margin parameter C had no effect.

We also created two additional feature sets. The *all unigram* set contains only templates that are comprised of a single word literal. This set contained 272 features, and the most significant remained "black." The most significant anti-semitic feature of this set was "television," with a log-odds of 2.28. In the corpus we developed, television figures prominently in conspiracy theories our labelers found anti-semitic.

The *positive unigram* set contained only unigram templates with a positive (indicating anti-semitism) log-odds. This set contained only 13 features, and the most significant remained "television."

7 Preliminary Results

7.1 Baseline Accuracy

We established a baseline by computing the accuracy of always assuming the majority (not anti-semitic) classification. If N is the number of samples and N_p is the number of positive (anti-semitic) samples, accuracy is given by $(N - N_p)/N$, which yielded a baseline accuracy of 0.910.

7.2 Classifiers

For each of the majority and gold corpora, we generated a model for each type of feature template strategy, resulting in six classifiers. We used SVM^{light} (Joachims, 1999) with a linear kernel function. We performed 10 fold cross validation for each classifier and recorded the results in Table 2. As expected, our results on the majority corpus were not as accurate as those on the gold corpus. Perhaps surprising is that unigram feature sets out performed the full set, with the smallest feature set, comprised of only positive unigrams, performing the best.

8 Error Analysis

Table 3 contains a summary of errors made by all the classifiers. For each classifier, the table reports the two kinds of errors a binary classifier can make: false negatives (which drive down recall), and false positives (which drive down precision).

The following paragraph is clearly anti-semitic, and all three annotators agreed. Since the classifier failed to detect the anti-semitism, we use look at this example of a false negative for hints to improve recall.

> *4. That the zionists and their american sympathizers, in and out of the american media and motion picture industry, who constantly use the figure of "six million" have failed to offer even a shred of evidence to prove their charge.*

Table 1: Example Feature Templates

unigram	”W+0:america”
template literal	”W-1:you W+0:know”
template literal	”W-1:go W+0:back W+1:to”
template part of speech	”POS-1:DT W+0:age POS+1:IN”
template Brown sub-path	”W+0:karma BRO+1:0x3fc00:0x9c00 BRO+2:0x3fc00:0x13000”
occurs in ±10 word window	”WIN10:lost W+0:war”
other labels	”RES:anti-muslim W+0:jokes”

Table 2: Classification Performance

	Accuracy	Precision	Recall	F1
Majority All Unigram	0.94	0.00	0.00	0.00
Majority Positive Unigram	0.94	0.67	0.07	0.12
Majority Full Classifier	0.94	0.45	0.08	0.14
Gold All Unigram	0.94	0.71	0.51	0.59
Gold Positive Unigram	0.94	0.68	0.60	0.63
Gold Full Classifier	0.93	0.67	0.36	0.47
Human Annotators	0.96	0.59	0.68	0.63

Table 3: Error Report

	False Negative	False Positive
Majority All Unigram	6.0%	0.1%
Majority Positive Unigram	5.6%	0.2%
Majority Full Classifier	5.5%	0.6%
Gold All Unigram	4.4%	1.8%
Gold Positive Unigram	3.6%	2.5%
Gold Full Classifier	5.7%	1.6%

The linguistic features that clearly flag this paragraph as anti-semitic are the noun phrase containing *zionist ... sympathizers*, the gratuitous inclusion of *media and motion picture industry* and the skepticism indicated by quoting the phrase *"six million"*. It is possible that the first feature could have been detected by adding parts of speech and Brown Cluster paths to the 10 word occurrence window. A method for detecting redundancy might also be employed to detect the second feature. Recent work on emotional speech might be used to detect the third.

The following paragraph is more ambiguous. The annotator knew that GT stood for gentile, which left the impression of an intentional misspelling. With the word spelled out, the sentence might not be anti-semitic.

> *18) A jew and a GT mustn't be buried side by side.*

Specialized knowledge of stereotypical language and the various ways that its authors mask it could make a classifier's performance superior to that of the average human reader.

The following sentence was labeled negative by annotators but the classifier predicted an anti-semitic label.

> *What do knowledgeable jews say?*

This false positive is nothing more than a case of over fitting. Accumulating more data containing the word "jews" in the absence of anti-semitism would fix this problem.

9 Conclusions and Future Work

Using the feature templates described by Yarowsky we successfully modeled hate speech as a classification problem. In terms of f-measure, our best classifier equaled the performance of our volunteer annotators. However, bigram and trigram templates degraded the performance of the classifier. The learning phase of the classifier is sensitive to features that ought to cancel each other out. Further research on classification methods, parameter selection and optimal kernel functions for our data is necessary.

Our definition of the labeling problem could have been more clearly stated to our annotators. The anti-immigrant category in particular may have confused some.

The recall of the system is low. This suggests there are larger linguistic patterns that our shallow parses cannot detect. A deeper parse and an analysis of the resulting tree might reveal significant phrase patterns. Looking for patterns of emotional speech, as in (Lipscombe, Venditti and Hirschberg, 2003) could also improve our recall.

The order of the paragraphs in their original context could be used as input into a latent variable learning model. McDonald (McDonald et al, 2007) has reported some success mixing fine and course labeling in sentiment analysis.

Acknowledgments

The authors are grateful for the data and the support of Matthew Holtzman of the American Jewish Congress. We would also like to thank Belle Tseng, Kim Capps-Tanaka, Evgeniy Gabrilovich and Martin Zinkevich of Yahoo! for providing data as well as for their financial support. Without their support, this research would not have been possible.

References

[Choi et al 2005] Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan, *Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns*. In *HLT '05* Association for Computational Linguistics Stroudsburg, PA, USA, pp. 355-362, 2005

[Yarowsky 1994] David Yarowsky, *Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French*. In *ACL-94*, Stroudsburg, PA, pp. 88-95, 1994

[Yarowsky 1995] David Yarowsky, *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*. In *ACL-95*, Cambridge, MA, pp. 189-196, 1995.

[Nockleby 2000] John T. Nockleby, *Hate Speech*. In *Encyclopedia of the American Constitution* (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan, 2000), pp. 1277-1279 (see `http://www.jiffynotes.com/a_study_guides/book_notes/eamc_03/eamc_03_01193.html`)

[Stephens-Davidowitz 2011] Seth Stephens-Davidowitz, *The Effects of Racial Animus on Voting: Evidence Using Google Search Data* `http://www.people.fas.harvard.edu/~sstephen/papers/RacialAnimusAndVotingSethStephensDavidowitz.pdf`

[McDonald et al 2007] McDonald, R. Hannan, K. Neylon, T. Wells, M. Reynar, J. *Structured Models for Fine-to-Coarse Sentiment Analysis*. In *ANNUAL MEETING- ASSOCIATION FOR COMPUTATIONAL LINGUISTICS* 2007, CONF 45; VOL 1, pages 432-439

[Pang and Lee 2008] Pang, Bo and Lee, Lillian, *Opinion Mining and Sentiment Analysis*. In *Foundations and Trends in Information Retrieval*, issue 1-2, vol. 2, Now Publishers Inc., Hanover, MA, USA, 2008 pp. 1–135

[Pang, Lee and Vaithyanathan 2002] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar *Thumbs up?: sentiment classification using machine learning techniques*. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002 pp. 79-86

[Qiu et al 2009] Qiu, Guang and Liu, Bing and Bu, Jiajun and Chen, Chun *Expanding domain sentiment lexicon through double propagation*. In *Proceedings of the 21st international jont conference on Artificial intelligence*, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2009 pp. 1199-1204

[Joachims 1999] *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[Koo, Carreras and Collins 2008] *Simple Semi-supervised Dependency Parsing* In *Proc. ACL/HLT* 2008

[Xu and Zhu 2010] *Filtering Offensive Language in Online Communities using Grammatical Relations*

[A Razavi, Diana Inkpen, Sasha Uritsky, Stan Matwin 2010] *Offensive Language Detection Using Multi-level Classification* In *Advances in Artificial Intelligence* Springer, 2010, pp. 1627

[Attenberg and Provost 2010] *Why Label When You Can Search?: Alternatives to active learning for applying human resources to build classification models under extreme class imbalance,* KDD 2010

[Lipscombe, Venditti and Hirschberg 2003] *Classifying Subject Ratings of Emotional Speech Using Acoustic Features.* In *Proceedings of Eurospeech* 2003, Geneva.