Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Locate the Hate: Detecting Tweets against Blacks

Irene Kwok and Yuzhou Wang
Computer Science Department, Wellesley College
21 Wellesley College Rd, Wellesley, MA 02481
ikwok, ywang5@wellesley.edu

Abstract

Although the social medium Twitter grants users freedom of speech, its instantaneous nature and retweeting features also amplify hate speech. Because Twitter has a sizeable black constituency, racist tweets against blacks are especially detrimental in the Twitter community, though this effect may not be obvious against a backdrop of half a billion tweets a day.[1] We apply a supervised machine learning approach, employing inexpensively acquired labeled data from diverse Twitter accounts to learn a binary classifier for the labels "racist" and "nonracist." The classifier has a 76% average accuracy on individual tweets, suggesting that with further improvements, our work can contribute data on the sources of anti-black hate speech.

Introduction

Regardless of a Twitter user's identity, access to Twitter means access to freedom of speech and a voice. However, in negotiating censorship and free speech, Twitter should be aware of how racially charged dialogues on their platform will impact human lives. Yet there is barely any research on detecting hate speech, with only one recent case on anti-Semitism (Warner and Hirschberg 2012).

In their 2010 study on "Twitter Usage in America," Edison Research reported that while blacks accounted for 13.5% of the general population, they represented 25% of the Twitter population.[2] According to a 2012 poll conducted by The Associated Press, 51% of Americans are overtly anti-black, and with an additional implicit racial attitudes test, that percentage has increased to 56%.[3] With so many black Twitter users shaping trends daily, juxtaposed with the active accounts of anti-black individuals and organizations, such as those of KKK members and allies, racial conflicts and anti-black sentiments are inevitably erupting on Twitter. The problem is that the effects of this hate speech, though substantial, is not always evident given Twitter's instant feeds.

On November 1, 2012, a Twitter user tweeted, "So an 11year old nigger girl killed herself over my tweets? ^_^ thats another nigger off the streets!!"[4] This was retweeted 77 times with 17 favorites, and the user presently has 14,959 followers. Given these numbers, the expected circulation of her similarly racist tweets is alarmingly extensive. Therefore, it is both significant and relevant to investigate racism against blacks on Twitter.

Data, Results, Analysis[5]

Survey

We designed a survey to gauge the complexity of identifying hate speech by applying a statistical measure, Fleiss' Kappa, to assess the reliability of agreement. We compiled a hundred tweets that contained keywords or sentiments generally found in hate speech, and asked three students of different races (but of the same age and gender) to classify whether a tweet was offensive or not, and if classified as offensive, to rate how offensive it was on a scale of one through five (with five being the most offensive). The calculated percentage of overall agreement was only 33%, indicating that this classification would be even more difficult for machines to do accurately, which is consistent with previous research (Razavi et al. 2010).

[1] http://www.complex.com/tech/2012/10/twitter-ceo-dick-costolo-reveals-staggering-number-of-tweets-per-day

[2] http://www.edisonresearch.com/home/archives/2010/04/twitter_usage_in_america_2010_1.php

[3] http://usnews.nbcnews.com/_news/2012/10/27/14740413-ap-poll-majority-harbor-prejudice-against-blacks?lite

[4] https://twitter.com/AntiDARKSKINNED/status/264126778153529344

[5] An extended version of this paper with a more detailed description can be found at http://tempest.wellesley.edu/~ywang5/aaai/paper.html.

Data Collection and Classification

We chose to implement the Naïve Bayes classifier, which performs just as well as more sophisticated classifiers (Huang et al. 2003), to distinguish between racist and nonracist tweets. In order to train one, we built a balanced training dataset composed of sample tweets that would already be classified and contain overlapping features.

Racist tweets were selected from Twitter accounts that were self-classified as racist or deemed racist through reputable news sources with regards to anti-Barack-Obama articles. We then searched for prominent features found in the racist dataset and spelling variations of those features, and determined which tweets and accounts were nonracist. We processed this balanced training dataset of 24582 tweets by eliminating URLs, mentions, stopwords, and punctuation; lowercasing; and equating alternative spellings of slurs to its properly spelled equivalent.

Upon our first analysis of the survey's tweets, we derived labels for why we found each tweet to be racist. These initial labels included "contains offensive words," "reference to painful historical contexts," "stereotypes," "threatening," etc. Of the tweets that were racist against blacks, we found that 86% were labeled as racist because they contained offensive words. Thus, we focused on unigram features when constructing our vocabulary from the processed tweets in our training dataset, obtaining 9437 unique words in the racist training dataset and 8401 unique words in the nonracist training dataset.

Results and Discussion

We evaluated the accuracy of our classification by using the 10-fold cross-validation method, achieving an average accuracy of 76% and an average error rate of 24%.

Because we only employ unigrams, information such as text sentiments are not considered but should be. For example, according to our results, features like "black," "white," and "filthy" are likely used in hate speech against blacks. Yet outside of context, these words bear no racial undertones of their own. Because our classifier does not use bigrams to capture the relationship between words, it may mistakenly classify any tweet containing these racist features as racist, thereby reducing accuracy.

Furthermore, hate speech can be expressed in a subtler manner without the presence of race-related features at all (Pang et al. 2002). For example, the tweet "Why did Obama's great granddaddy cross the road? Because my great granddaddy tugged his neck chain in that direction" contains no single word that is obviously negative by itself, yet this is a racist twist to a traditional American joke.[6] Thus, hate speech seems to require more understanding than the usual topic-based classification.

Certain features also require a social framework and knowledge of a Twitter user's racial identity. "Niggers" and "nigger" bear the greatest racist feature counts at 3040 and 2363 respectively, whereas "niggas" and "nigga" are only found in our nonracist dataset 516 and 763 times respectively. These numbers imply that "niggers" and "nigger" are standard for insulting blacks, averaging to one "nigger"-related word per two tweets against blacks. In contrast, we observe "niggas" and "nigga" to be limited to informal speech, generally confined within the black Twitter community. According to our findings, "nigga" and "niggas" are the only two features out of the top ten who have an association with race. While the two words do not embody racial connotations in meaning, acceptable usage of these words is restricted to blacks and approved allies of blacks. In the black Twitter community, "nigga" has been observed to function as a synonym for the word "person" while implying the male gender. Therefore, humans may also consider a tweet to be racist or nonracist depending on the racial identity of the tweet's owner, in which case the tweet's contents are rendered as secondary in classifying racism.

Conclusion

We have shown that our bag-of-words model is insufficient to accurately classify anti-black tweets. Although the discrepancy found in our survey reflects the difficulty in achieving this accuracy, this challenge should serve as motivation for searching for ways to further refine our classification. Our algorithms need to include bigrams, as well as sentiment analysis and classification, word sense disambiguation, etc. Future explorations may include how often we need to incorporate new vocabulary, how we may utilize popular hashtags to collect more training data, how we may predict and classify deliberate misspellings, how we may involve the racial identity of Twitter users, whether anti-black tweets are targeted to individuals or to groups, how often anti-black tweets are woven into various conversations, etc. As more and more people participate in social media networks, platforms like Twitter become an intersection for diverse groups and individuals, which in turn makes our research increasingly relevant.

References

Huang, J., Lu, J., Ling, C. X. 2003. Comparing naïve bayes, decision trees, and svm with auc and accuracy. In *Proc. of the 3rd IEEE ICDM03*.

Pang, B., Lee, L., Vaithyanathan, SH. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of the Conference on EMNLP*, 9-86.

Razavi, A., Inkpen, D., Urisky, S., Matwin, S. 2010. Offensive language detection using multi-level classification. In *Advances in Artificial Intelligence Springer*, 1627.

Warner, W., Hirschberg, J. Detecting hate speech on the World Wide Web. 2012. In *Proc. of the 2012 Workshop on LSM*, 19-26.

[6] https://twitter.com/Walken4GOP/status/265928481572007938