Detecting Racism in Tweets

Fiete Botschen, Sean Noran

December 2015

1 Abstract

Racism is a ubiquitous violation of human rights that often goes undetected in on-line communities such as Twitter. This paper explores the application of Machine Learning approaches to detect racism in Tweets. We examine several models, including Naive Bayes, Support Vector Machine and Random Forest combined with various features such as bag-of-words features, bigram features and term-frequency inverse document-frequency scores. We achieve an F1 score of 92.6% using a Random Forest with all of the features over a balanced dataset containing 50% racist Tweets.

2 Introduction

Racism is a world-wide ubiquitous problem that violates essential human rights. In recent years, the world-wide web has become a platform for racist comments due to the anonymity it promises. In particular, many people have expressed hatred and xenophobia toward the recent influx of refugees. On-line social networking communities such as Twitter have been flooded with racist remarks in response to such race-related events, including the offensive Sigma Alpha Epsilon fraternity chant, the shooting of Michael Brown in Ferguson, Missouri and the subsequent riots, and the Black Lives Matter movement. Because those who make hateful remarks are often protected under First Amendment rights, it is essential that we examine the roots of such hatred. Are there systemic sources of racism that should be addressed? Can we identify respected institutions that propagate racist beliefs? These questions have largely been addressed qualitatively. This work aims at quantitatively tackling these issues, by examining racist comments among a community of Twitter users. This work focuses on the initial step of identifying racist Tweets; we address future exploratory analysis in Section 7 Discussion and Future Work.

3 Related Work

Greevy and Smeaton [1] used a Support Vector Machine with a polynomial kernel to classify documents as racist over both bag-of-words features and word

bigrams separately. They achieve 92.78% precision and 90% recall on 200 documents after training their model on 800 documents. At the same time, the documents they are using are web pages or articles and not tweets, meaning that their documents are of much bigger size, but the number of available documents is much smaller.

Kwok and Wang [2] detect racism specifically against Blacks in Tweets with 76% accuracy using a Naive Bayes model over bag-of-words features. However, they conclude that their classification is insufficient applied to other sources than their test set. They propose to use the racial identity of a Twitter user which is likely to produce better results. However, this information is rarely provided in the Twitter account and would thus need to be predicted.

Rauch and Schanz [3] took a look at the acceptance and susceptibility of Facebook users to messages with racist content. Their work demonstrates that the amount of racist content on Facebook increases and that an increasing number of people are susceptible to it. This corroborates the significance of our work.

Warner and Hirschberg [4] use a Support Vector Machine to detect hate speech over labeled data provided by Yahoo! and the American Jewish Congress. Offensive language is defined by them as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation". In particular, they developed a classifier for anti-Semitism, which is a subset of our problem space. Their accuracy is 94% and their F1-score is 0.63. They used unigram and bigram features, as well as part-of-speech tags over windows and log-likelihood ratios between counts of positive and negative instances.

Razavi et al. [5] use a combination of statistical models and rule-based approaches to detect offensive language in text. However, they do not focus solely on racism, but also look to detect homophobia, crude language and provocation, a much broader task than we aim to achieve.

Mahmud et al. [6] similarly look to detect flames in text and use semantic dependencies to determine whether the comment is directed at an individual or not. They, too, examine more general instances of offensive language without focusing on racism.

Correa and Sureka [7] research the situation that more and more radical groups use the Internet to coordinate and radicalize collectively. This applies of course also to certain groups who have racist views against others in common. They are not only focusing on how to detect the radical part of documents in social media, but also how to get information how they connect to each other. This would also be applicable and interesting in our setting, although it is out of the scope of our work: Do Twitter user who tweet racist content connect with each other and if yes, how? A lot of information (such as retweets) is available within the data stream of Twtter.

Chae DH et al. [8] quantify racism by U.S. region defined in terms of the number of Google searches of the N-word proportional to the number of people in the area. They also map the mortality rate of African Americans by region and find correspondences between their results. However, they only address racism towards African Americans and they disregard the fact that not all uses of the N-word are racist.

Xu and Zhu [9] filter offensive words using a parse tree to determine the syntactic role of candidate words. Their model is able to identify unseen offensive words as well as remove inoffensive words that could allow users to easily identify the censored offensive word. They evaluate their system on over 11000 text comments from YouTube. Of these comments, 1739 comments contain at least one of 368 offensive words. The overall accuracy of their semantic-based filter is 90.94%. While this identifies offensive language, it does not find racism that lacks offensive words.

4 Data

4.1 Labeled Data Collection

Due to a lack of freely available labeled data, we manually established a labeled dataset of approximately 15000 Tweets from three sources: a filtered subset of a substantially large Twitter dataset which spans over 20 days in January 2011 and 20 days in March 2015; a Twitter user that retweets primarily racist content; and a Twitter user that posts racist jokes. The first source of Tweets was filtered by terms related to race, including derogatory words referring to various ethnic groups and commonly used words such as *racist* and *black* which might appear in racist contexts. Only 8% of the filtered Tweets were labeled as racist. In contrast, approximately 50% of the racist user's retweets were indeed racist and 87% of the joke tweets were labeled as racist.

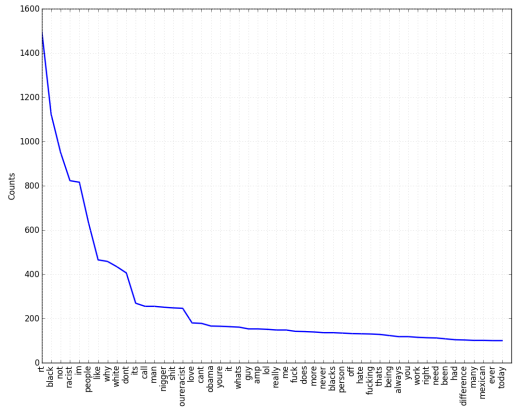

Figure 1: Frequency of most common words, excluding 100 most frequent English words

Of the 15000 tweets, 2157 were labeled as racist. The average Tweet contains about 30 characters and may contain links to images, news articles and other web content. Figure 1 shows the frequency of the most common words found in these tweets, excluding the 100 most common English words.

5 Methods

First, we apply a Topic Model [10] as a way to explore the dataset. For classification we use Bag-of-Words features and Term-Frequency Inverse-Document Frequency [11].
We apply several standard Machine Learning algorithms in order to detect racism. Specifically, we are working with Support Vector Machines [12], Random Forests [13] and Naive Bayes.

5.1 Preprocessing

The raw data we were starting with consists of json [14] data. The .json file contains geo-tagged information and Tweet information. The most relevant aspect of the geo-tagged information is the language, the most important part of the Tweet information is the Tweet content itself. Additional information such as the user of the tweet, the number of times the Tweet was favorited, etc. is available, but was not used in this work.
We filter the Twitter stream data by a set of selected racism-related words (see 6.1) obtained from several web sources such as Wikipedia [15], the racial slur database [16] and the Hammerskin forum [17]. The filter contains 94 keywords which tend to be used against several ethnic groups.

Table 1: list of racism related keywords

nigger	crow	jew	tacohead	blackie
goy	heeb	polack	kyke	racist
jew	jewish	jewing	mulatto	spic
spik	spig	spick	jap	gypsies
gypsy	kaffir	kaffer	kafir	kaffre
jypped	jipped	gyp	jip	gypped
chong	supremacy	latino	hispanic	natives
chink	negro	slave	bombing	racial
nihao	ethnicity	ethnic	mosshead	aye-rab
beano	blackie	bog-rat	bog-hopper	buckwheat
canuck	chico	chinaman	chink	clod-hopper
coon	darkie	dink	gook	goy
goyem	greaser	grease-ball	gringo	half-breed
heeb	hick	hillbilly	honky	injun
jap	jigaboo	kyke	kraut	limey
paki	polack	rusky	shit-kicker	sambo
shyster	sand-nigger	spic	spook	towel-head
wasp	whitey	wop	yid	zebra-head

This preprocessing step only applied to the stream of daily Tweets. It was not necessary to use this on the collection of racist Tweets from the re-tweeter and the user who tweets racist jokes.

We extract the Tweets out of the .json files and filter out the 100 most common English words [18]. Later, we discovered the Mallet Stop Words [19], which did not change our results. We also convert all the input to lowercase since we made the assumption that capitalization does not influence the significance of the classifiers. Plus, punctuation characters are removed.
Here are some results of the preprocessing step:

- Raw tweet: 'Obama', 'you', 'piss', 'me', 'off...', 'yes', 'because', 'your', 'a', 'nigger.
- Result: 'obama', 'piss', 'off', 'yes', 'nigger'
- Raw tweet: "I'm", 'not', 'racist', 'but', 'I', 'H8', 'Mexicans', 'they', 'need', 'to', 'ship', 'the', 'ass', 'back', 'on', 'the', 'other', 'side', 'of', 'the', 'border'
- Result: 'im', 'not', 'racist', 'h8', 'mexicans', 'need', 'ship', 'ass', 'side', 'border'

5.2 Topic Modeling

We perform Topic Modeling (TM) [10] to explore our dataset. The core principle behind TM is the assumption that each word in a document, in this case Tweet, is generated by a certain topic. Hence, for generating a word, first a topic

has to be drawn from a distribution over all topics, and then the word has to be sampled out of that topic distribution. There exist a lot of different methods to implement TM; we chose the most common approach called Latent Dirichlet Allocation (LDA) [20]. TM does no hard clustering, meaning that each Tweet consists of potentially all topics, most of them presumably with a low probability. Finally, it should be mentioned here that LDA is a randomized process, meaning that running LDA several times on the same dataset can lead to different topic distributions.

5.3 Naive Bayes

Naive Bayes(NB) is a supervised learning technique which basically models probability distributions for classes and uses Bayes Rule [21] to classify new data. Since we are working with words, the probability distribution is discrete in this case. The probability distribution is modeled by counting how often a word occurs within one class.
Hence, the training part of a NB classifier consists of counting each word in the vocabulary. For classification, the probability of each document appearing in a class is being computed by multiplying the likelihood (or adding the log-likelihood) and normalizing over all classes.
The strength of a NB classifier is that it is a very simple and efficient algorithm. The weakness is that it assumes that all the words are independent of each other, so it does not consider the fact that given some surrounding racist words the probability of another word being racist might be actually higher.

5.4 Support Vector Machine

The problem of dependency between words can be tackled by several more complex algorithms, such as the Support Vector Machine (SVM). A SVM projects the data (e.g. Bag of Word features(BoW)) into a higher dimensional space in order to separate them by a linear hyperplane. The objective of a training a SVM is to maximize the margin of the separating hyperplane. The classification is quite simple: Given a new sample and its projection into the space of the classifier, the sample will be either on one or the other side of the separating hyperplane. Since not all (especially our) data is perfectly separable, a SVM can be tuned to use either a soft or hard margin, meaning that it "allows" a certain bias to converge or not. Since the SVM works with all BoW features as a single data-point, it automatically considers the dependency between those features.

5.5 Random Forest

Random Forests(RF) [22] is a set of algorithms out of the family of graphical models. The word "random" is proverbial in that case: The trees created by a Random Forest are created randomly, therefore, training a random forest usually means trying several forests out and choosing the best one. We don't

want to discuss reasons for going for a SVM instead of a RF or the other way round (since this is highly debatable). Since there are a lot of publications in which a RF beats a SVM and also the SVM beating a RF, working with a RF seemed natural. Specifically, the randomness refers to the feature selection. So each time a new RF is created, it uses (potentially) a different subset of the features.

5.6 Bag of Word features

As a starting point, we decided to work with Bag of Word(BoW) features. Bag of Word features are quite simple: After the preprocessing step, each word occurring in the Tweet is counted. Hence, the feature vector for a BoW of a Tweet is simply a mapping from each possible word to the number of occurrences in the Tweets. It is essential to mention that some dependency is lost because information like the order of the words or the surrounding context of the words is not encoded in this counting scheme. However, BoW features still contain some informative non-sequential contextual information.

5.7 Term-Frequency Inverse Document-Frequency Features

Term-Frequency Inverse-Document Frequency Features (TFIDF) [11] capture the importance of a word in a document, given a larger corpus of documents. The TF-IDF value of a word increases proportionally with the count of the word in its document, offset by the count of the word across all documents. In this particular case, a Tweet constitutes a document. TF-IDF scores address the noise introduced by commonly seen words that are generally irrelevant to racism.

6 Results

6.1 Results of preprocessing and data selection

Due to our data selection methods, where we had to switch from filtering tweets out of the Twitter stream to use tweets from the retweeter or the "racist jokes tweeter", we ended up with a dataset which consists out of tweets in a highly racism related setting. Most of the tweets are either pro- or against racism (which is also shown by the Topic Modeling, see below).
This basically changed our task significantly: Instead of assuming random tweets, we are now training on a set of racism - related tweets and our goal is to distinguish between racism and no racism within that context. (see 7.1).

6.2 Topic Modeling

In the following, we present the most common words for the topics depending on the dimension of the Topic Model for several runs (since LDA is a randomized process (see 5.5)):

Table 1: list of racism related keywords

No. topics	run	topic id	Most common words for topic
1	1	1	racist im not like people black white skin its dont
1	2	1	racist im not like people black white dont skin slave
1	3	1	racist im not like black people white skin dont slave
2	1	1	slave skin arab mexican dago latino lol di chico job
2	1	2	racist im not people like white black yesyoureracist racial its
2	2	1	racist white im arab people dago jewish skin black slave
2	2	2	racist im not like dont black its yesyoureracist people latino
2	3	1	racist im not people white black like jewish amp why
2	3	2	racist im skin slave lol dago not like arab nigger
3	1	1	arab dago chico skin kkk di dont jap haha love
3	1	2	racist im not people like black white slave yesyoureracist dont
3	1	3	racist like lol white job amp chicos black 2 its
3	2	1	mexican race chico im asian jewish d gypsy ya skins
3	2	2	arab dago slave skin chicos kkk job di jobs jap
3	2	3	racist not im people like black dont white yesyoureracist lol
3	3	1	slave racist amp its race more black crow not racial
3	3	2	skin dago arab latino kkk chico im di de negro
3	3	3	racist not im like white people black dont yesyoureracist shit

Since we are in a binary setting, the Topic Model with two topics is most interesting. However, the results for one and three topics are interesting in terms of our data analysis.

First of all, one clearly sees that there seem to be some dominant words in all the Tweets. The different runs for the TM with one topic show that the word "racist" occurs very often. This is presumably because of a lot of people writing racist comments actually state that they are *not* racist. This also explains why the words "im" and "not" also occur in at least one topic of all the topics in the runs of the TM with two topics. Looking at the TM with one topic also shows that apparently amongst the English tweets there is a trend to use skin color (black or white) in racist context. This is not surprising, however the results of TM with one topic show that apparently most of the racism we are working with is within the context of insulting people of darker skin color. Since most of the Tweets are of U.S. origin, we would have expected to see words referring to the Hispanic-American population. From manual labeling, we know that there are a lot of racist Tweets against Hispanic-Americans as well.

Nevertheless, the TM with two topics reflects racism against other ethnic groups such as Hispanic-Americans as well. Words which refer to a lot of ethnicities like "arab", "jew", "latino", "black", "nigger" show up amongst the top ten of all the topic models of size two. Unfortunately, the TMs of size two are, from our perspective, not meaningful regarding distinguishing between racism or no racism. In the given examples, none of the top words of a topic seem to represent the racism or no racism class. This does not change for TMs with higher numbers of topics. We tried up to five topics.

This is the reason why we used Topic Modeling only for data analysis and not as a tool for classification.
Finally, one thing about the TM with three topics is remarkable: Each of the three runs contains one topic with the following top words, where only one or two are different: racist, not, im, like, white, people, black, dont, yesyoureracist, shit, lol, slave. This again shows, as stated before, that there seems to be a trend of racism against blacks while stating that one is not racist. Unfortunately, the other topics are not not-racist which is why this information is interesting, but not usable in the following work.

6.3 Quantitative Evaluation

Table 2 shows the classification results with balanced class data.

Training size: 2000 Tweets. Test size: 2000 Tweets. Class balance: 50% racist.

Table 2: Precision, recall and F1-score on different classifiers and features

Algorithm	**Features**	**Precision**	**Recall**	**F1-score**
SVM	Unigram	86.1%	97.5%	91.4%
SVM	Bigram	93.6%	77.8%	85.0%
SVM	TF-IDF	86.7%	98.2%	92.1%
SVM	Unigram + Bigram	87.4%	97.8%	92.3%
SVM	Unigram + Bigram + TF-IDF	87.3%	97.8%	92.3%
RF	Unigram + Bigram + TF-IDF	88.7%	96.8%	92.6%
NB	Unigram + Bigram*	79.7%	99.9%	88.6%

*Note that TF-IDF features were not compatible with the Naive Bayes implementation we were using because it required an integral feature vector.

However, balanced data is not representative of the actual distribution of racist Tweets. Therefore it is essential to evaluate our model on a realistic distribution. The results are shown below in Table 3.

Number of racist tweets: 2157. Number of normal tweets: 10429. Training size: 2000 Tweets. Test size: 2000 Tweets. Class balance: 5% racist.

Table 3: Precision, recall and F1-score on different classifiers and features

Algorithm	**Features**	**Precision**	**Recall**	**F1-score**
SVM	Unigram	23.0%	98.0%	37.2%
SVM	Bigram	38.5%	72.0%	50.2%
SVM	TF-IDF	23.3%	98.0%	37.6%
SVM	Unigram + Bigram	25.2%	99.0%	40.2%
SVM	Unigram + Bigram + TF-IDF	25.0%	99.0%	39.9%
RF	Unigram + Bigram + TF-IDF	26.6%	98.0%	41.8%
NB	Unigram + Bigram	15.4%	100.0%	26.7%

As is clear in the tables above, the precision drops significantly when we account for the prior distribution. This suggests that we are over-fitting to the training data.

6.4 Qualitative Evaluation

In order to understand the limitations of our model, it is insightful to examine some of the errors the model makes. In general, the model has a high false positive rate and a low false negative rate. Some of the false positives include images and links. In these cases the ground-truth annotations are determined only by the text and not by the links contained in the text. Figure 2 is an example of a Tweet, which, given the image contained in the Tweet, is clearly racist; however, because the ground-truth annotations do not account for the image, it is labeled as not racist.

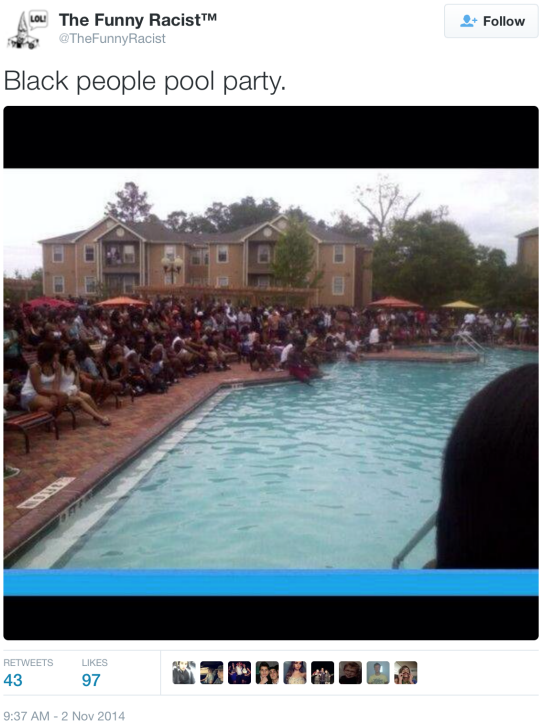

Figure 2

Another example of a false positive is shown below. This was labeled as not racist because it does not refer to a specific group of people. However, many people may consider this Tweet racist.

Figure 3

The following Tweet was jokingly intended to sound like a racist comment, but as far as we can tell, nothing racist was said.

Figure 4

These Tweets could have been racist in a particular context. It is understandable that the model misclassifies Tweets that are ambiguous to human readers.

The false negatives are more interesting because they capture the weaknesses of our model. While the following Tweet is clearly racist, the model may be confounded by the choice of language. The phrase *just don't like* is much less direct than most racist Tweets. The word *mfs* (motherfuckers) is very clearly derogatory but is not a commonly used acronym and may not have appeared at all in the training data.

Figure 5

The following Tweet is an example of blatant racism that went undetected by the classifier. It is surprising that even the bigram features failed to capture the racist content of this Tweet. For instance, the bigram *communist nigger*, which occurred several times in the dataset, should unambiguously identify this Tweet as racist.

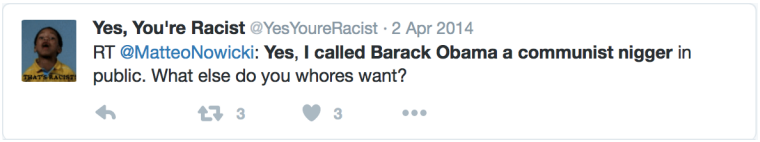

Figure 6

7 Discussion and Future Work

7.1 Bias of the dataset

As stated in 6.1, our setting changed from classifying Tweets of the Twitter stream to Tweets which are already in a highly racism-related context. This means that if we would use the introduced classifier on the random Twitter stream, the classifier would be biased towards racism since it hasn't learned many normal tweets. From our perspective, this has rather positive than negative implications. For example, in a racial context, President Obama has a very high probability of being the subject of a racist remark. If one does not know anything about the context, this changes of course significantly, because the President of the United States is a person a lot of people talk about in a variety of contexts.
So, given that the strength of our classifier is to distinguish between racism or no racism in a racism-related context, we would be interested to see how our classifier would perform on filtered data - this would demonstrate its ability much more comprehensively.
This additional filter / classifier would of course need to be more sophisticated than the simple filter described in 5.1, which failed to cover all racist contexts.

7.2 Ground-Truth Annotations

A significant flaw in this work is that the ground-truth annotations are subjective and, in particular, the annotators are two white males university students. Thus, the labels reflect the definition of racism from the perspective of this particular demographic and disregards the perspective of minority groups who are generally targets of racism.

Additionally, the words that were used for filtering the Tweets are subjective and enormously impact the distribution of the ethnic groups referred to in the annotated data. For example, although the word *black* is found in many racist Tweets, it is primarily found to refer to the color. By omitting the word *black*, the filtered Tweets have a higher proportion of racist content, but the number of racist Tweets referring to blacks will drop. As another example, we did not filter using the word *wetback*, a highly derogatory term used to refer to Central Americans who illegal enter the United States. Thus, we make no claim that our data reflects a realistic distribution of demographics over racist Tweets.

These issues beg the question: What is racism? The U.S. Commission on Civil Rights defines racism as "any action or attitude, conscious or unconscious, that subordinates an individual or group based on skin colour or race. It can be enacted individually or institutionally" [23].

However, there is vast disagreement regarding how to define *race* and *subordinate*. As an example, the two annotators disagreed on the content of the following Tweet:

> RT @DeadlySushi: @YesYoureRacist Instead of complaining about police fix youre neighborhoods first and stop doing crimes! Point the finger ...

After seeing many Tweets referring to this context, the shooting of African American Michael Brown in Ferguson, Missouri, one annotator felt that this Tweet was racist, while the other did not.

We would like to extend this work so that the annotations reflect a more comprehensive notion of racism. One possible approach to address this issue is to gather annotations over many people of different backgrounds and to use the average label as the ground-truth. It may be more appropriate to examine the distribution of responses for various ethnic groups and develop a metric that captures some high-level aspects of these distributions. This might entail predicting whether a particular group might find a Tweet offensive or incorporating a non-binary measure that describes how offensive the Tweet is.

7.3 Grammatical Features

While not informative alone, it may be useful to examine the grammatical structure of the Tweet in addition to the other features. It is possible that racism co-occurs with poor grammar; therefore, it would make sense to use features that capture grammatical errors. On the other hand, this may be a noisy feature due to the general divergence from Standard English grammar on Twitter. Empirical analysis would be necessary in determining the extent of its usefulness.

7.4 Non-textual features

The Tweet alone is often not enough to identify racist content. Often information about the Twitter user, the geographic location, and the social context will be informative in conjunction with the text. For instance, the following Tweet is racist, because *Kanye West* is a black artist and the *frat song* refers to a racist chant sung by the Sigma Alpha Epsilon fraternity at the University of Oklahoma.

> Kanye West should rap the frat song

Without enough training data, a bag-of-words model may be able to address this issue, but if the training data consists only of Tweets prior to the incident

at University of Oklahoma, then this particular example would not be flagged by our model. Thus, for uncommon bigrams such as *frat song* and proper nouns such as *Kanye West*, it may be useful to use features learned from a web crawl, such as incrementally learned context vectors.

Additionally, the date of the Tweet may be informative, because there are often temporal dependencies between Tweets and events they reference. See

7.5 Clustering Racist Tweets by Location

There is a tremendous amount of exploratory analysis into racism where this work can have enormous implications. It is of great interest whether there are geographic dependencies that govern racist thought. For example, many believe that racism is much more common in the U.S. south than in the north, but there has been little quantitative analysis of these claims. Chae DH et al. [8] quantify racism by area according to the number of Google searches of the N-word proportional to the number of people in the area. They only address racism towards African Americans and they disregard the fact that not all uses of the N-word are racist.

We plan to detect racism over a substantially large set of arbitrary Tweets and cluster the Tweets which are detected as racist. The clusters will then be normalized by dividing by the number of Twitter users in those locations. This will give an idea of where racism is concentrated in the United States and whether there are significant geographic dependencies.

7.6 Finding Dependencies between Racist Tweets and World Events

This work can be used to explore whether current events that relate to race align with spikes in the number of racist Tweets. Costa et al. [24] showed that each wave of the Brazil protests of 2013 corresponded to a significant peak in relevant Tweets. Would we expect to see analogous trends for the Black Lives Matter movement or for the influx of Syrian refugees?

References

[1] Edel Greevy and Alan Smeaton. Classifying racist texts using a support vector machine. *SIGIR*, 4:468–69, 2004.

[2] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. *AAAI Conference on Artificial Intelligence*, 27, 2013.

[3] Shannon Rauch and Kimberly Schanz. Advancing racism with facebook: Frequency and purpose of facebook use and the acceptance of prejudiced and egalitarian messages. *Computers in Human Behavior*, 29.3:610–15, 2013.

[4] William Warner and Julia Hirschberg. Detecting hate speech on theworld-wideweb. *LSM Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, 2012.

[5] Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive language detection using multi-level classification. *Canadian Conference on Artificial Intelligence*, 6085:16–27, 2010. http://link.springer.com/chapter/10.1007/978-3-642-13059-5_5.

[6] Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. Detecting flames and insults in text. *Proc. of 6th International Conference on Natural Language Processing*, 2008.

[7] Denzil Correa and Ashish Sureka. Solutions to detect and analyze online radicalization : A survey. 2013. http://arxiv.org/pdf/1301.4916.pdf.

[8] David H. Chae, Sean Clouston, Mark L. Hatzenbuehler, Michael R. Kramer, Hannah L. F. Cooper, Sacoby M. Wilson, Seth I. Stephens-Davidowitz, Robert S. Gold, and Bruce G. Link. Association between an internet-based measure of area racism and black mortality. *PLoS ONE*, 10, 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122963#pone.0122963.ref001.

[9] Zhi Xu and Sencun Zhu. Filtering offensive language in online communities using grammatical relations. 2010.

[10] David Blei and J.D. Lafferty. Topic models, 2009. http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.

[11] Jones Spärck. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, pages 11–21, 1972.

[12] Bernhard Schölkopf and Alexander Smola. Learning with kernels: Support vector, machines, regularization, optimization, and beyond. 2002.

[13] Tin Kam Ho. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 278–282, 1995.

[14] Yahoo!, java script and json, 12 2015. http://www.programmableweb.com/news/yahoo-javascript-and-json/2005/12/16.

[15] List of ethnic slurs, 12 2015. https://en.wikipedia.org/wiki/List_of_ethnic_slurs.

[16] The racial slur database, 12 2015. http://www.rsdb.org/.

[17] Forum 38, 12 2015. http://www.crew38.com/forum38/forum.php.

[18] Wikipedia: Most common words in english, 12 2015. https://en.wikipedia.org/wiki/Most_common_words_in_English.

[19] Machine learning for language toolkit, 12 2015. `http://mallet.cs.umass.edu/`.

[20] David Blei, Andrew Ng, Michael Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, pages 993–1022, 2003.

[21] Michel Hazewinkel. Bayes formula. *Encyclopedia of Mathematics*, 2001.

[22] Leon Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[23] Racism in america and how to combat it. *U.S. Commission on Civil Rights*, 1970.

[24] Jean M. R. Costa, Rahmtin Rotabi, Elizabeth L. Murane, and Tanzeem Choudhury. It is not only about grievances - emotional dynamics in social media during the brazilian protests. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 594–597, 2015.