A Look at Hate Speech on Reddit

Noopur Gupta
Northwestern University
Evanston, Illinois
noopurgupta2016.1@u.northwestern.edu

Ries Guthmann
Northwestern University
Evanston, Illinois
r-guthmann@northwestern.edu

Philip Meyers
Northwestern University
Evanston, Illinois
meyers@u.northwestern.edu

Content Warning

This project focuses on hate speech, which by its very nature is hateful, offensive, and vulgar. This paper contains text with graphic language, slurs, and other forms of hate speech targeting many peoples and identities. We have chosen not to censor or remove any offensive content to enable the most transparent analysis possible.

1. Introduction

1.1. Reddit

Reddit is composed of many *subreddits* (denoted *r/subreddit* after the website's URL schema `www.reddit.com/r/subreddit`) that are communities revolving around a specific topic. Subreddit topics could be anything from a point of view (*r/conservative* for conservative politics, *r/vegan* for vegans, *r/PCMasterRace* for PC enthusiasts) to a general type of content (*r/HighQualityGifs* for high FPS and high resolution gifs, *r/funny* for humorous content, *r/babyanimals* for pictures and videos of baby animals). As of 2018, Reddit is home to over 1.2 million subreddits and the rate of growth has increased since Reddit's inception over 10 years ago (Figure 2).

Subreddits are not created and owned by Reddit. Instead, they are developed by individual users and grown by capturing the interest of other users on the platform. Subreddits are generally public, meaning that any user (or non-user) can freely browse the community's content and comments. However, certain subreddits are made private in order to limit membership. For example, *r/CenturyClub* only permits users with over 100,000 *karma* (Reddit's points system). Larger subreddits typically have one or more moderator users (mods) responsible for ensuring that content

Figure 1: A popular post on *r/AgainstHateSubreddits* linking to racist content on the Donald Trump fanbase subreddit *r/The_Donald* (abbreviated T_D in the post title).

and behavior aligns with the specific subreddit's guidelines. When enforcing a subreddit's guidelines, a mod may lock a post (preventing further interaction with the post and its content), remove inappropriate content, or even ban an offending user (preventing the user from accessing the content of the subreddit). Mods act at their own discretion and hence subreddits are self-regulated.

The Reddit platform also has the ability to moderate content on the site by deleting content or globally banning users and subreddits that do not comply with the site's rules. Such tools are rarely used by platform itself as doing so is often met with criticism and retaliation from its user base for limiting free speech on the platform. In June 2015, site administrators chose to ban subreddits *r/hamplanethatred*, *r/transfags*, *r/neofag*, and *r/shitniggerssay* for violating the

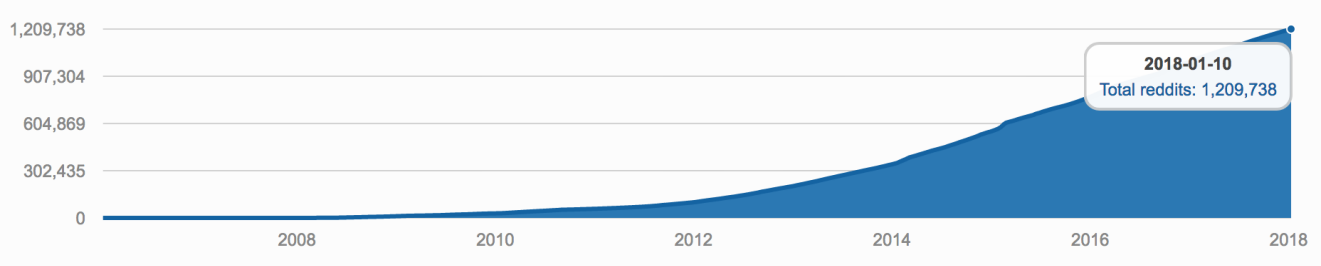


Figure 2: Accelerating growth of number of subreddit communities on Reddit since 2006 [1] .

company's harassment policy. In response, supporters of the banned sites rallied users to leave Reddit for the nearly identical platform *Voat.com*. Reddit admins also face critique from some users that they have not done enough to suppress toxicity across the site. *r/AgainstHateSubreddits* is a community of over 45,000 users who wish to remove hateful subreddits and hate speech from the social media platform. Posts on the subreddit frequently highlight content and subreddits that they believe to be against Reddit's policies (Figure 1). To the delight of the former group and the dismay of the latter group, Reddit has not recently acted to address this kind of speech.

A recent work by Chandrasekharan *et al.* examined the effect of the aforementioned 2015 ban. The authors found the ban to be a success and observed that "other subreddits did not inherit the problem [of hate speech]" after the four subreddits were removed. We agree with the findings and believe that the authors sufficiently covered the ban and its impacts. Hence we are not concerned with the local effects of the 2015 ban, rather we are interested in a broader look at hate speech on Reddit. The Oxford English Dictionary offers the following definition for hate speech:

> Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation.

We believe that hate speech exists far beyond the four subreddits that were the focus of [4]. Hence we endeavor to take a more global look at hate speech on Reddit. In doing so, we hope to answer the following questions:

1. How can we detect hate speech on Reddit?
2. How has the volume of hate speech changed throughout time?
3. What are the common themes of hate speech on Reddit?

2. Data

The data on the Reddit website is nicely organized as a tree structure. Reddit is home to many individual subreddits. Each post belongs to a single subreddit. Comments on a post can either be made in direct reference to the post (top-level) or in response to other comments (replies) (Figure 3).

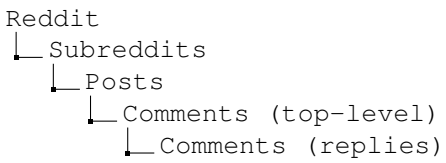


Figure 3: Tree structure of Reddit data.

Our analysis is entirely based upon comments on public subreddits Reddit. Comments are the largest source of data because there can be up to tens of thousands of comments for a single post. They offer insight into the ideas and opinions of the user base through interactions among its users. We source our comments from the publicly available dataset `reddit_comments` in Google BigQuery[1]. The dataset contains every available public comment (at the time of scraping) from December 2005 to March 2018. The dataset contains over 4.1 billion comments and metadata like author (user), post, and subreddit. The following is a sample comment with all available metadata:

We are only concerned with the content of individual comments and not how they relate to content in other comments and the post. Hence we only use the data from the comment body, author, created_utc, and subreddit fields.

2.1. Comments

Our first challenge was figuring out how to detect and isolate hate speech comments. Recall that the original dataset contains over 4.1 billion comments in BigQuery, a high performance SQL database. Attempting to locate relevant comments by hand was inconceivable. Our ability to perform complex NLP on the comments was also limited by the capabilities of and inexperience with SQL. We decided that the best course of action would be

[1] `https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments`

”body”:”ŏ026gt; I never said Israel, i said that place in the middle east thatś full of jews. No no no. He’s in Israel now, having just gotten ”back from the middle east.” ”,
”author”:”YourFairyGodmother”,
”created_utc”:1495493511,
”subreddit_id”:”t5_2cneq”,
”link_id”:”t3_6cnw5y”,
”parent_id”:”t1_dhw1qpp”,
”score”:1,
”retrieved_on”:1496742044,
”controversiality”:0,
”gilded”:0,
”id”:”dhwn6wz”,
”subreddit”:”politics”

```
SELECT
  *
FROM
  `fh-bigquery.reddit_comments.20*`
WHERE
  REGEXP_CONTAINS(author, r"(?i:\buncivilised\b)")
  OR REGEXP_CONTAINS(author, r"(?i:\bgypo\b)")
  OR REGEXP_CONTAINS(author, r"(?i:\bgypos\b)")
  OR REGEXP_CONTAINS(author, r"(?i:\bcunt\b)")
  OR REGEXP_CONTAINS(author, r"(?i:\bcunts\b)")
  OR REGEXP_CONTAINS(author, r"(?i:\bpeckerwood\b)")
```

Figure 4: Screenshot of the top lines of the first query over all Reddit comments in BigQuery.

to collect comments with terms frequently used in hate speech. Hatebase [2] is non-profit website that advertises itself as ”the world’s largest online repository of structured, multilingual, usage-based hate speech”. They offer a free API that can be used to look up hate speech terms, usages, meanings, and target groups. We use the original lexicon from Hatebase to find comments containing hate speech. The lexicon contains 1034 terms that encompass a variety of hate speech. The following is a random sampling of 15 terms from the lexicon:

Honyock	teabagger	Rhine monkey
dune coon	niggresses	trailer park trash
mockies	blaxicans	proddywhoddies
brownie	redneck	beach nigger
nitches	nigors	Merkin

Using a Python script, we generated a SQL query whose WHERE clause is the disjunction of regular expression (RegEx) matches for each of the 1034 terms. Each term RegEx follows the pattern `(?i:\bhate_term\b)`. The term is padded with the word boundary token `\b` to ensure that the a match is for the full hate speech term and not simply for a substring of an unrelated word. Without the word boundary tokens, the comment ”I like puppies!” would be a match for the hate speech term ”ike.” For non-unigram terms, spaces between words are replaced with the whitespace token `\s+` to enable matches for terms in comments where users put more than one space between words. Finally, the case insensitive flag `i` is used to match against all ways of casing the term.

The query (Figure 4) took 48 minutes to complete and returned 78.7 million rows (42.2 GB). After inspecting the results, we saw that although hate speech was being captured, many of the comments in the results were using terms that could be hate speech in unoffensive ways. For example, the term ”yellow” can be used as a slur towards people of Asian descent but in the results we found it mainly being used to refer to the color in an non-hate fashion. In other words, finding matches based on the list of 1034 terms offered very high recall but very low precision. To improve our precision and refine our results, we performed a second query on the 78.7 million results for the occurrence of 178 common hate speech n-grams from [5]. This list of n-grams was during analysis of hate speech on Twitter. Through hand labeling, the authors found that least 50% of the time that an n-gram occurred, the n-gram was being used in hate speech. We used the same technique as above to compose a query of disjunctive RegEx matches for all n-grams. The second query completed in 10 minutes and returned 4 million rows (3.1 GB). We exported all 4 million rows into a JSON file and downloaded it for local use.

2.2. Labels

Although we have a high confidence that the 4 million comments collected are some form of hate speech, we are interested in a finer level of granularity for analysis and prediction. In other words, we would like to know whom specifically does the hate speech target. We are able to get more specific hate speech labels for comment by using the n-gram list as a map. For each of the 178 n-grams, we identified the target group and general category of hate speech. We found 10 target groups (*Muslim, Black, Asian, Homosexual, Rage, Women, Hispanic, Mexican, White, Overweight*) and 3 target categories (*Racism, Rage, Homophobia*). Figure 5 shows the frequencies of target groups and categories for all 178 hate speech n-grams. To map from group to categories, *Homosexual*-targeting n-grams were mapped directly to *Homophobia*, *Rage/Women/Overweight* were mapped to *Rage* and all others were mapped to *Racism*. Note the *Rage* category is a ’catch-all’ category used to capture both general undirected hate speech as well as hate speech directed at lower frequency groups. The following are group and category mappings for 4 of 178 hate speech n-grams:

[2] `http://hatebase.org/`

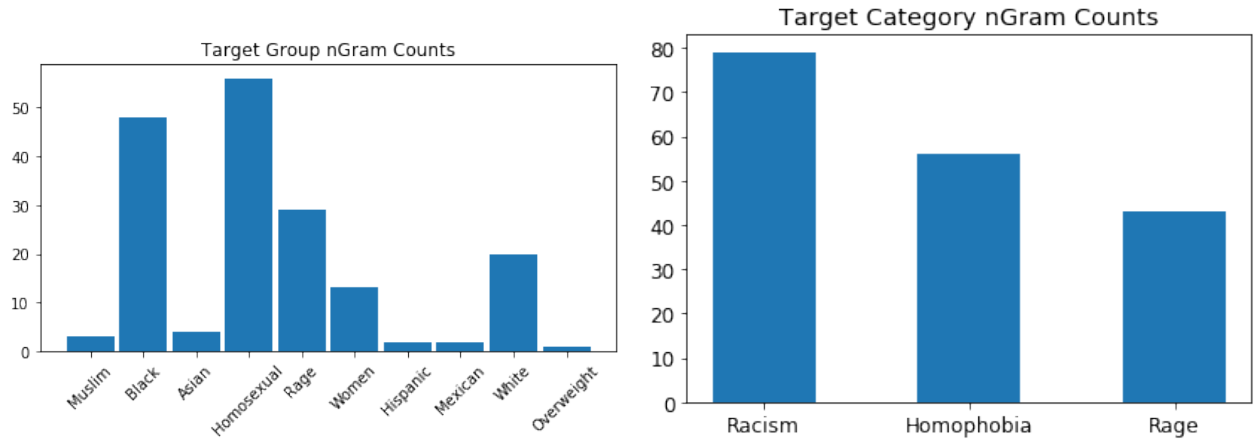


Figure 5: Frequencies of target groups and categories for 178 n-grams.

n-gram	Target Group	Category
homo	Homosexuals	Homophobia
short bitch	Women	Rage
hate all	Rage	Rage
chink eyed	Asians	Racism

Determining the target group label for a comment is a accomplished via a simple routine. We check if each of the 178 n-grams is in the comment body using the same RegEx pattern described previously. If there is a match, the comment receives a vote for the n-gram's group. For example, if an attempted RegEx match 'short bitch' returns positive, that comment receives a vote for *Women*. After all n-grams have been tested for matches, the group with the most votes is chosen as the target group label. The same process is repeated for categories.

3. Classification

Our first research question revolves around the detection of hate speech in comments. Detection is an essential task a social media platform that wants to provide an environment free of hate speech. Automatically identifying and tagging content as hate speech allows a platform to remove offending content and, if necessary, the offenders themselves.

3.1. Model

For our hate speech classifier we chose to use an RNN. There are numerous available approaches to building a classification model for language, but we decided to work with a neural-net based approach given the course's emphasis on the successes of deep learning. Our model is a mix of recurrent and fully connected layers in order to leverage the strengths of each: recurrent layers for text processing and fully connected layers for classification. We used LSTMs in order to avoid the vanishing/exploding gradient problem commonly associated with vanilla recurrent layers. Our fully connected layers are vanilla dense layers with ReLU activations. A softmax activation is placed at the end of the net in order to produce a prediction distribution over the possible classes. We placed dropout layers with $p = 0.5$ between each layer to avoid overfitting to our data. Finally, in order to incorporate more information about word meaning we are using Facebook's pre-trained fastText embeddings [3]. Figure 6 shows the final model architecture, which has has two LSTMs and two fully connected layers. The model is implemented with Keras using the TensorFlow backend.

3.2. Data

We created a new dataset by combining a random sample of comments from the 4 million hate speech dataset and a random sample of comments from the original 4.1 billion comments dataset on BigQuery. The comments sampled from the hate speech dataset were assigned labels corresponding to the category of hate speech. The comments sampled from the original dataset were all assigned the *None* (not hate speech) label as they are all assumed to be normal speech. The ratio of hate speech comments to normal speech comments in the original BigQuery dataset is less than 0.01%. Using the same ratio in our training data would yield a classifier that always achieves an accuracy over 99.9% by classifying all comments as not hate speech. We experimented with artificial ratios between 10%-50% and found that a 50-50 split between hate speech comments and normal speech comments yielded the lowest validation error despite portraying a wildly inaccurate distribution. That the training distribution heavily skews the ratio of hate speech to normal speech is favorable in our case. When detecting hate speech, it is preferable to have a model with higher recall in order to not accidentally miss a potentially harmful comment. Using all 4 million hate speech comments and 4 million randomly sampled normal comments

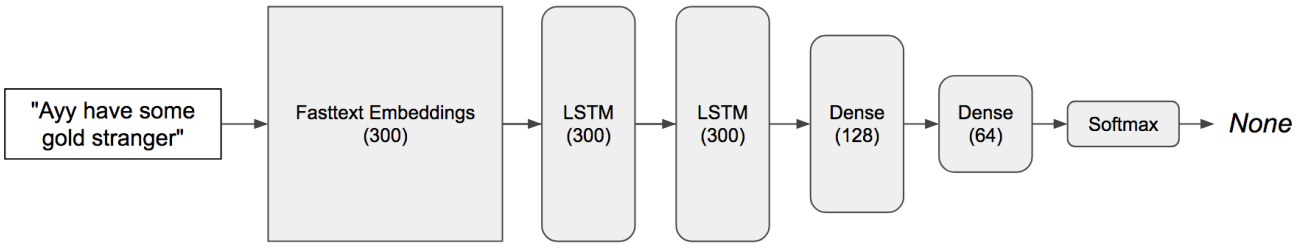


Figure 6: Comment classification neural net structure. Numbers in parenthesis indicate number of units per layer. Not included are dropout layers between all layer pairs except the embeddings and first LSTM.

was feasible[3] yet not practical given our time constraints. Each epoch would have around a day to complete. Instead, we opted for a training dataset of 500,000 comments from each dataset, which still took almost four hours to complete each epoch.

We tokenized each comment, built a vocabulary from all comments, and transformed each into a sequence of one-hot vectors corresponding to the index of each word in the vocabulary. Comments were padded by start (<s>) and end tokens (</s>). Comments longer than 1000 tokens were shortened cutoff to 1000 and comments shorter than 1000 were zero-padded. An embedding matrix with fastText embeddings was created for the vocabulary. Hate speech comments were labeled with their previously determined category (*Homophobia, Rage, Racism*). All normal speech comments were assigned the *None* label. Recall that the hate speech comments were collected by looking for the existence of 1034 and then 178 terms. The labels of each comment are also determined by the existence of the latter terms. We did not want our model to simply memorize these terms and predict a label based on whether a term exists. Instead, we want to encourage our model to learn more global properties of hate speech (and normal speech) like sentence structure and tone. To prevent the model from relying too heavily on the presence of these hate speech terms, we removed them from 50% of the hate speech comments.

3.3. Training

A 70/20/10 training/validation/test split was used on the 1 million comment training dataset. The model was trained with the ADAM optimizer (learning rate $= 0.001$) for 7 epochs and the weights with the best performance on the validation set were saved after each epoch. The model continued to improve after each epoch, but the improvements were marginal after the 5th epoch (Figure 7). The final training/validation/test accuracies were all $94\% \pm 0.5\%$. This leads us to believe that our model is very well performing and did not overfit to our training data.

[3]We sampled 20 million random normal comments for a theoretical training dataset of 24 million comments with a hate speech to normal speech ratio of 1:4.

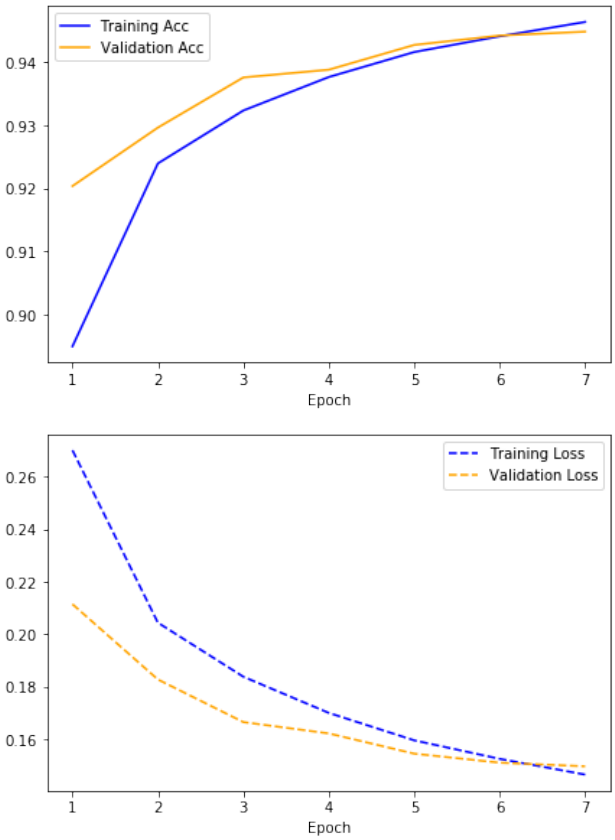


Figure 7: Accuracy and loss during training.

We experimented with less complex models before settling upon the current architecture. We started off with a single LSTM layer, added a second LSTM layer, and finally added dense layers. Table 1 shows the evaluations on the reserved test set for all three models. We did not explore trainable embeddings or a model without embeddings because we believe that using non-specific embeddings helps to generalize our model to work on unseen words.

Architecture	Params (total)	Params (trainable)	Test Accuracy
Embeddings + LSTM	207,096,004	722,404	91.1%
Embeddings + 2 LSTM's	207,817,204	1,443,604	91.5%
Embeddings + 2 LSTM's + 2 Dense Layers	207,863,044	1,489,444	94.5%

Table 1: Evaluations of the three model architectures explored to develop a classifier. Note that all use an embeddings layer and the last architecture is the one used in the paper.

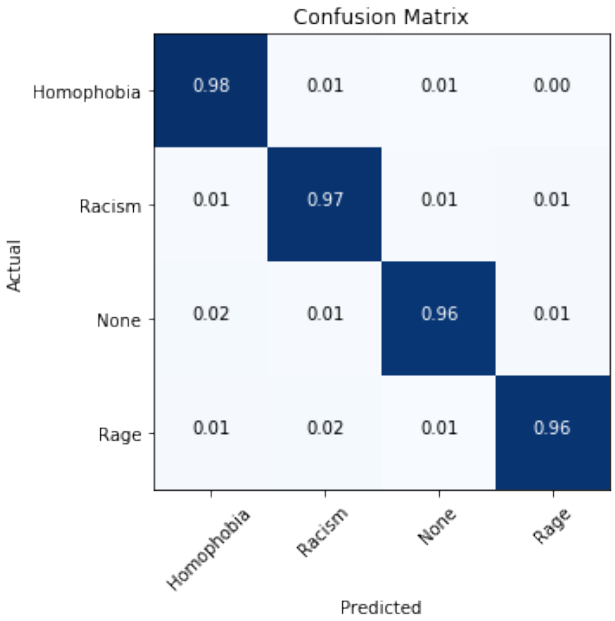


Figure 8: Confusion matrix on 2 million unseen comments.

3.4. Evaluation

Accuracy often does not reveal the entire story. To get a better understanding of the performance of the model, we evaluated it on 2 million unseen comments (50% hate speech comments and 50% normal speech comments). The confusion matrix in Figure 8 shows that model strongly agrees with the labels assigned for training. Yet the confusion matrix does not tell the whole truth. Disagreements between the model and the initial labels indicate that our model actually outperforms out the labels assigned via looking for RegEx matches. Consider the following comment:

> The only difference between the two of them is that she happened to be wearing shorts the day they recreated the picture. Doesn't really look like she's trying to be sexual. Glad we could establish that I'm a 'faggot' by not immediately thinking about her vagina.

This comment was labeled *Homophobia* via the same approach used for labeling training data but predicted to be *None* by the model. Although the comment includes the homophobic slur "faggot," the slur was being used to reference what the author had been called and not as hate speech by the author. Similarly, another comment discussing golf is mislabeled by the RegEx approach as *Racism* but properly labeled by the model as *None*:

> Tough hole. 200 yards from the whites, with around 180 needed to carry some garbage and land on the front fringe. Decently large green that isn't too slopey but you're almost always hitting into the wind, so if you don't strike it well and hit it straight, your shot might go flying off the hill to the right and end up on the 15th fairway.

Finally, we can see an instance where the comment was labeled as *Homophobia* but the model predicts it to be *Racism*:

> Hes like a gay homophobe. But hes a racist nigger.

Both the original and predicted label agree that the comment is a form of a hate speech. However, they focus on different parts of the comment. We believe that the model correctly chose the hate speech label associated with the more salient hate speech portion of the comment.

4. Comment Breakdowns

Our model allows us to accurately distinguish hate speech from normal speech in Reddit comments. We use this model to analyze the content of the largest sources of hate speech on Reddit from two perspectives: users and subreddits. In doing so we hope to answer a few questions: Do the users who are largely responsible for hate speech post exclusively hate speech? If so, do they only post hate speech for a specific category or for multiple categories? If not, what percentage of their comments are hate speech?

To answer these questions we counted the number comments per user in the 4 million hate speech comment dataset. We removed likely automated users with 'mod' and 'bot' in their names and randomly selected 100 of the top 1000 users by highest volume. All comments from the 100 random users were retrieved from BigQuery to produce a new set of 1.5 million comments. The model was used to classify the content of each comment. Finally, we calculated the proportion of each comment type and plotted the distributions of proportions per type.

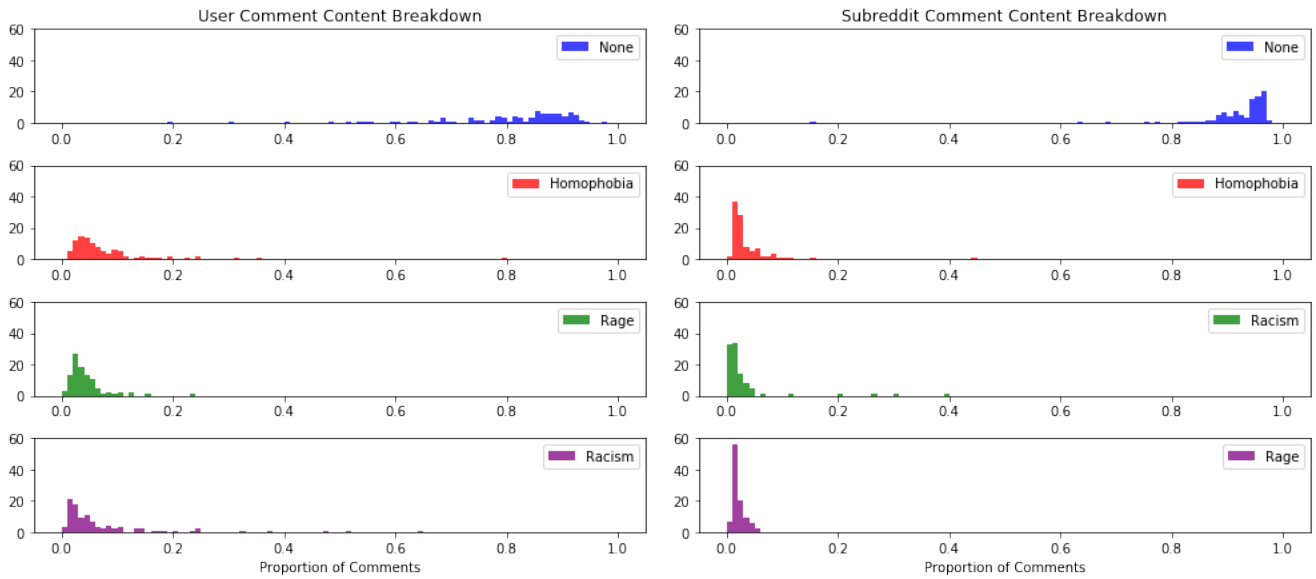


Figure 9: Comment type distributions for users and subreddits.

We performed a similar analysis on the comments of subreddits. We counted the number of comments per subreddit in the same dataset and selected the top 100 subreddit by volume of hate speech. These 100 subreddits have 1.5 total of comments of all types. It would be infeasible to download and predict on a dataset of that volume so we chose a random sample of 0.1% to produce another dataset of 1.5 million comments. The same process of calculating the distribution of comment types was applied to each subreddit.

Figure 9 shows the proportions for the 100 users and 100 subreddits. The results are very similar for users and subreddits. Hate speech generally makes up a very small portion of the overall comments written by users or to subreddits. The decline in number of users as proportion of hate speech increases is more gradual than that of the subreddit plots. This indicates that these users post higher proportions of hate speech. The distribution of normal speech comments for users is far more spread than the distribution of normal speech comments for the subreddits. The latter distribution is more concentrated at the top end and increases as it approaches 100% (but drops off around 97%). We can conclude that even though these subreddits are home to the most hate speech content on Reddit, they are far more oriented towards normal speech.

5. Time Series Analysis

The Reddit time series analysis proposed weekly aggregations of 178 n-grams within 623 weeks of subreddit comments. The aggregates were produced by by performing MapReduce on all 4 million comments in a MongoDB instance. The *map* function mapped a comment to an object with binary vector indicating which of the 178 n-grams occur in the vector. The key of each object is the timestamp of the object rounded to the first day of the week. The *reduce* function aggregated all binary occurrence vectors into a vector of how many times all 178 n-grams occurred in all comments posted during the given week. Figure 10 shows the smoothed time series for each of the 10 target groups. These trends were then aggregated into the three categories based on the previously discussed mappings. We analyze three categories over the time that Reddit has existed as well as the period consisting of Barack Obamas two terms as President of the United States. Figure 11 shows plots for both sets of time series after a 4 week smoothing window has been applied.

The analysis of this aggregation reveals that this period between 2009 and 2017 results in overall rise in racism, homophobia, and a consistent rise in rage. An exchange in the direction of hate speech is noted by a progressive rise in 2013 where incendiary commentary identified as homophobia exceeds racism. In contrast, 2015 events and rhetoric denote a substantial rise in racism and a decline in homophobia a as a focus of hate. By 2016 both racism and homophobia note a decline experience a lockstep increase. If we assess this transition beginning with 2014 up to the start of 2015, the value of the n-gram ”blacks” rises from 337 instances to a dramatic 496 The n-gram ”dyke” stays relatively constant from 63 to 66 instances, ”fucking faggot” drops to 90 from 108 instances during this particular 52 week period where racism seems to be on a particular climb.

The general term ”blacks” drops interestingly enough

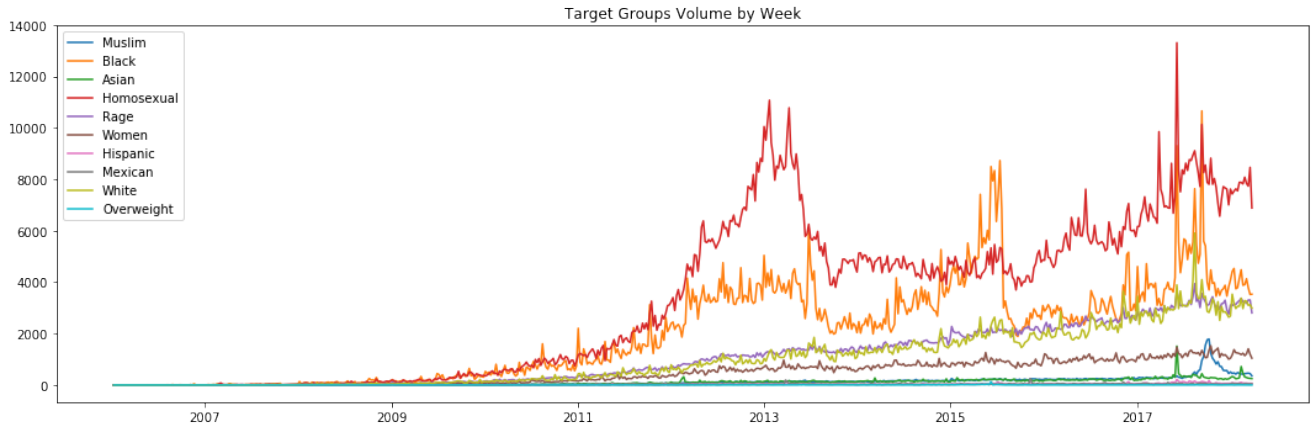


Figure 10: Volume of comments per target group per week.

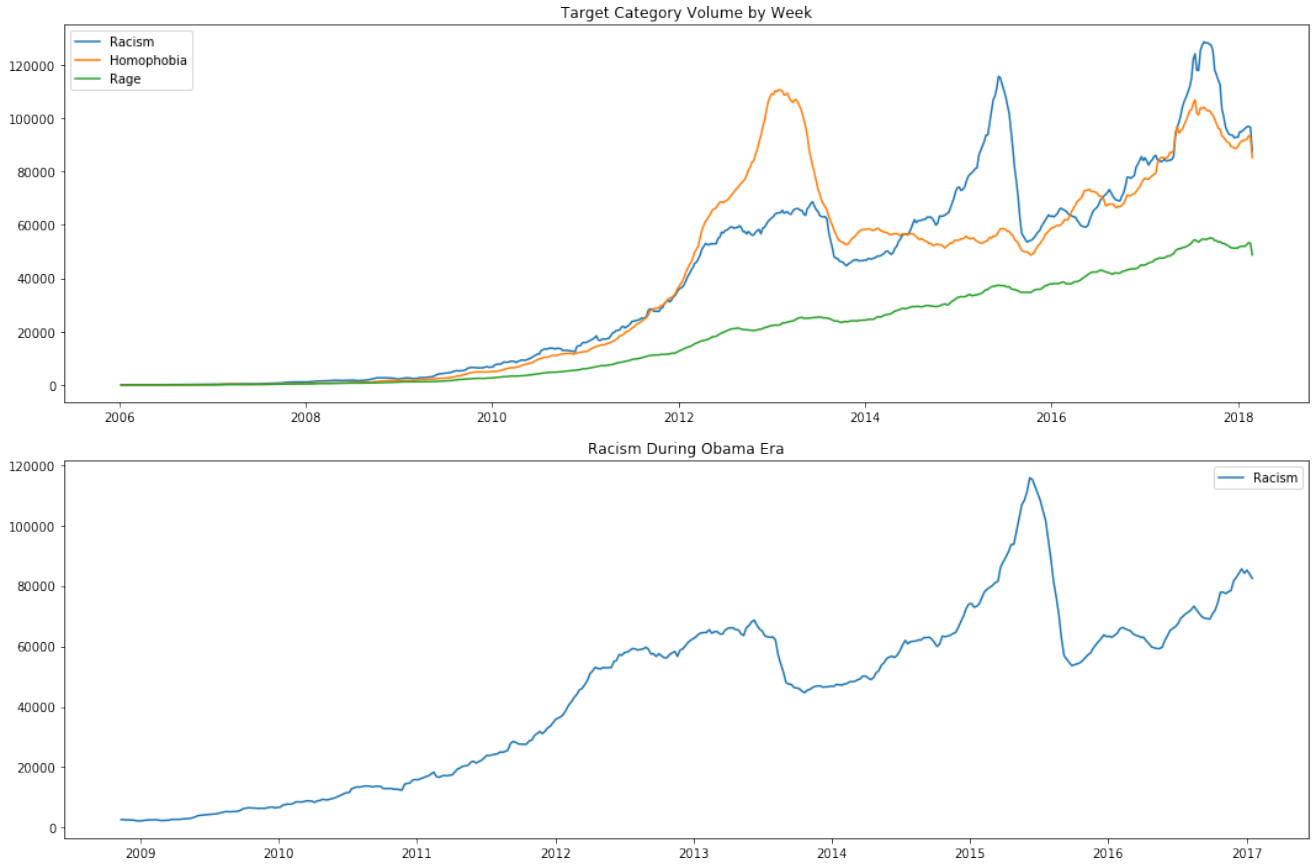


Figure 11: Above: Weekly frequency of hate speech comments per category. Below: Weekly frequency of *Racism* comments around the Obama presidency. Both plots are smoothed using a 4 week averaging window.

from 2258 instances to 1603, which may be assessed as a lack of civil reference in relation to escalating racism. This can be observed in that the n-gram ”all niggers” doubles toward the end of end of 3rd quarter 2014, October 6, 2014 to a value of 20 from an earlier start of the year January 6 at 11 instances. Related news of the period consisted of headlines such as ”Fifth Ebola patient flies home to the U.S. from Africa for treatment”, whereas additional items the same day such as ”Scenes of Exultation in Five States as Gay Couples Rush to Marry” may have been under less

scrutiny given a scourge such as Ebola; only a few days later on October 17th 2014 President Obama appointed an Ebola czar due to the gravity of the concern. It is this particular type of correlation that may be a worthwhile bit of incite to gauge mass sentiment and social impact during a particular crisis.

In terms of developing a times series prediction model, the subreddit data can be viewed as a sequence prediction problem. Hence we may select homophobia and racism as values that may be prone to interchangeability given public sentiment over a large enough media topic. This future work would build upon the patterns utilized in the analysis we have worked through. For instance many general business production problems are often viewed in this manner in that demand for a given product is seasonal, such as sales of swimming trunks dropping off during winter while scarf sales rise. Modeling these factors for freeform headline text from the NY Times API would be a close future work. Nevertheless language shifts themselves may also play a role and we do note this case in certain low counts of particular terms. Thus reviewing terms and specific decline and rise and usage in popularity of a term during a given year would be a further refinement that may reveal findings regarding term sensitivity to popular vernacular. For instance the phrase "coon shit" includes the term "coon" which is associated as a slur against blacks however is aggregated to 178 occurrences across 623 weeks. Meanwhile overall instance aggregates of the term nigger amount to 506,068. Isolating present day usage to similar terms reveals that over this 8 year span taking the n-gram "fucking nigger" compares 9,585 instances against 38,139 instances of the n-gram "fucking faggot".

Identifying hate through an aggregate of terms such as "fuck you too" reveals a subset of the continuous rise in general hate, where this particular n-gram is identified in 6871 subreddit comments. During the same period the n-gram trailer park is aggregated to 18,034 thus perhaps denoting a specific form of rage directed at economic status.

We however do not trace this analysis to reflect parallels with the general use and adoption of the Reddit platform itself and this may be viewed as an approach for further work.

6. Topic Modeling

Topic Modeling provides a convenient way to analyze big unclassified text. With the use of large amount of text data on social media, with varied length and multilingual support, this technique enables clustering of such text documents with semantic similarity and represent these clusters in few topic words. We chose to apply topic modeling to the hate speech Reddit collected by us, so as to analyze the clusters of the these posts into different hate speech categories.

Our Basic Topic Modeling approach can be represented through Figure 12. The number of topics to cluster our documents were randomly chosen in advance before using any topic modeling algorithm. Each topic is represented by a vector of words following that topic. Numerous methods have been listed in Alghamdia and Alfalqi's Survey of Topic Modeling techniques [2]. We chose LDA and NMF for our topic modeling analysis.

6.1. Dataset

Our 4.5 million data comprised of all Reddit posts that includes both hate speech and non-hate speech. Our primary focus is to bring down topics only for Reddit posts that categorizes these topics to 3 different hate speech categories of *Racism*, *Homophobia* and *Rage*. We chose 10 subreddits because they were either banned in the aforementioned Reddit hate speech ban or our analysis revealed them to be a large source of hate speech. The subreddits are (in no particular order): *r/beatingwomen*, *r/fatpeoplehate*, *r/transfags*, *r/hamplanehatred*, *r/neofag*, *r/shitniggerssay*, *r/niggers*, *r/Coontown*, *r/Physical_removal*, and *r/incels*. These subreddits provided us with a set of 75,000 samples of hate speech comments.

6.2. Data Cleaning

As a part of text processing, we observed the Reddit posts include web links and some words which would not contribute to hate speech analysis. A bag of words approach is applied where every document is represented as term vector along with the number of times a term appeared in that document. Tokenizing of words helped analyze each word independently and later with bigrams for word embeddings, to understand the closeness between the resultant topics. An important part of the processing involving the high and low occurring words, to avoid our results from frequency bias. Lastly, lemmatization of the text normalized our data for better results.

6.3. Methodologies

Tf-Idf Vectorizer: The Term Frequency-inverse document frequency vectorizer is a world level computation method to denote total number of distinct documents containing a term. We chose to fit our bunch of subreddits together in the Tf-Idf vectorizer, to find the importance of each term occurring in the document. The resultant matrix gives weights for each term over the distribution of all the documents.

LDA / NMF algorithm: We chose to implement both Latent Dirichlet Allocation(LDA) and Non-Negative Matrix Factorization to compare the results of the clustered topics. Both the methods have been widely used earlier on Twitter data and we wanted to further extend it to the Reddit. An earlier comparison [7] helped us out to check which algorithm gives us better results.

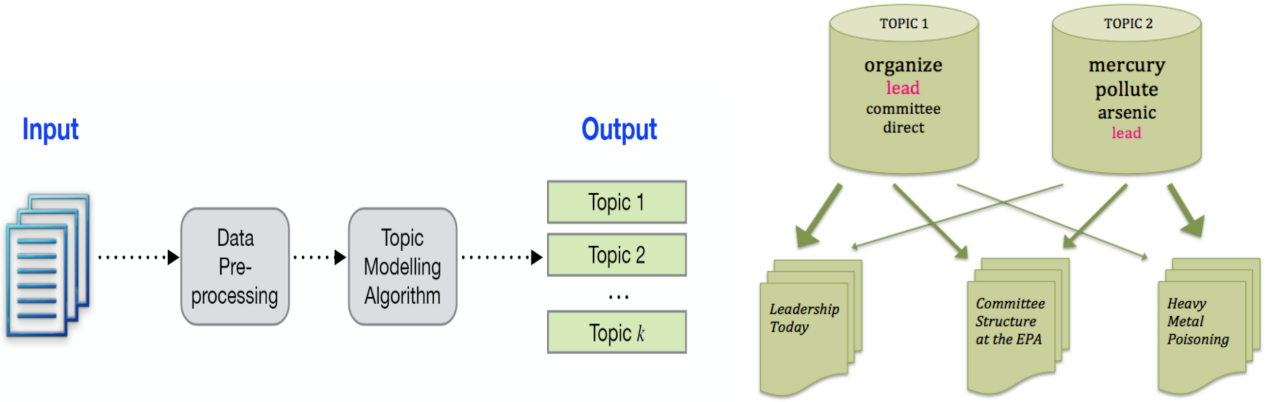


Figure 12: Topic modeling pipeline.

6.4. Results

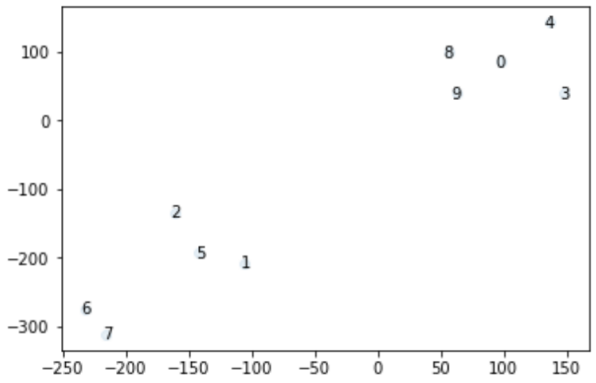
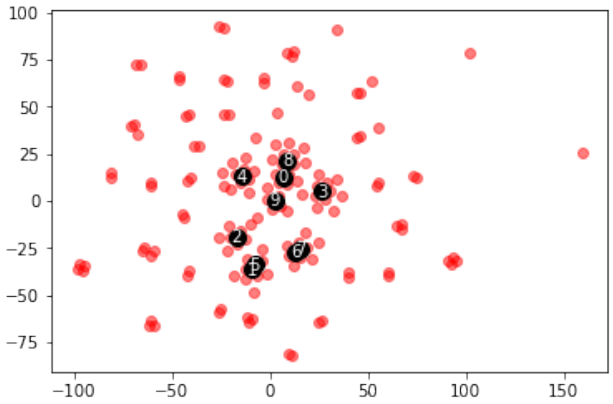


Figure 13: Topic groups visualized with TSNE.

A collection of 75k data points includes Reddit tweets. Unlike Twitter, Reddit does not have character limit and hence each post can have varied lengths. Applying only LDA and NMF on such large stream of data on social media like Reddit is not sufficient. Hence, an improvised approach of combining the Tf-Idf vectorizer resultant weights were applied, to ensure the positions of words in the documents do not affect our results, as explained in a t-LDA method in previous study [6]. Tables 2,3 show the top words from 10 topics generated using each methodology. We observed varying results for different Perplexity value in the LDA/NMF , and chose a value for $p = 2$ for best results. Further to understand the semantic closeness of all the 10 topics we clustered, word embeddings was applied to plot and analyze the distribution of these topics. A lower p value gave us better results because of the sparse aggregated words that represent each topic. These results in Figure 13 shows the 3 clusters formed by 10 topics, categorizing each to the 3 hate-speech categories: *Racism* (6, 7), *Homophobia* (1, 5, 2), and *Rage* (0, 3, 4, 8, 9).

We also observed that the results of LDA were more semantically interpretable while NMF performed 10 times faster among the two. Finally, we attempted to visualize the average embedding for each topic along with the embeddings for all component words. We performed TSNE dimensionality reduction on the embeddings for the words and the average embeddings for each topic with perplexity $p = 3$:



We can gain more insight into each cluster by looking at the 5 words nearest to the average embedding for each group:

1. Group [4]: girl, man, dog, woman, smell
2. Group [9, 8, 0]: think, say, do, really, just
3. Group [2]: faggit, faggot, r9k, phaggot, mod
4. Group [5, 1]: bitch, asshole, fucker, dickhead, fucking
5. Group [7, 6]: negro, nigger, hispanic, racist, white

6. Group [3]: guy, kid, friend, day, year

We can clearly see Group 5,1 closely expressing *Rage* (gender) and while Group 7,6 expressing *Racism*.

7. Conclusion

We have shown that although ugly and irregular, hate speech can be effectively detected and removed from the Internet. The volume of hate speech is generally not constant, rather it is triggered by popular and public events that unfortunately can spawn large magnitudes of *haters*. Furthermore, the content of online hate speech is not as sparse as one might initially assume. Instead, it generally targets a few common groups of people for a relatively small set of topics.

This project was a fantastic learning experience. It offered a lot of insight into how to tackle many problems commonly associated with social media mining like managing extremely large datasets and processing text. We are grateful to our TAs for providing guidance and valuable feedback throughout the quarter.

All code is publicly available at `https://github.com/feelmyears/reddit_hatespeech_analysis`. We are actively looking for a free way of hosting and distributing the labeled dataset of 24 million comments that we collected for our analysis.

References

[1] http://redditmetrics.com/history.

[2] A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[4] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. 1:1–22, 12 2017.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[6] L. Huang, J. Ma, and C. Chen. Topic detection from microblogs using t-lda and perplexity. In *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, pages 71–77, Dec 2017.

[7] P. Suri and N. R. Roy. Comparison between lda nmf for event-detection from large text stream data. In *2017 3rd International Conference on Computational Intelligence Communication Technology (CICT)*, pages 1–5, Feb 2017.

Topic	Top Words
0	['people', 'think', 'know', 'make', 'shit', 'want', 'good', 'thing', 'ha', 'time', 'gt', 'doe', 'work', 've', 'come']
1	['fuck', 'lover', 'shit', 'stupid', 'bitch', 'fat', 'idiot', 'cunt', 'retard', 'man', 'god', 'shut', 'piece', 'holy', 'dumb']
2	['faggot', 'lol', 'post', 'ban', 'sjw', 'gaf', 'gay', 'thread', 'beta', 'mod', 'love', 'sub', 'game', 'guy', 'little']
3	['wa', 'year', 'time', 'school', 'kid', 'tell', 'guy', 'saw', 'day', 'think', 'ago', 'friend', 'cop', 'shoot', 'old']
4	['like', 'look', 'act', 'sound', 'shit', 'guy', 'treat', 'girl', 'human', 'talk', 'feel', 'dog', 'animal', 'woman', 'smell']
5	['hate', 'fat', 'people', 'sub', 'god', 'bitch', 'jew', 'coontown', 'cunt', 'really', 'hat', 'racist', 'reason', 'gay', 'crime']
6	['white', 'trash', 'people', 'woman', 'man', 'race', 'asian', 'jew', 'men', 'kill', 'negro', 'girl', 'country', 'privilege', 'hispanic']
7	['black', 'people', 'crime', 'race', 'behavior', 'racist', 'culture', 'person', 'difference', 'gt', 'commit', 'american', 'negro', 'mean', 'population']
8	['just', 'want', 'mean', 'bad', 'll', 'wait', 'stupid', 'think', 'dumb', 'point', 'really', 'maybe', 'doe', 'typical', 'happen']
9	['say', 'word', 'racist', 'mean', 'll', 'talk', 'thing', 'hear', 'guy', 'point', 'lover', 'lol', 'comment', 'ask', 've']

Table 2: 10 topics generated using NFM method.

Topic	Top Words
0	['fuck', 'shit', 'like', 'kill', 'wa', 'day', 'shoot', 'stop', 'rap', 'watch', 'gun', 'just', 'face', 'video', 'police']
1	['white', 'country', 'asian', 'american', 'america', 'africa', 'population', 'african', 'people', 'murder', 'race', 'slave', 'racism', 'war', 'european']
2	['people', 'black', 'white', 'just', 'like', 'hate', 'racist', 'think', 'say', 'race', 'know', 'liberal', 'want', 'doe', 'stupid']
3	['wa', 'year', 'live', 'kid', 'cop', 'just', 'fight', 'house', 'come', 'city', 'run', 'happen', 'old', 'leave', 'like']
4	['wa', 'say', 'think', 'just', 'like', 'jew', 'know', 'make', 'use', 've', 'time', 'word', 'tell', 'thing', 'good']
5	['human', 'rape', 'pay', 'sjw', 'free', 'child', 'dog', 'kike', 'animal', 'attack', 'business', 'mother', 'deserve', 'mention', 'order']
6	['black', 'white', 'gt', 'school', 'ha', 'crime', 'average', 'high', 'race', 'culture', 'study', 'child', 'year', 'iq', 'higher']
7	['work', 'people', 'life', 'like', 'know', 'make', 'better', 'need', 'hard', 'think', 'care', 'just', 'thing', 'shit', 'fuck']
8	['faggot', 'fuck', 'post', 'like', 'lol', 'gt', 'shit', 'ban', 'nice', 'just', 'fag', 'game', 'feel', 'thank', 'sex']
9	['woman', 'fuck', 'men', 'fat', 'hate', 'white', 'bitch', 'like', 'girl', 'look', 'guy', 'female', 'want', 'love', 'man']

Table 3: 10 topics generated using LDA method.