Jonathan Schaeffer, Jonathan Mendez, & Jacob Sutter

Hate Speech Across Public Reddit Comments

Introduction

Our idea for our project stems from the increase in online social interactions and the rise in hateful speech, which has become more prevalent in the media. An example of hateful activities would be like the event that took place on Saturday, October 27th. On that day, the deadliest attack on American Jews took place at the Tree of Life in Pittsburgh, PA. Days before the massacre, a resident from Florida sent bombs to high ranking politicians and news sources labeled as "Fake News" by the current presidential administration. Accompanied by the massacre and bomb scare are talks about increased hate speech not only in America but throughout the world.

This topic is discussed in detail from this article by the Washington Post. Social media platforms such as reddit, twitter, and facebook have been taking measures to ban groups and individuals who push their hate speech on said platforms but has there been a general rise of hate speech across social media? A redditor has complied all the public comments across reddit from 2005 to March of 2017 which adds up to just over 300 gigs of workable data. This topic has been brought up earlier from an opinion article written by the New York Times. Furthermore, CNN correlates hate speech with social media . There has even been an academic study done on the

Equality and Freedom of Expression: The Hate Speech Dilemma so there must be some type of rise in hate speech on social media.

This led us to a question of why is there a rise, if any, and what are the impacts of hateful speech. To put an emphasis on the online aspect, we decided to use reddit as a basis for online social interactions. In the datasets subreddit, we managed to find a dataset of the reddit comment data. This ultimately led us to our analysis question of has there been a rise in hate speech across reddit comments and what are the impacts of this.

Even though the dataset we are using ranges from the years of 2005 to 2017, our team decided that there were too many holes in the data to included 2005 and 2006, so we left those values out of our analysis. As of now, our analysis only covers up to 2015 as well. This was due to the fact that when the reddit comment data is uncompressed, it is almost a full two terabytes in size. Since the focus of our study is determining if there's a rise in hate speech on reddit, we needed to remove any excess data that was irrelevant. In order to do this, we needed to parse through the data and remove anything that did not relate to the question we were trying to answer. This led to our team removing everything that was not considered a hateful comment or hate speech.

In order to determine if a comment was hateful or considered hate speech, we need to find a tool that would both determine this and provided us with a unique confidence value. Once this process was done and we deleted the unnecessary data, we then create a new csv file with the relevant data. This in turn made it much more manageable. We then used the newly created csv file and compared it to different world events that we felt would have an impact on our analysis. To our surprise, we did not see all of the correlations we expected between world events and the reddit comment data.

Some of the challenges we ran into during this final project were the lack of computing power and time needed to perform our analysis. This was primarily caused by the size of our data, which made it extremely difficult to get enough data for our team to find correlations with world events. Another challenge we are facing is our inability to really know how accurate our data is. We would ideally like to run our tests multiple times to insure the accuracy of our analysis and to ensure our work is actually replicable. With our current amount of processing power, it would be impossible to run our processes multiple times due to how long it would take. The only way we would have enough processing power is if had access to a ryzen threadripper or if we had a paid tier of the amazon aws service. Besides these minor set backs, our team believes we did the best of our ability, given the circumstances, and came out with very interesting results.

Sorting the Data

Our data was sourced from this reddit post https://www.reddit.com/r/datasets/comments/65o7py/updated_reddit_comment_dataset_as_torrents/ and includes every publicly availiable reddit comment from December of 2005 thourgh March of 2017. The comments were downloaded in a very heavily compressed .b2z format which totalled over 300 gigabytes of data. Fully uncompressed, the data reached well over 1.2 Terabytes of json files. In the early stages, our original approach was to extract from the json files the comment, author of the comment, and subreddit. Then, we would be able to keep track of not only hateful and offensive comments, but the subreddits that could possibly have the most hateful or offensive comments.

To determine which comments are hate speech or offensive language, we turned to an open source python tool called HateSonar which can be found here https://github.com/Hironsan/HateSonar. The developers behind HateSonar collected tweets which were labeled hate speech and then polled a group of users to categorize the tweets as hate speech or offensive language. The tool was written in python, so we were limited to using python as our main language. As our data got much larger with each year, we had to scale back the amount of content we were analyzing. So we decided only to focus on comments (not subreddits or authors) and only worry about identifying hate speech. This significantly cut down on file size and our processes started to run much faster.

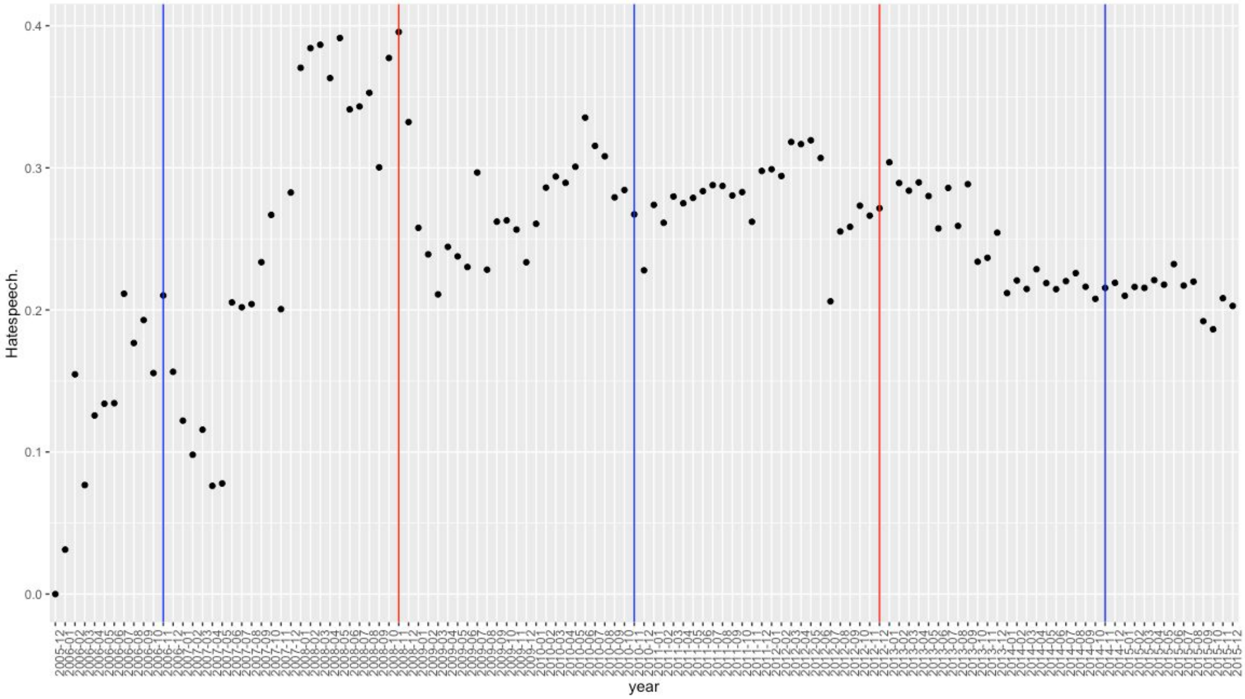
We had to write a suite of bash scripts to automate a lot of what we did. For example, cleanUpData.sh uses jq (a command line json manipulator) to extract all of the comments from our large json files into much smaller json files that only have comments in them. To understand the scale, we could take a 23 gig file and if we only extract the comments, it could top 6 gigs in size. Now, we would run the getCommentNumber.sh script that uses yet another linux command line tool to get the amount of lines in the file. This would equate to how many comments there are per file since each comment (no matter the length) only takes up one line. Now, since python is terrible with memory allocation, we had to split each large monthly file into multiple smaller 1 gig files. If we didn't do this, python would use excess of 16 gigs of ram to sort through the files. Even using packages that claimed to only read N amount of lines at a time to save on ram still didn't work so this was the only option. Speaking of python, multicore processing on python is not straightforward. You need to use external libraries that can run the same function on multiple cores. To achieve this, when a 1 gig data file was loaded into python, it was loaded as a list and then it would be split up evenly N times for the amount of cores that are available.

Now, the python program would be able to run fast enough to get through a majority of the data. Another way to help is that we cut down on the amount of I/O interactions the program made. Each write/read to/from disk would take up considerably more time and eliminating the I/O became a big help. Next step is to take the output of the data, which stores the hateful counts of each month and then import them into an excel spreadsheet where we can analyze the data and import it into R studio to manipulate it to our liking.

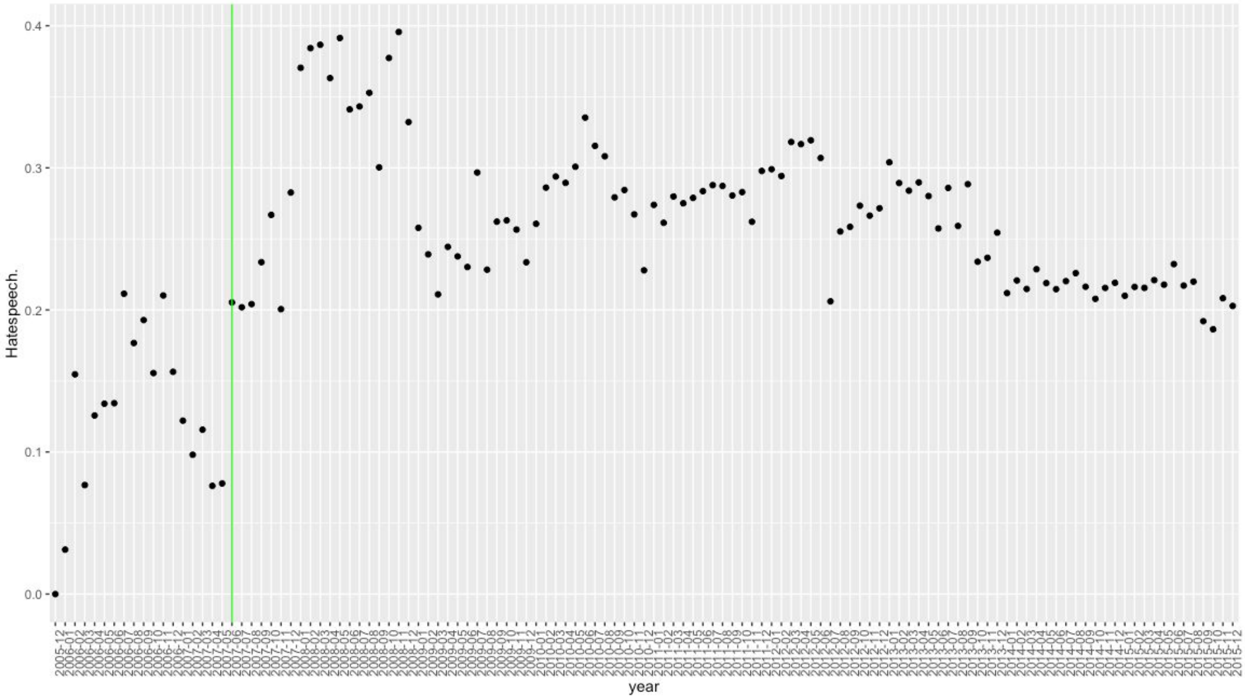
Analysis of the Data

The following graphs show the percent of reddit comments that contain hate speech per month from 12-2005 to 12-2015. Marked on each graph is a line that represents an event that occurred within the month where the line is placed on the x-axis. The goal of this is to see if certain events were either triggered by or resulted in a rise of hate speech.

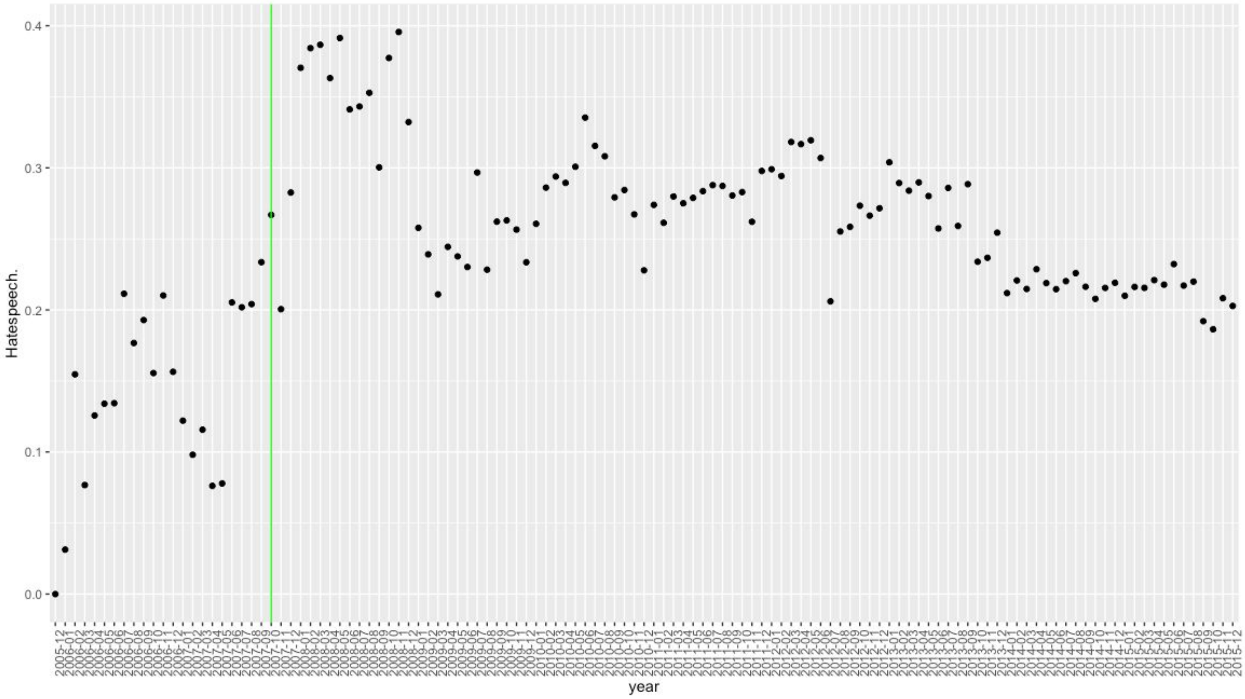
Hate speech % per month by year with election years and midterms marked



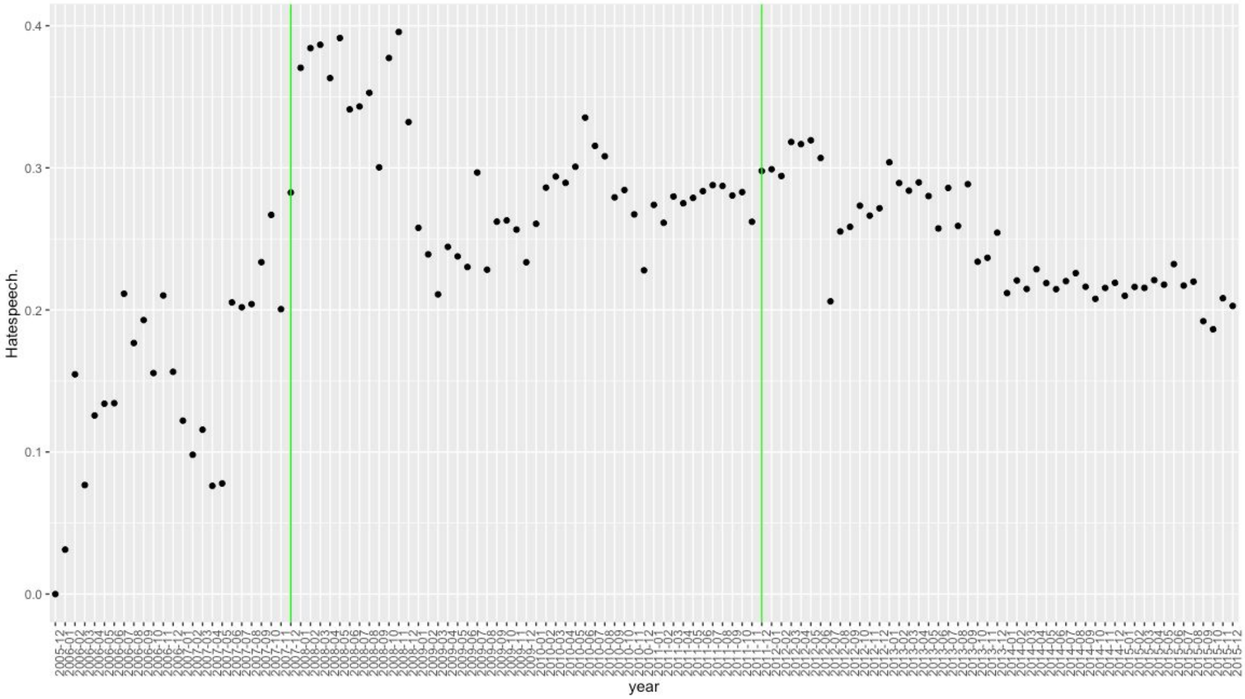
Hate speech % with iPhone introduction marked



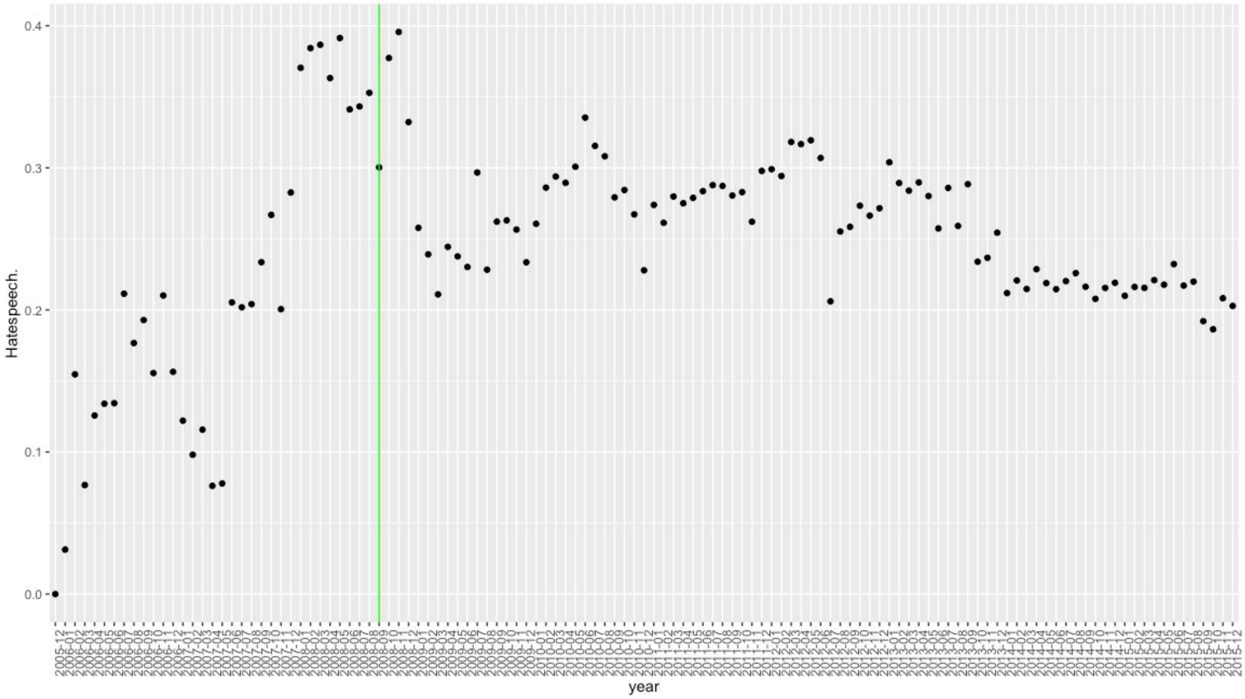
Hate speech % with Al Gore wins Nobel Peace Prize for climate change work marked



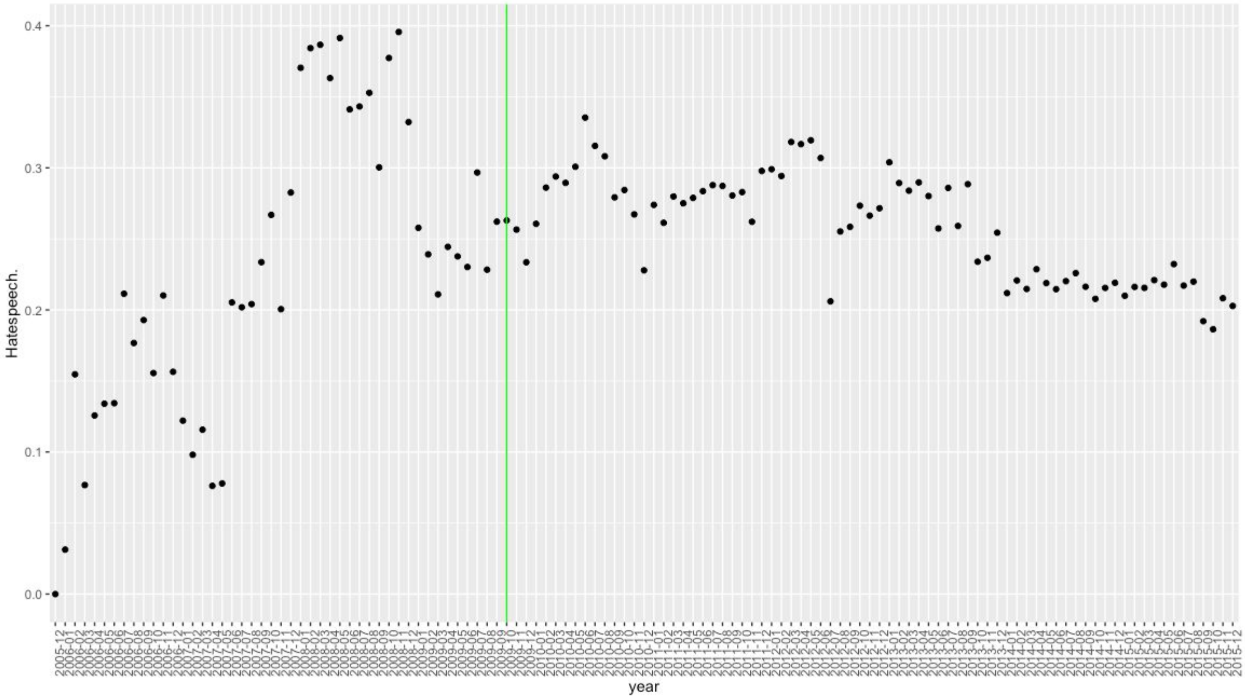
Hate speech % with Iraq troop withdraw start and end marked



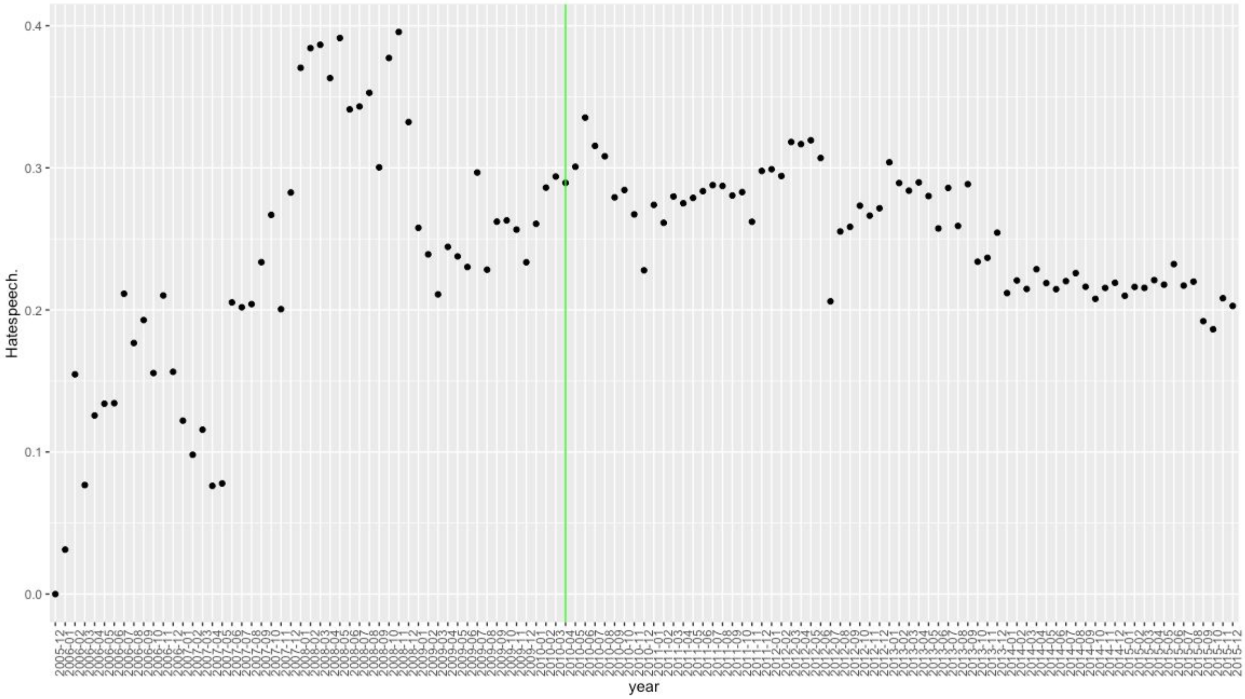
Hate speech % with stock market crash marked



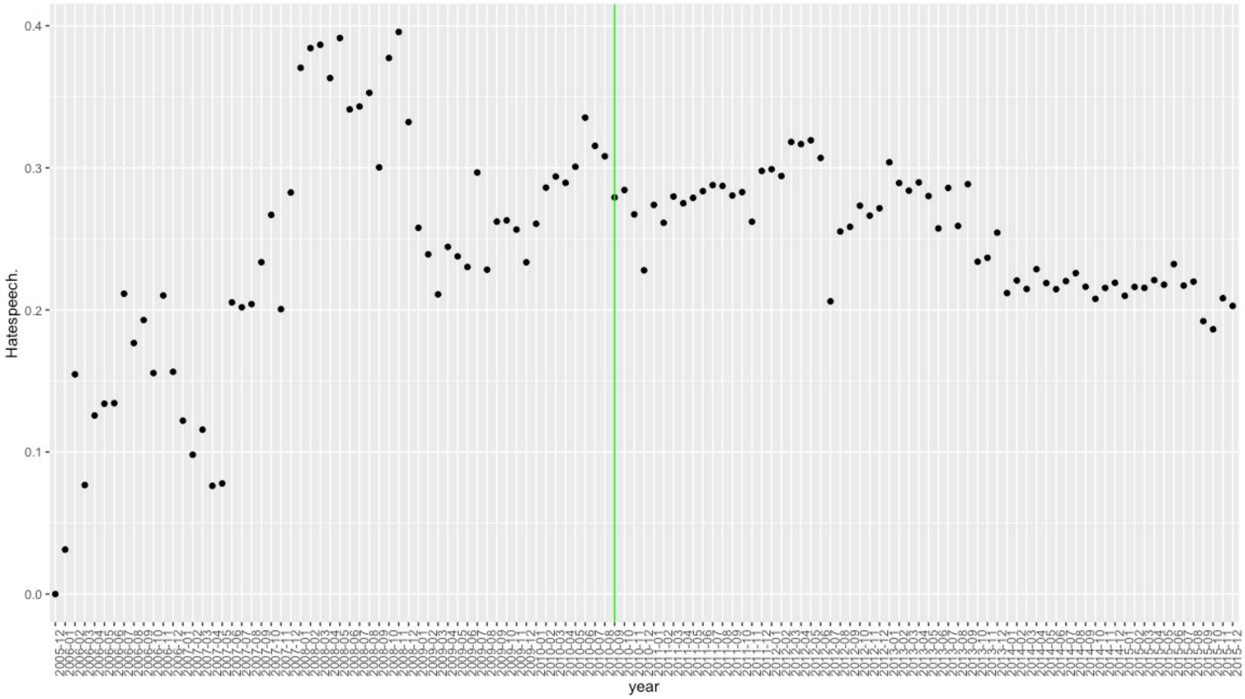
Hate speech % with US unemployment at 10% marked



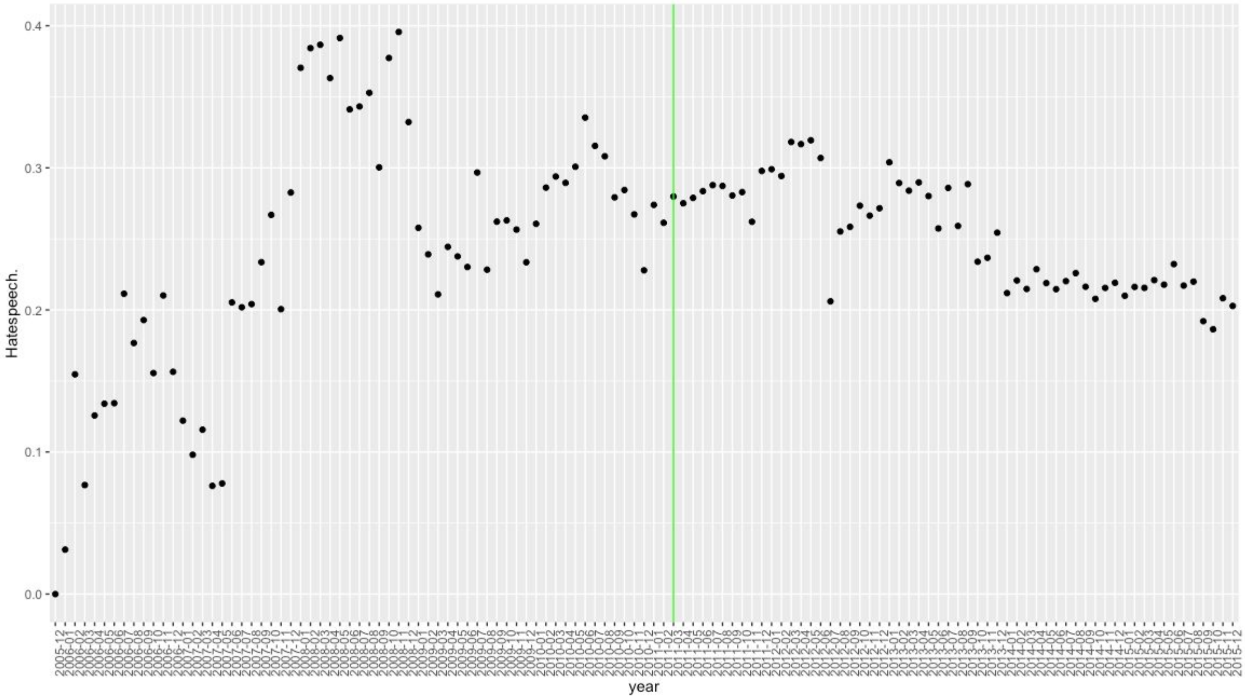
Hate speech % with Deepwater Hozizon oil spill marked



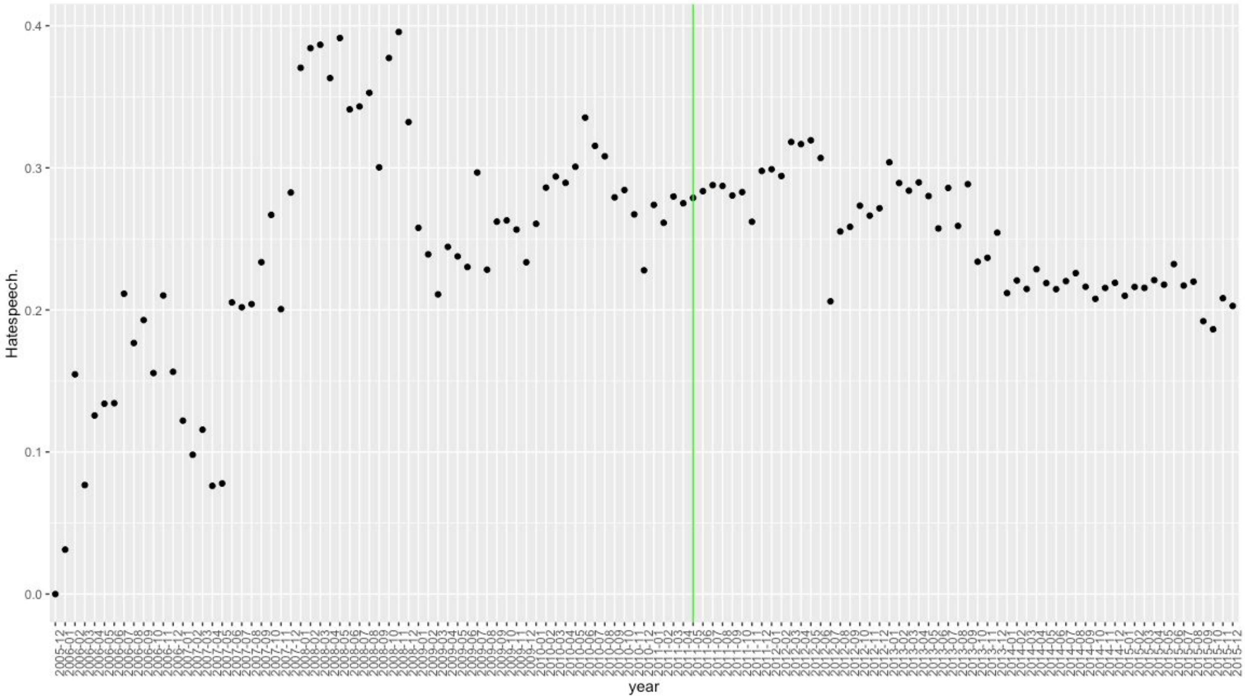
Hate speech % with Gasland anti-fracking documentary release marked



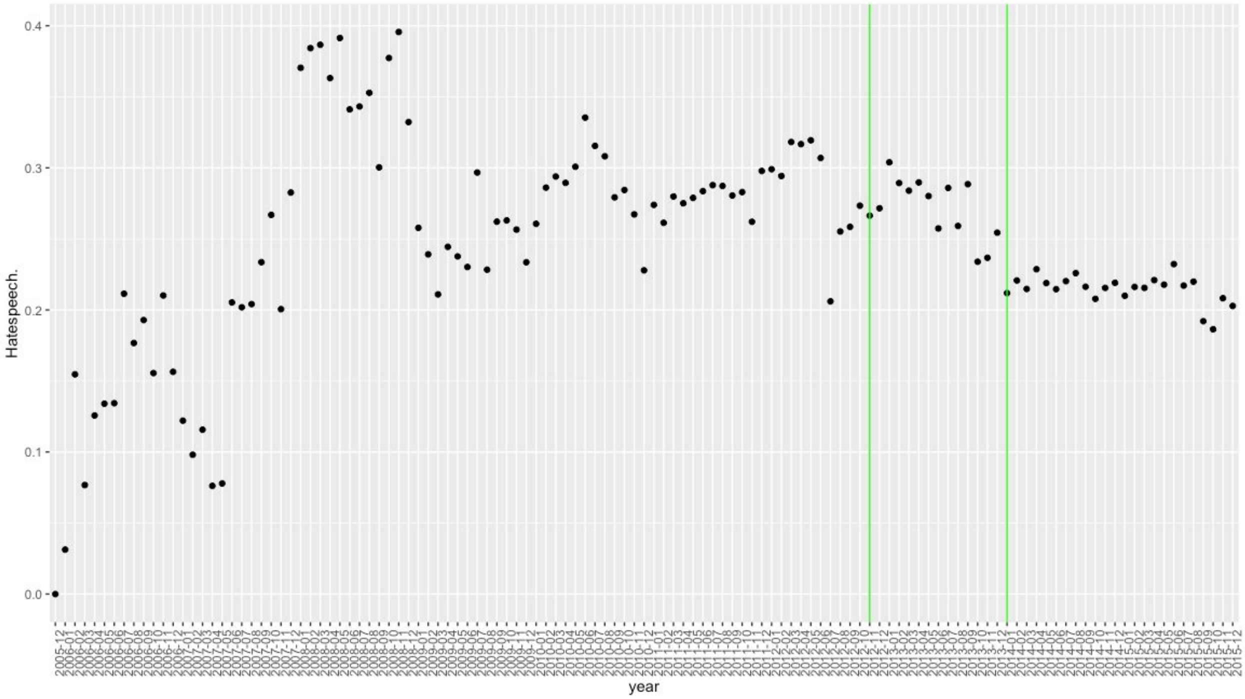
Hate speech % with start of Syrian civil war marked



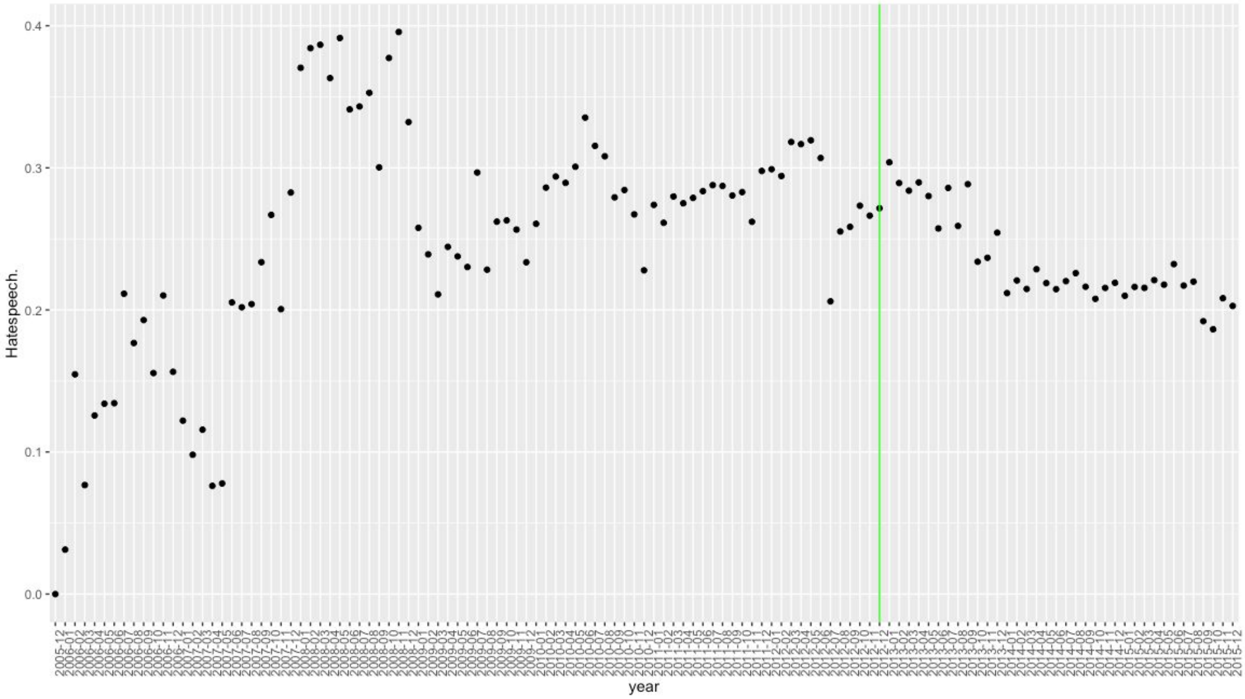
Hate speech % with Osama bin Laden killing marked



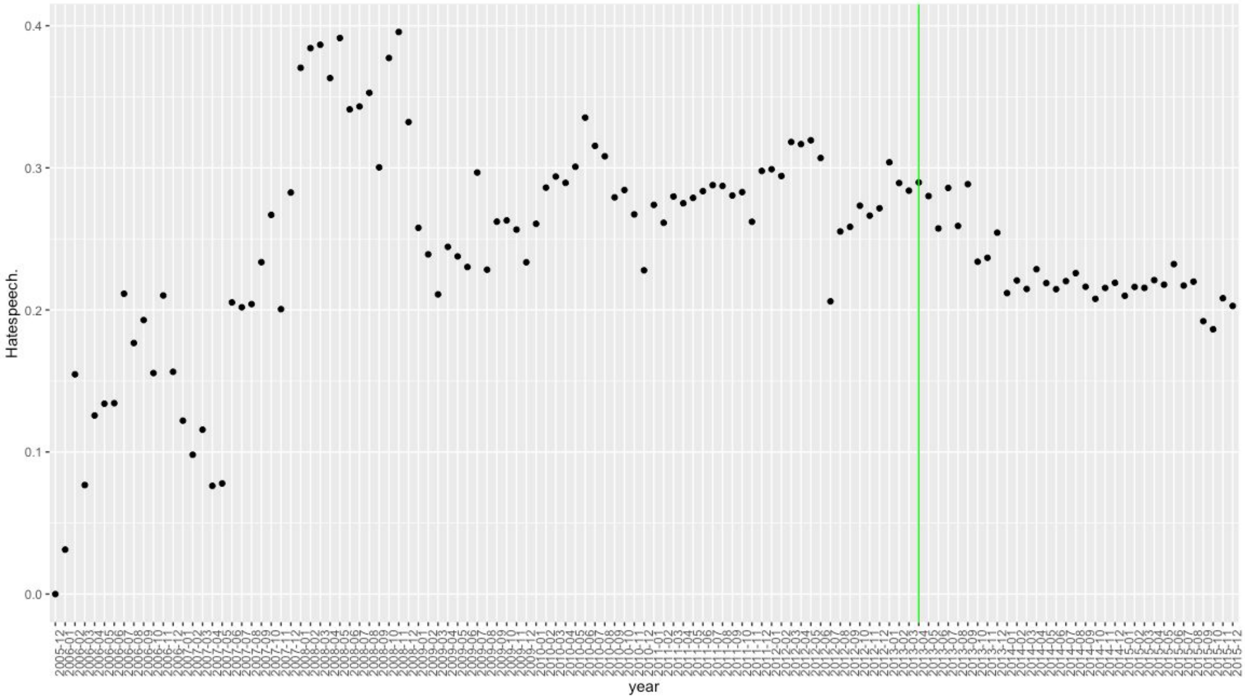
Hate speech % with US weed legislation and legalization marked



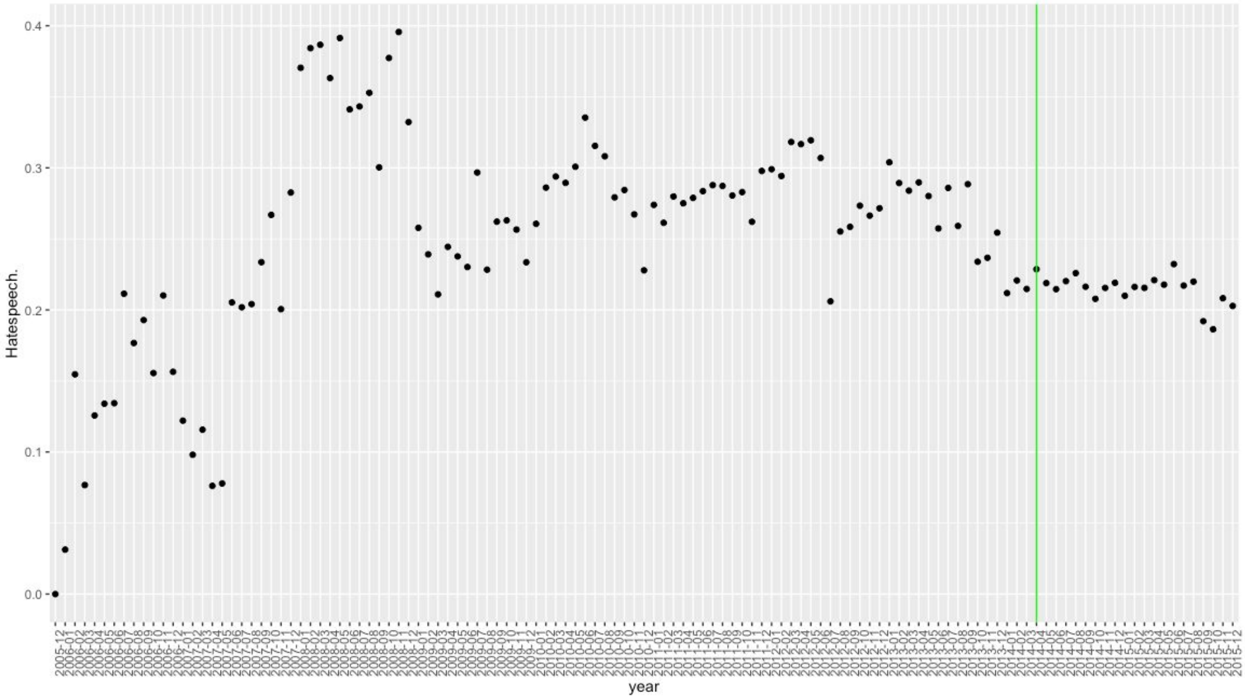
Hate speech % with Sandy Hook shooting marked



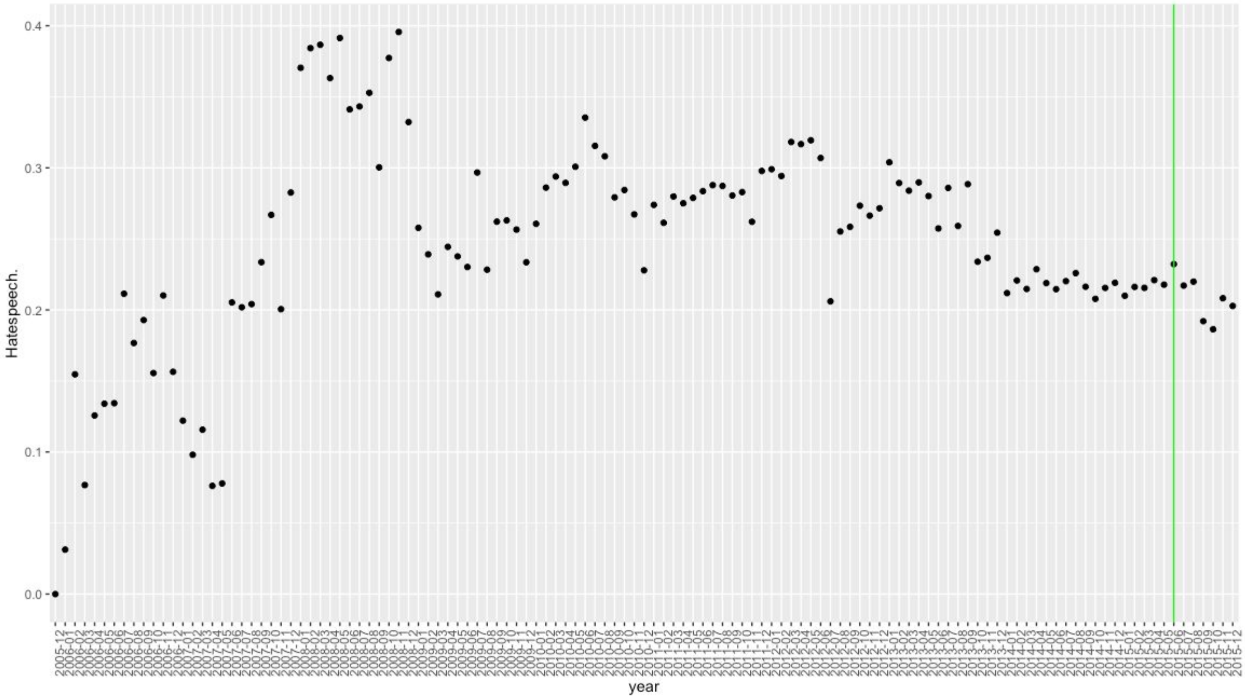
Hate speech % with Boston Marathon bombing marked



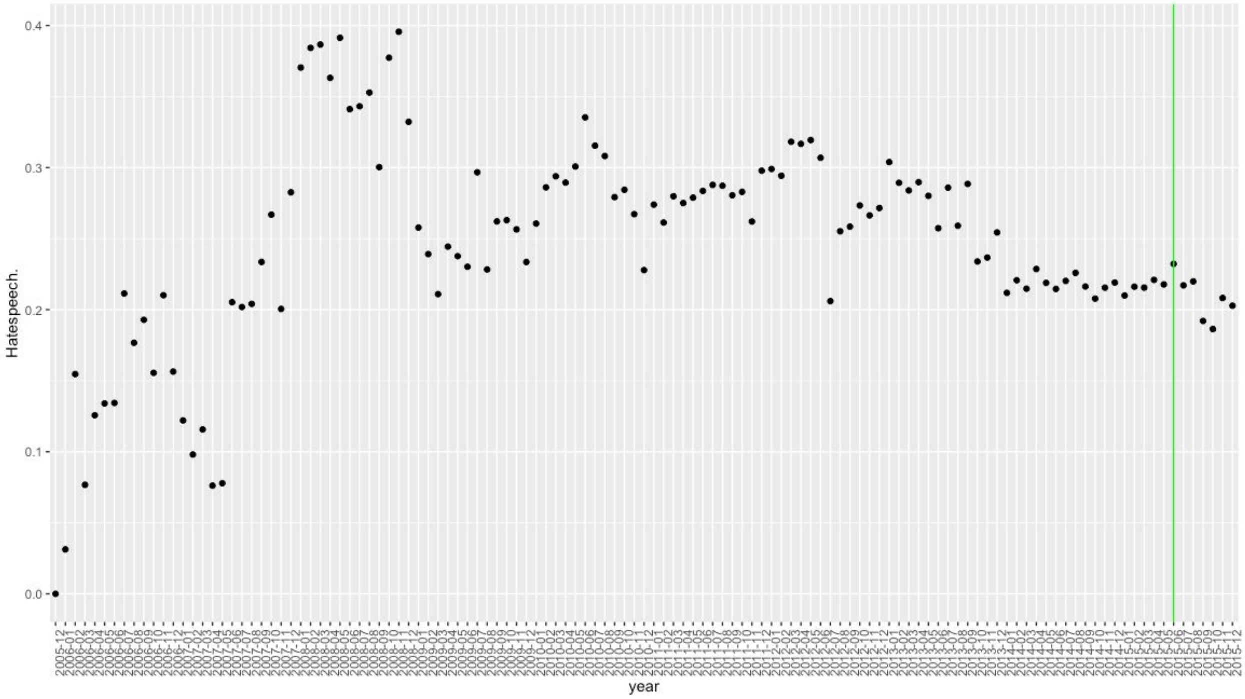
Hate speech % with ISIS offensive start marked



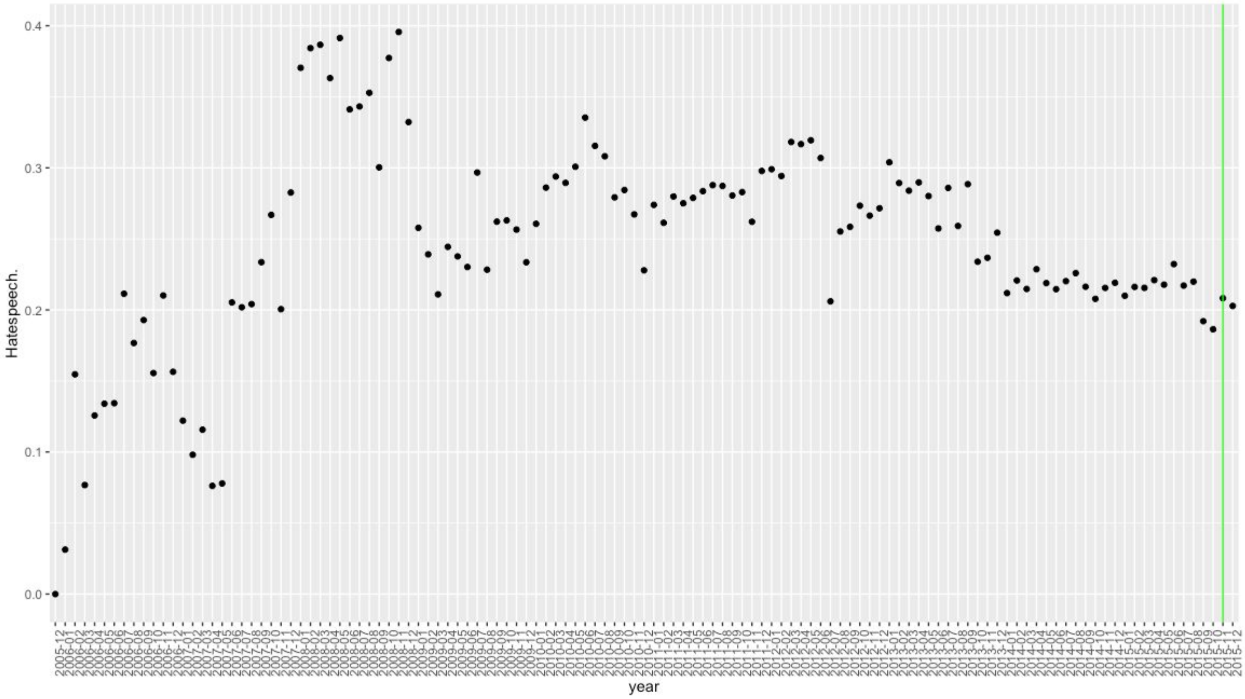
Hate speech % with Charlestion chruch shooting marked



Hate speech % with US same sex marriage ruling marked



Hate speech % with Paris terrorist attacks marked



Conclusion

Based on all these graphs, it can be seen pretty clearly that there are spikes and dips in the percent of hate speech that coincide with different events. While it is very possible that these events have nothing to do with the changes in hate speech, it is possible considering that almost all of the chosen events have some change in percentage associated with them. As part of our analysis, we also looked at offensive language. However, due to limitations with hardware and time, we only have half of the months analyzed for offensive language that we do for hate speech. In our R code, we have provided some of the same graphs that show events with offensive language, but have not included them here due to them being incomplete.

An overall trend of leveling can be seen in the data. As time moves on, the changes from month to month are less extreme. It would seem that as the reddit platform has become more popular, the percentage of people with polarizing opinions gets outweighed by more positive comments overall.