Hate-Index by States
(2015/01 – 2016/12)

Nationwide Temporal Pattern in Hate-Index

Quantifying structural factors affecting racial
tensions during 2016 US Presidential election

Bjarke Felbo, Richard Kim, Nick Obradovich, Iyad Rahwan

ABSTRACT

Throughout the 2016 US Presidential election cycle and the months following the election, many media outlets reported increased in racial and religious hate speech in the US on social media platforms. We hypothesize that the rise in racial tensions was in large part due to preexisting structural factors. Using geo-tagged tweets generated between January 2015 and December 2016, we quantify a measure of racial tensions that we use to analyze the connection between such tensions and economic/demographic variables.

The contributions of our research are threefold:
1. We show how advanced deep learning algorithms can be applied to identify hate speech in 1.5 billion tweets.
2. We quantify the measure of hate speech in an important political time period.
3. We analyze the relationship between structural variables and racial tensions at a high spatial resolution.

METHODS

Once trained, we use the DeepMoji algorithm to identify hate speech in our subset of candidate tweets.

In order to address the prevalence of bot accounts, which contribute disproportionately large number of posts with advertisements, news, and weather announcements, we also train a random forest classifier to identify the bot accounts. We compile a training data set of 1000 Twitter accounts labeled by Mturk workers. After training, our bot detection algorithm achieves accuracy of 98% and F1-score of 0.6.

To analyze temporal and geospatial patterns of hate speech in tweets across the country, we introduce hate-index:

\[
\text{hate-index}_t = \frac{\# \text{ of hate speech tweets at period } t}{\# \text{ of all tweets at period } t}
\]

Hate-index allows us to quantify the level of hate speech present in Twittersphere and track its variation across time and geospatial dimensions. Whether or not hate-index can be used as a proxy measure for racial tensions in a broader structural analysis is left open for exploration; nevertheless, we use it as a measure to capture the racial tensions in Twittersphere during the 2016 US Presidential election cycle.

DATA

From a database containing over 5.0 billion geotagged tweets, we isolate tweets generated between January 2015 and December 2016 within the US. We select January 2015 to December 2016 as appropriate months to capture Twitter activities for the US Presidential election and the Democratic and Republican primaries.

Using HateBase.org API, we compile a list of race related hate words. From this list, we identify tweets that are candidates for racialized hate speech. As some speech containing these words are not necessarily hateful depending on the context, we use a machine learning algorithm to identify at scale whether the candidate tweets indeed contain hate speech.

MODEL

We use DeepMoji, a pretrained deep learning algorithm that has demonstrated success in accurately identifying complicated displays of emotion in texts across multiple domains.

We employ Amazon Mechanical Turk (MTurk) to have human workers identify and label 5,000 candidate tweets.

After training, the algorithm is able to accurately identify hate speech in tweets with F1-score of 0.62.