HateCorp: Cross-Platform Data Collection for Online Hate Groups

Presenters: Shruti Phadke Rohit Kumar Chandaluri
Course: CS6604 Digital Libraries Fall 2019
University: Virginia Tech
City: Blacksburg, VA, 24060
Date: 12/05/2012
Motivation
Social media can provide an efficient and fast communication for hate groups to exchange information and spew radical beliefs and activism, amplifying what are otherwise fringe opinions.

Why the Holocaust Story Was Invented

Why the Holocaust Story Was Invented - TBR writer John Wear has been asked the questions: “Why was the Holocaust story invented? Who benefits from this falsification of history?” This article will answer these questions. By John
barnesreview.org
In a mounting number of hate crimes against various minorities, perpetrators sought support and publicized their actions on various social media platforms.

Bryan Fischer
@BryanJFischer

Right Wing Watch helps me get word out: if we ban Confederate flag, should ban LGBT flag for same reasons.
Church shooting at Charleston, synagogue shooting in Pittsburgh and New Zealand mosque shooting are just a few of the many incidents that have recently reinforced concerns about organized hate on social media.

To this date, the research community has not developed a strong understanding of how hate groups use social media to frame their messages and share information.
Research Questions

How do hate groups use social media?
How do hate groups use different social media spaces?
HateCorp : Digital Library of Cross-Platform Hate Group Data

Research Task 1: Creating a dataset of hate group communication across multiple platforms

Research Task 2: Observing various linguistic, informational and social engagement trends in the cross-platform data
SPLC (Southern Poverty Law Center) is a non-profit social justice organization dedicated to monitoring hate group activity in the United States.
**ACT for America**
Brigitte Gabriel, born Hanan Qahwaji, claims ACT was launched as a response to the 9/11 attacks and “educates citizens and elected officials to impact policy involving national security and defeating terrorism.” ACT has stayed true to its mission by working to advance anti-Muslim legislation at the...

**Alliance Defending Freedom**
Founded by some 30 leaders of the Christian Right, the Alliance Defending Freedom is a legal advocacy and training group that has: Supported the recriminalization of homosexuality in the U.S. and criminalization abroad Defended state-sanctioned sterilization of trans people abroad Linked...
We focused on following 5 ideologies.

Along with the names of the hate groups, SPLC also signifies the hate ideology they identify with.

SPLC (Southern Poverty Law Center) is a non-profit social justice organization dedicated to monitoring hate group activity in the United States.
Mapping Social Media Accounts Across Platforms

SPLC contains 367 hate groups list with their ideologies.

We needed to identify social media accounts of hate groups across different social media platforms.

Manually searched for the accounts existence and verified the websites content with the ideology of the group.

For every organization we searched on
- Twitter
- Facebook
- YouTube
- Instagram
- Pinterest
## Accounts and Data Points [1 April 2019 - 1 Oct 2019]

<table>
<thead>
<tr>
<th>Accounts</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>🎥 Twitter</td>
<td>98</td>
</tr>
<tr>
<td>🎥 Facebook</td>
<td>108</td>
</tr>
<tr>
<td>🎥 YouTube</td>
<td>92</td>
</tr>
<tr>
<td>🎥 Instagram</td>
<td>17</td>
</tr>
<tr>
<td>🎥 Pinterest</td>
<td>3</td>
</tr>
<tr>
<td>🎥 Other</td>
<td>20</td>
</tr>
</tbody>
</table>
Collected public tweets posted by the hate accounts using HTML page scraping.

Our code was inspired by the GitHub repository of Henrique Jefferson.

CrowdTangle API
Social Science Research One
Grant
Google YouTube API

25K videos

621K comments
Trends

Linguistic

Informational

Social Engagement
Indexing collected data into Elasticsearch

- Elasticsearch is based on Map reduce concept with master and slaves to work with big data.
- Elasticsearch is mostly used as a search engine.
- It also provides various analytical tools that work on big data.
- Elasticsearch indexes all the content that we pass it in a JSON format.
- Text cleaning
  - To enable word-based visualizations, we first need to clean the text of all punctuation, stopwords and inline URLs.
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>The name of the hate organization as obtained from the SPLC website</td>
</tr>
<tr>
<td>Ideology</td>
<td>Ideology of the organizations as described in the Background section</td>
</tr>
<tr>
<td>Text</td>
<td>Text of the message cleaned using the steps explained above</td>
</tr>
<tr>
<td>Links</td>
<td>URL domains extracted from the links embedded in the text</td>
</tr>
<tr>
<td>Replies</td>
<td>Number of replies on Twitter and the number of comments on Facebook and YouTube</td>
</tr>
<tr>
<td>Reactions</td>
<td>Number of likes for tweets and Facebook posts, and upvotes for the YouTube videos</td>
</tr>
<tr>
<td>Created UTC</td>
<td>The epoch timestamp in seconds for every post</td>
</tr>
</tbody>
</table>
Word Count Visualization
Most used word usage in each ideology visualization
Reactions count over time in each ideology visualization
Advanced Linguistic Trends

Sparse Additive Generative Models of Text [SAGE]

- Unlike Multinomial Dirichlet distributions, SAGE distributions are added in logarithmic space.
- AGE is especially useful in constantly evolving datasets—such as this—where a significant portion of the probabilities in LDA might not be well-calibrated.

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>Twitter</th>
<th>YouTube</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Supremacy</td>
<td>prison, liberals, stupid, negro,</td>
<td>racial, whites, race, difference,</td>
<td>monthly, pledge, buchanan, wilson, steve</td>
</tr>
<tr>
<td></td>
<td>nazis</td>
<td>superior</td>
<td></td>
</tr>
<tr>
<td>Anti-LGBT</td>
<td>prayer, join, jesus, god, abortion</td>
<td>predator, pedo, church, life,</td>
<td>militant, church, join, subscribe,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>homo</td>
<td>media</td>
</tr>
</tbody>
</table>
Advanced Information Trends: Domain Networks

Connect two URL domains if they are shared by the same account

Color the edges based on where they are shared

**Blue**: Only Twitter

**Red**: Only Facebook

**Green**: Both
Religious Supremacy Domain Network
Future work

A real-time data collection pipeline can be built using Kibana and Elasticsearch to update the visualizations created every 15 minutes. This can give an overview of the latest trends in hate group activity online.

More complex machine learning models can be incorporated with the Kibana visualizations to have complex analyses such as SAGE and domain graphs in near real-time.

By adding differential privacy to the data, this platform can be made public to be used by other researchers and enthusiasts.
THANK YOU