Project: Hate Speech Detector
CS 410 Fall 2018
10/21/2018

Team Members:

Project Coordinator	Project Member	Project Member
David Jamriska	Ayan Putatunda	Chien-Yi Chen
Djj3@illinois.edu	ayanp2@illinois.edu	chienyi2@illinois.edu

What is the use case?

Hate crime is growing in the United States and law enforcement is devoting ever increasing resources to track and investigate hate crimes. Recent experience has shown that hate crime actors use social media and website comments to promote their ideals. Currently law enforcement organizations (LEOs) manually search and review the web for these comments and attempt to identify individuals or organizations involved in coordinated hate promotion. LEOs would be better served to have an automated tool to perform these searches. An open source tool would help organizations with tight budgets better serve the community.

What is the function of the tool?

Our project specifically will address the hate and threatening comments posted in the comments section of online newspapers. The tool will consist of the following sections:

- Article Filter – This is a two-part deliverable, first the ability to define article keywords. These keywords will be used to scan a website for key words in article titles to identify which should be scanned. For example, the system should not scan an article about housing prices, but should scan one related to an assault or attack. The second part is the scanner that scans the websites and builds a list of articles (URLs) for the comment extractor to process.
- Comment Extractor – This portion of the tool also consists of two parts, the first with a list of keywords that trigger a extraction of the comment. The second part is the extraction tool. Our tool will focus on news sources that use the Disqus commenting tool. The keywords can also have a weight applied to them.
- Search Results Presentation – Once the articles and comments have been evaluated those results will be stored in a database and a list of articles and comments will be presented to the user for review. The initial results will return two views for the user, a ranked list of results by article. The second will be a list of comments by user profile.
- CFR28 Purge – Law enforcement must comply with many rules, in this case we will have to respect and implement a CRF28 purge tool that will purge comments not specifically marked as being under investigation.

Who will benefit from such a tool?

While initially intended for LEOs other non-profit organizations can also use these tools. Although not planned, the tool could be configured to also target other types of speech such has child exploitation or threatening speech against specific individuals. Businesses could also use this tool to determine if their brand is being promoted or attacked as related to an article. A review of most newspapers when posting an article about a sensitive topic will be filled with inappropriate comments. After some review you can see patterns of the same people.

Does this kind of tool already exist? How is your tool different from them?

There are some existing discussions and tools in the literature e.g. Ref. [1,2]. However, in our work we intend to focus not only on classifying hate and non-hate speech but also on tracking the sources of messages to determine whether they are directly used to a person or group of people targeted. Moreover, we want to identify the pattern of the speech and the source of the people. This would involve developing method to distinguish between different uses and look more closely at the social context and conversations in which hate speech occurs. We also intend to study more closely the people who use hate speech focusing both on their individual characteristics and motivations and on the social structure they are embedded in.

What existing resources can you use?
The team has identified a number of existing Python libraries that deliver portions of the extraction and web scrapping tool. These resources will allow the team to focus on the search result delivery.

The team has also established a relationship with a local law enforcement organization that will assist with testing and validation of the extract and ranking functions.

What techniques/algorithms will you use to develop the tool?
The team is currently exploring building a custom version of the BM25 tool, adding in a special ranking for keywords. We don't want to list the keywords, but some more offensive / threatening words will drive comments towards the top of the list.

How will you demonstrate the usefulness of your tool?
We have already had discussions with a top 10 law enforcement organization that will asking for this specific tool and has agreed to be a tester/validator. Brief conversations with additional large scale organizations have also expressed interest. In being involved.

For the class presentation we will prep a list of keywords for both articles and comments and build a video showing the tool in action.

A very rough timeline to show when you expect to finish what.
The team expects to begin design and development on 10/22

Activity	Expected Delivery Date	Status
Build Project proposal	10/21/18	Complete
Validate extraction tools	10/25/18	In Process
Database Design	10/28/18	Not started
Prototype UI	10/30/18	In Process
Article Keyword UI/DB	10/30/18	In Process
Article Keyword Scraper	11/15/18	Not started
Comment Keyword UI/DB	11/5/18	In Process
Comment Keyword Scraper	11/20/18	Not started
Search/Ranking Tool Design	11/15/18	Not started
Search/Ranking Tool UI	11/25/18	In Process
Testing with End Users	11/28/18	Not started
CFR28 Purge (*maybe)	12/2/18	Not started
Final Testing	12/5/18	Not started
Presentation	12/8/18	Not started
Package and deployment	12/8/18	Not started

<u>In Scope for Project</u>

1. Filtering words to control which articles are searched. If the top level article does not contain one of these words, or a word it would not be searched. (static search terms)
2. Ability to enter limited term search words with start and end dates, for example related to a current event, topic or person. This would be a list of constant keywords that would trigger a review. (dynamic search terms, think current events)
3. Ability to search for specific words or phrases in comments sections of websites based on the articles flagged in #1.
4. Information presented/stored should include link to original article, title of original article, publication, date.
5. Information stored for comments requiring review include poster, links to post, content of post, user name, date discovered, link to original article, contents of comments.

Stretch Goals (probably won't be done, but would be great)

1. Ability to flag comments as keep or delete
2. Ability to view all comments for a profile
3. Information must be purged in accordance with Federal laws for example, CRF 28 Part 3.

Future Goals (out of scope)

1. Analysis of multiple comments across sites to try to link comment profiles.
2. Ability to review comments that appear to be similar showing coordination
3. Ability to manage the searches sources, websites, including structure of how the comments are stored. (hard coded for now)

References

[1] **Automated Hate Speech Detection and the Problem of Offensive Language**. Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber. https://arxiv.org/pdf/1703.04009.pdf

[2] **Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter.** Ziqi Zhang and Lei Luo. https://arxiv.org/pdf/1803.03662.pdf