Monitoring and Visualizing Hate Crime in the News

Victor Ardulov
Aida Mostafazadeh Davani
{ardulov, mostafaz} @usc.edu

1 Introduction

In recent years Hate Crimes have come to dominate a large portion of the socio-political discourse occurring in the world. Findings from the past year indicate that hate crimes are on the rise in the United States, and the United Kingdom. It is also beginning to arise as a more severe issue in countries which previously underreported these events. To better understand and mitigate violence and discrimination, as well as, inform policymakers and emergency services responding to these acts, we find it necessary to have a unified and structured system for inspecting these types of events throughout time. As such, we present a methodology for using the semantic web and knowledge-graph technologies for collecting news event data, constructing a unifying news story ontology, and linking stories from numerous source into unique news event identification.

Our presented work allows for better indexing, monitoring, and visualization of hate crimes in the media. Integrating information from numerous structured and unstructured sources these methods lend themselves to the thorough detection and analysis of events worldwide. In this report we present the details of how our system was implemented, defining the methods and decisions made as we constructed a unified news story ontology, and a unique news event identifier and exposure rating mechanism to better understand the reach and coverage of specific news stories.

2 Background

The FBI defines "Hate Crime" as:

> Criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity

and tracks hate-crime annually based on state reports from local police departments. One major issue is that these local townships heavily under-report hate crime to the federal level. As a result foundations and non-profit organization such as *Propublica* have set up small databases logging hate-crime news on different news outlets.

Previous work on hate-crimes have mostly used sociological and psychological analysis to understand the under-pinning that motivates and drives perpetrators to commit these crimes. However most of these endeavors have not attempted to computational model or predict indicators of these crimes.

Recently, strong efforts have been applied to social media platforms to better identify and extract "hate-speech" in text. In a similar study, moral polarity in Social Media discussions have been used to predict violence outbreaks during protests.

In the field of automated news story processing, EMBERS was a recent attempt to build a 24/7 continuous system for forecasting civil unrest across 10 countries of Latin America using open source indicators such as tweets, news sources, blogs, economic indicators, and other data sources.

Finally, most closely related, Cardiff University was recently awarded $500,000 for research in Los Angeles Hate-Crime detection and prevention using Twitter indicators.

3 Data

When choosing the data there while there is an abundance of news websites that continually generate content, hate-crime news stories are generally rare relative to all of the news that gets generated everyday. As a result we had to turn to more specific outlets that had refined streams of news specifically curated to be structured around discrimination and hate-crimes.

Below we present a more detailed discussion of the specific sub-tasks that had to performed with the data, and a detailed outline of the steps taken to ultimately construct a date-base with over 20,000 hate-crime news instances.

3.1 Collection

Collecting the data was initially very challenging. Just scraping news websites was an option, but as previously stated, few actually have a detecated source of stories relating to hate-crime, and filtering the remaining news stories proved to be too time consuming as the examples available in the data were very sparse. Furthermore former open news aggregate APIs, like Google News, no longer exist and very few news websites provide robust APIs capable of narrowing articles down to subjects like "hate-crime".

The New York Times (NYT) was a crude exception to the rule, in that they had a dedicated web-page which contained all of their news stories written on hate-crime and hate-crime related events (e.g. police activity, and court proceedings). However, the national and even global reach of the news site made it predisposed to writing only about high profile cases. Also using only the NYT limited us in the number of authors and headlines from which we could detect events. This suggested that while reaching globally, NYT may miss out on more localized instance of hate-crime events and be biased in the reporting of certain events.

To combat and diversify our sources of information, we turned to Event Registry, and NewsAPI. Both APIs contain a pay-free limited stream of news article data from various sources for a query. Both only provide a certain prefixed time frame of data, either temporally in or in volume per request. While being limited they did provide a number of specific smaller news sources that were usually smaller and more local to a specific region. Furthermore both data source would make geographical data accessible when it was available, this opened the possibility to using co-location methods to better predict the similarity between 2 news stories.

Further still, ProPublica is a non-profit newsroom organization, which explicitly manages a group dedicated to the collection and aggregation of new stories on the subject of hate-crime. They do not have an API which is directly accessible but do provide access to a weekly updated CSV file which stores the headline, location, date of publication, and original story URL. This data was very high quality as it is largely human curated, rather than the results of a query, and was available for news stories since the beginning of the project in February of 2017 (i.e. mora than a year of data)

Finally, there was a belief that web searches might be indicative of critical events occurring in the news. So Google Trends (GT) data was desired to track the interest in a subject over time. At the time, no API is available and while a number of complicated solutions occur, GT provides a functionality for downloading CSV files which contain data pertaining to Google search queries made over a time range. Due to the length of time and specific queries we were looking for, the data was available at per-week resolution.

3.2 Cleaning

The first step in cleaning the data was normalizing the fields which we tracked. In particular the fields associated with each news story were:

- Date of Publication - This was a temporal marker for when a particular news story occurred, and thus when it first appeared on the news source
- Story Headline - The headline of the news story
- Location - If available the city, state, and country of each news story was recorded
- URL - a url which pointed to the website where the news story was published

- Source/id - the API or website from which the data was made available plus an numeric indexing which helped identify the uniqueness of a particular story from a source

Using the KARMA interface, we created a context that leveraged the `schema.org` OWL ontology, and applied it the various collections of data that we had accrued. With the data fields all normalized, we could begin using Natural Language Processing (NLP) technologies to analyze and resolve entities.

4 Entity Resolution

When resolving entities, we were posed with a number of hierarchical resolution problems:

- Named Entities - in particular were there widely-identifiable entities such as People, Geo-Politica Entities, or Locations that could help identify whether 2 news stories were related to each other
- Agents and Actions - identifying agents and actions in the new story title would be indicative of whether or not 2 news stories, without explicitly named entities could be the same event
- Unique Events - now with the agents, entities, and actions identified, we can use a scoring function to find data in our records which is related to uniqueness of an event, and can aggregate a number of stories into a single event in time, rather than having large amounts of redundant information.

Below we outline and discuss the methodologies and practices that enabled us to accomplish these tasks.

4.1 Named Entities

Detecting Named Entities has a long well studied history in NLP and a number of endeavors to create open and publicly available Named Entity Recognition (NER) software have been made. Of particular interest in the NLP library for Python SpaCy which integrates together a number of well known algorithms and solved many different established problems in the domain. SpaCy is robust and fast and lends itself nicely, as it has built-in to its parser a NER system.

Each headline was processed and parsed using the SpaCy pipeline, and named entities were extracted. Afterwards, a list of these entities was loaded into the graphical data processing tool OpenRefine, which was able to effectively cluster similar named entities (e.g. Donald Trump, and Trump usually refer to the same person) and reduce then reduced the number to the most frequent named entities available. Following the refinement and processing of named entities, a new field was added to every data point for news stories that stored the named entities associated with that particular new story.

4.2 Predicates

Similarly to the methods described above, we turned to the Semantic Role Labeling (SRL) as a method to extract agents or participants and actions from sentences. This was harder to accomplish as SpaCy does not provide a SRL tool out-of-the-box, and many of the available tools require a tremendous amount of hand-labeled data to be used freely. Fortunately, University of Washington's Know it All research group, provides a SRL tool called ReVerb which works using Java.

The Java library is made available as well as pre-built JAR file which can take plain-text files (or an IO pipe) on input and outputs the semantic labels (noun-phrases, verb-phrase dependency linkages) it identifies in the sentences it finds in the file.

Initially the approach was to run the `java` command line from python on a temporary file which contained a single sentence with the headline. However the amount of time needed to load the binary per processing cycle made the task excruciatingly long. It would take about 3 - 5 seconds a headline across a total of 22,000+ headlines so a different method was needed. So instead a single file with all headlines was created and processed in one single batch, then each headline was tested for each sentence extracted and each SRL triplet was assigned a list of the IDs for the news stories that corresponded with them. As ReVerb only needed to load its NLP pipeline once and perform a batch processing, it was able to process all 22k+ headlines in a matter of seconds.

Once the SRL was complete the triplets were loaded into OpenRefine. In OpenRefine the set was reduced using clustering and filtering. Many of the noun phrases and verb phrases contained adjectives and adverbs which or were in deferent tenses, which were then normalized. This reduced set made it simpler to identify what was happening based on the subjects and actions in the headline.

4.3 Events

An event is defined as a real-world happening and can be represented in several variations across news sources. To find these different representations related to the same event, we define a score function. According to the real world attributes of event entities, the score function uses time, location, headline word similarity, and named entities in identifying overlapping news stories.

To this end, we compare each news article against previously detected events. The news is considered to be associated with an event if:

- They are co-located: if the news article has happened in the same state and city as the event.
- They are simultaneous: whether there is at most a two-day delay between the news article and the event.
- They are similarly represented: if the news article has a headline which is textually similar to the headlines of other news articles associated with the event. To this end we calculate the ratio of the intersection to the union of the words in two headlines.
- They include similar named entities: if there is a significant intersection between the sets of named entities in the news article and the event. Here again, the ration of the intersection to the union of named entities is calculated.

```
similarity in headline
Man shouts 'go back to Lebanon' to Sikh-American girl on NY subway
White man shouts 'go back to Lebanon' to Sikh-American girl

similarity in named entities
Dip in Admissions of US Universities { Hate Crime Effect?
US Universities register decline in Indian Applicants due to rising Hate Crimes,
   concerns over changes in visa policies by Trump Administration

similarity in named entities
Bible, rocks thrown through doors of Colorado mosque
Colorado Mosque Attack Suspect Arrested,
    Accused of Throwing Bible and Rocks in Hate Crime

similarity in named entities
Hate Crimes In LA Hit Highest Mark Since 2008, With Marked Increase Against LGBT Community
Hate crimes in LA up 15 percent in 2016 compared to 2015, study finds
2016 Saw Rise in Hate Crimes in LA, Many Against LGBTQ People: Study
Hate crimes rise 15 percent in LA with uptick in LGBT victims
```

Figure 1: Examples of News Stories Linked into single Events

As a specific example from Figure 1, the two following sentences the ratio of intersection to union is > 0.5 and the two headlines are considered similar:

- Man shouts "go back to Lebanon" to Sikh-American girl on NY subway.
- White man shouts "go back to Lebanon" to Sikh-American girl.

Similarly in the following two sentences the ration is > 0.5 and the two headlines are considered similar:

- Bible, rocks thrown through doors of Colorado mosque.
- Colorado Mosque Attack Suspect Arrested, Accused of Throwing Bible and Rocks in Hate Crime

Once the association between a news article and an event is recognized, the relation between the two is stored in the event entity and the news article enters the set of articles that are the representations of the particular event. If the news article is associated with none of the existing events, we consider it as a new event and add it to the events set.

Finally we store the data into a JSON file which aggregates the news stories into a headline file with an array of event objects containing information about all of the relevant new stories. We also compute and store an "exposure score" which is the ratio of unique URLs reporting on an event to the average number of URLs per event detected.

5 Data Exploration and Visualization

With the data assembled in to a number of time series streams, we would like to now be able to process, monitor, and visualize insights about the data. In order to do that. To accomplish these goals, we turn to the Elastic Stack a set of programs that allow for robust data indexing and search.

ElasticSearch is the backbone of the system, built on top of Apache Lucene, it is a framework for creating searchable content by indexing JSON documents. Each of the news stories, events, and GT data points a JSON document is created. For the news stories each news source (API or website) was assigned a unique index, as well as, separate indices for events and GT data. This made it easier to explore smaller sets of data, or to isolate data sources to understand how the data varied. As an example only after loading the data into its own indices and looking at it from there did I realize that despite querying for a years worth of data, Event Registry only returned data since the beginning of this particular year.

Having loaded the JSON documents and indexing them, I was interested to see how the news data evolved over time and which locations (if any) were particularly hot spots for hate-crime news. To visualize we used another Elastic Stack service, Kibana, which provides a graphical dashboard for visualization and querying your ElasiticSearch indices.

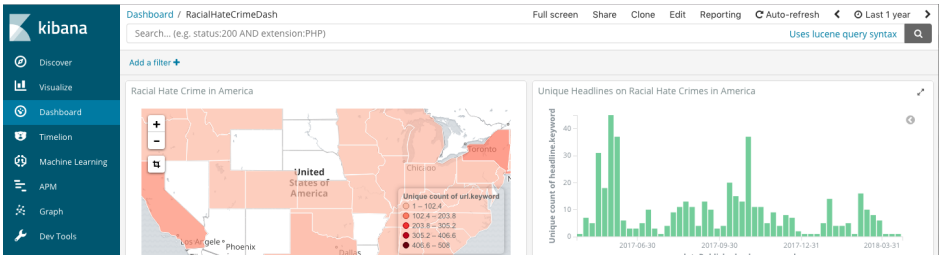

Figure 2: Visualization Dashboard using Kibana to understand "Hate-Crime" News in around the US over the last year

Figure 2 demonstrates how this data can be explored using Kibana and the Elastic Stack. Allowing for quick and easy analysis of the events and news stories happening. The example demonstrates quickly the under reporting of hate-crime in specific location, and also shows temporally how those stories have been occurring (or reported) most recently.

Elastic Search and Kibana provide researchers and data scientist a strong ability to effectively analyze and quickly gain insight into the data available, as well as, interface with it with a wide variety of tools and libraries.

6 Conclusion

We presented a methodology and process for collecting, cleaning, indexing, and exploring data on hate crime. The described work outlines the difficulties, and challenges of this type of work. However, in spite of the problems we have demonstrated that with the correct tools and frameworks, the data lends itself nicely to analysis and interpretation.

Demonstrating the power of Semantic Web Technologies, Entity Resolution Strategies, and Natural Language Processing, we have contributed an important and meaningful first step towards better understanding hate-crime and its underpinning nature.

7 Future Work

In this research, we mainly investigate the representation of hate crime events in the most popular news sources in the United States. This data is very coarse and misses smaller or under-values minor (less covered) instances of hate-crimes. To improve this imbalance, we suggest that social media data can be considered as a source for exploring and detecting hate crime events. By integrating news articles with social media data we would have a more reliable tool for detecting real-world events.

Moreover, by tapping into and collecting from generally smaller news sources like local news stations in American counties and townships, geo-social analysis can be conducted to better understand how hate-crime representations vary by different social foundations. When compounded with more localized social media data and text, we can enrich our understanding of how cultural backgrounds can change the interpretation and motivations of hate crimes.