Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes

Aida Mostafazadeh Davani, Leigh Yeh, Mohammad Atari, Brendan Kennedy
Gwenyth Portillo-Wightman, Elaine Gonzalez, Natalie Delong, Rhea Bhatia,
Arineh Mirinjian, Xiang Ren, Morteza Dehghani

University of Southern California

Abstract

Official reports of hate crimes in the US are under-reported relative to the actual number of such incidents. Further, despite statistical approximations, there are no official reports from a large number of US cities regarding incidents of hate. Here, we first demonstrate that event extraction and multi-instance learning, applied to a corpus of local news articles, can be used to predict instances of hate crime. We then use the trained model to detect incidents of hate in cities for which the FBI lacks statistics. Lastly, we train models on predicting homicide and kidnapping, compare the predictions to FBI reports, and establish that incidents of hate are indeed under-reported, compared to other types of crimes, in local press.

1 Introduction

Hate crimes are defined as crimes of violence either against a person or their property that display evidence of prejudice based on the victims' race, gender or gender identity, religion, disability, sexual orientation, or ethnicity (Jacobs et al., 1998). According to the results of a new Department of Justice hate crime report released in 2017 (Masucci and Langton, 2017), approximately 54% of hate crime victimizations were not reported to police during 2011-2015. Despite the recent efforts of advocacy groups, policy makers, and researchers to create reliable, national data to understand the extent and severity of hate crime victimization, the existing estimates continue to fall short (Pezzella et al., 2019).

It stands to reason that hate crimes induce local disturbance, and as a result, might be likely to get local coverage. Therefore, local news agencies can be considered a unique source of information for detecting these incidents. Here, we use a corpus of local news articles, collected from the Patch[1] website. The Patch data contain independent, hyper-local news articles compiled from local news sites.

We apply event extraction methods to identify incidents of hate crime reported in the Patch corpus for cities with no representation in FBI reports, and analyze the frequency of the extracted events compared to the number of incidents reported by the FBI. The task of labeling each article as a hate crime or not is defined as a Multi-Instance Learning (MIL) problem since each article is modeled as a sequence of sentences. Instead of predicting a label for each sentence, we use the information embedded in all the sentences of an article to determine whether the article is reporting a hate crime.

After testing the model on a set of annotated articles, we apply the trained model to cities for which the FBI does not have any reports, and we provide a lower-bound estimate on the occurrence frequency of hate crimes in those cities. Lastly, we compare the coverage of incidents of hate as reported in local news sources with coverages of two non-hate crimes, namely homicides and kidnappings, and contrast the overlap of the extracted incidents with those reports by the FBI.

Our results show that applying MIL for event extraction can help approximate the missing reports, especially in cases in which publishing the comprehensive event set faces challenges and is influenced by subjective bias.

2 MIL for Event Extraction

In this paper, we perform event detection and extraction on news articles based on taxonomies of acts of crime. We adapt the MIL approach for event detection developed by Wang et al. (2016), which identifies key sentences for a given article. We then use these key sentences to perform event

[1] https://www.patch.com

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5753–5757,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

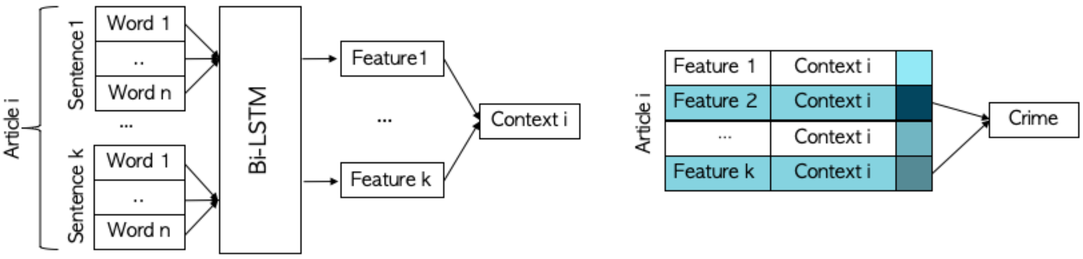

Figure 1: The event detection model using a MIL network. Local representation of each sentence are combined with context representation of its related article.

extraction, predicting the target and type of action for a given incident.

Event Detection

The MIL approach for document classification is illustrated in Figure 1. The two basic components are the creation of local features (representations of sentences) and the aggregation of these features into a document representation. Whereas Wang et al. (2016) applies Convolutional Neural Networks (CNNs) for the creation of local features, we use a bidirectional Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) network for representing each sentence of an article. Bidirectional networks (Graves and Schmidhuber, 2005) have been shown to provide a good semantic representation of textual data (Huang et al., 2015).

Local representations are then aggregated to form a "contextual" representation of the document, using a CNN layer. This context vector, which is the same for all sentences in the document, is then concatenated with *each* sentence's local representation.

Given the feature representation of the sentences in an article, the probabilistic score of each sentence in an article is calculated using a fully connected layer with sigmoid activation. This probabilistic score shows the extent to which the sentence contributes to predicting the crime label of the article. The label for a bag of sentences is calculated by averaging the *k* highest probabilistic scores. We checked the results with *k* being set to 2 or 3, since a few number of sentences in each article can determine the label.

Another prominent method which we compared the MIL results with is Hierarchical Attention Networks (HAN; Yang et al., 2016). HANs apply attention first at the level of words, then at the level of sentences, to produce representations of documents subject to local variations in textual importance. We also compare the results of neural network models with TF-IDF as a text classification baseline.

Event Extraction

The most challenging aspect of extracting events from a sentence is that the context of a document should be considered in order to interpret an entity and the type of triggered event (Chen et al., 2015). Approaches that exclusively use word features for the task usually lack comprehensiveness.

The event detection model in the previous section produces, for each positive prediction, a small set of sentences likely to influence the document's label. In the event extraction step, we use a bidirectional LSTM text classifier to predict the attributes of a crime event.

The attributes of a crime event are determined by the taxonomy proposed by Kennedy et al. (2018) for annotating hate rhetoric. In our case (see Section 3), we are predicting two attributes: the target of a crime event, and the type of crime.

Formulated as a multi-class, multi-task prediction, we train a biLSTM to produce a representation of the concatenation of the top two sentences and feed this to two separate feed-forward networks, one predicting the target categorization and one the crime type.

3 Data

The Patch website includes hyper-local news articles from 1217 cities based in the US. For this project, we scraped the articles in the "Fire and Crime" category of Patch, resulting in a corpus containing $\sim 370k$ unlabeled local news articles. For our experiments, we manually annotate subsets of the main dataset for training event detection models.

Our annotations consisted of a binary label

— whether the article represents a specific hate crime — as well as labeling the attributes of hate crime articles, which consist of the target of the action (whether the crime was based on the race, nationality, gender, religion, sexual orientation, ideology, political identification or mental/physical health of the target) and the type of action (whether the crime was an assault, arson, vandalism or hate demonstration).

For gathering a subset of articles for annotation, we filtered the news articles based on a set of 8 keywords (*swastika*, *hate*, *racial*, *religion*, *religious*, *gay*, *transgender*, *transsexual*) related to hate crimes, resulting in $\sim 3k$ patch articles, which were then combined with 500 randomly sampled articles to account for the high frequency of the hate crimes in the selected dataset. Each article was annotated for the presence and the attributes of the hate crime reports by one annotator. Annotators achieved 0.73 inter-coder agreement on a subset of 500 posts based on Cohen's Kappa (Cohen, 1968).

For hate crime articles that are not associated with the keywords, we expected the model's predictions to be sparse. To deal with this problem we applied an active learning approach introduced by Lewis and Gale (1994). In this approach, after training the model, we predicted the hate crime label for all the articles in the dataset and gathered their associated probabilities. We then selected $\sim 1k$ articles based on their probability score, using a normal distribution with a mean of 0.5 and standard deviation of 0.1. This set of articles, for which the model was uncertain about their labels, was then annotated by the same annotators and added to the training set.

We performed a similar procedure, without entity labeling and active learning, for homicide (keywords: *homicide*, *manslaughter*, *murder*, and *kill*) and kidnapping (keywords: *kidnapping*, *abduct*, *hostage*, *abduct*, and *shanghai*) events. The frequency statistics for these annotations are represented in Table 1.

Event Type	Positive	Negative
Hate Crime	1979	3192
Homicide	1664	1327
Kidnapping	1864	1104

Table 1: Frequency of events from Patch annotations.

Type and target of the hate crime was also annotated for each article. Crime type labels are distributed across assaults (900), arson (76), vandalism (450), and hate demonstrations (543). The most frequent target types were race (1029), religion (376), and sexual orientation (265).

	MIL	HAN	TF-IDF
Hate crime	**82.9**	82.6	81.6
Homicide	**81.3**	79.7	77.4
Kidnapping	**78.7**	75.6	73.9

Table 2: Event detection F1 scores for the test set

4 Experiment

All models were implemented with Tensorflow (Abadi et al., 2016). Hidden size of the LSTM cells was set to 50, filter sizes of the CNN were set to 2, 3 and 4, and a dropout layer was placed on top of the LSTM cell to set 25% of the values to zero. Each batch included 5 articles converted to their latent representation using 300-dimensional GloVe word embeddings (Pennington et al., 2014). Parameter tuning was performed with 70% of the dataset as the train set and 10% as development set and the learning rate was set to 0.00008.

All three models for predicting hate crime, kidnapping and homicide were trained for 50 epochs.

5 Results

The resulting F1-scores are calculated for the test set and represented in Table 2.

We apply the learned models to make predictions about the rate of hate crimes in cities for which the FBI lacks data. We also compare the relative rate of news coverage of hate crime with those of homicides and kidnappings.

Predicting Hate Crime

First, we compare the positive hate crime labels predicted for Patch with the FBI's city-level hate crime reports. After applying the trained model to the Patch dataset, we captured 3152 articles that report hate crime incidents. These articles include 678 reports from 286 cities that have no representation in the FBI reports. This suggests that the MIL model applied to the local news dataset can approximate missing statistics on hate crime in those cities. However, presuming a one-to-one relation between the news articles and hate crime incidents is not accurate, since there can be false positive results and duplicated articles about an incident. To provide an accurate set of unreported hate crime incidents we removed duplicated and misclassified articles from the set of 678 unrepresented hate crime incidents.

In order to account for the possible duplications, we utilize the event extraction model to capture the event entities, namely target and action type. Running the extraction model with the same hyperparameters yields the results presented in Table 3. We use the entities together with the time (mentioned in the dataset) and location (extracted with named entity recognizer of CoreNLP (Manning et al., 2014)) of the articles to detect duplicated events.

Label	Precision	Recall	F1
Target	63.9	65.3	63.9
Action	67.7	68.0	67.4

Table 3: Event extraction scores of MIL

After checking for pairs of articles from the same state and city, with the same reported target victim and crime action, reported at most one day apart from each other, we found 20 pairs of duplicated articles, indicating 658 unique incidents of hate in the cities with no representation in the FBI dataset.

Next, we manually checked these articles and found 315 articles that were correctly labeled as hate crime. Table 4 represents a few instances of false positives. Exploring the false positive results indicates that non-hate crime articles that mention minority social groups are often incorrectly labeled as hate crime. This issue can be explored further in future works to improve the accuracy of the predictions.

Comparisons to Other Crime

In order to compare the coverage of incidents of hate with coverage of homicides and kidnappings, we contrast the overlap of the extracted incidents with those reported by the FBI. Specifically, for 159 cities that have representation of the three crimes both in Patch and FBI crime reports, we calculate the ratio of Patch-based predictions to FBI reports for each crime.

To investigate the differences between the distributions of these ratios, we ran a Welch-type one-way ANOVA, which is robust to non-normal distributions, allowing for heteroscedasticity and extreme non-normality of ratios in our data (Field and Wilcox, 2017). The results indicates that the three crimes' distributions have significantly different medians ($F[2,214.28] = 102.03, p < 0.001$). Post-hoc tests suggested that Patch-based estimates of hate are significantly lower than homicide and kidnapping (both p's < 0.001).

	False Positive Examples
1	A former Ku Klux Klan leader in Ozark was sentenced Thursday to a decade in prison for sexually abusing a woman in southern Alabama.
2	The FBI is part of an investigation into a suspicious substance delivered to a Council on American-Islamic Relations office in Santa Clara on Thursday.
3	The hate from the violent white nationalist gathering that resulted in the death of an anti-racism protester in Charlottesville, can be found anywhere.

Table 4: First sentences of sample articles recognized as false positive results by our annotators.

6 Discussion

Hate crimes in the US remain vastly underreported (Masucci and Langton, 2017). For instance, only 12.6% of the agencies in the FBI report indicated that hate crimes had occurred in their jurisdictions in 2017, and agencies as large as the Miami Police Department reported zero incidents of hate (FBI, 2018), which seem unrealistic. The contributions of this paper are two-fold: First, we have shown that event detection can be applied to the study of hate crimes. Specifically, we demonstrated that using MIL for event detection, in conjunction with local news articles, can provide conservative estimates of the occurrence of hate crimes in cities with no official representation. A possible application of this method is creating a real-time hate crime detector based on online local news agencies, providing researchers and community workers a lower-bound on the number of hate crimes in such locations. Second, the statistical analyses suggested that the scope of hate crime coverage in local news is lower than that of violent, but non-hate crimes. This suggests that although local news sources can be used as an additional source for gathering better statistics about hate crimes, the predictions of our models are simply lower bound estimates.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 167–176.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

FBI. 2018. Hate crime statistics, 2017. https://ucr.fbi.gov/hate-crime/2017. Accessed: 03-04-2019.

Andy P Field and Rand R Wilcox. 2017. Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy*, 98:19–38.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

James B Jacobs, Kimberly Potter, et al. 1998. *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand.

Brendan Kennedy, Drew Kogon, Kris Coombs, Joseph Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh Davani, Mohammad Atari, and Morteza Dehghani. 2018. A typology and coding manual for the study of hate-based rhetoric. *PsyArXiv*.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR94*, pages 3–12. Springer.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Madeline Masucci and Lynn Langton. 2017. Hate crime victimization, 2004-2015. *Washington, DC, US Department of Justice Office of Justice Programs Bureau of Justice Statistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Frank S Pezzella, Matthew D Fetzer, and Tyler Keller. 2019. The dark figure of hate crime underreporting. *American Behavioral Scientist*, page 0002764218823844.

Wei Wang, Yue Ning, Huzefa Rangwala, and Naren Ramakrishnan. 2016. A multiple instance learning framework for identifying key sentences and detecting events. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 509–518. ACM.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.