Antisemitic Tracing on Reddit: An Intersection Approach Toward Dissecting Ideologies

Jason Michálek
Indiana Univeristy, Bloomington
jasomich@iu.edu

Chang-Han Chang
Indiana Univeristy, Bloomington
cc93@iu.edu

Abstract

The domain-specific communities that form in subreddits may seem at first glance to cohere around the topical label with which each subgroup is named. However, such a surface-level orientation may not be as transparent in framing what is discussed and how. Particularly in detecting hate speech and other potentially harmful discussions, such analysis requires more robust methods to detect and understand the emergence of ideologies. In this study, we present an intersection method of determining likelihood of harmful discussions. Our methodology replicates successful detection methods from previous studies and suggests opportunity for more discrete analysis.

1 Introduction

In this study, we employ the intersection method of comparing antisemitic terms identified as recently as 2019 by an organization devoted to stopping antisemitism on the internet, Global Jewish Advocacy (GJA) (2019). Our research draws from a variety of previous studies that offer approaches to analyzing and understanding instances of "hate speech," broadly construed (Pontiki et al. , 2020) (Hasanuzzaman, Dias, and Way, 2017). In our particular study, we offer a representative sampling of subreddit comments in order to demonstrate the statistical copresence of terms that could be denotated as "antisemitic" within a given comment. Though our results demonstrate the complexity of instances in which such terms occur, our overall analysis demonstrates that the list of terms we selected is statistically significant in determining the presence of hateful ideologies, despite authorial intent.

Using our list of terms and the intersection method of identifying copresence of multiple terms, we compiled all the instances in which a minimal amount of terminal intersections to generate a discursive exchange that likely included antisemitic sentiments. We then used these to generate a statistical representation of posts that fit our functional definition of antisemitic speech. We compiled charts of robustly populated subreddits with dual-segment bars to demonstrate both the total occurrence of any term as well as the unique occurrence of each term. To offer a perspective of how our method could be used, we show that the occurrence in our dataset varies over time in idiosyncratic modes and we also identified top subreddits as ranked by unique occurrence to evidence potential corpora for further application. We then manually reviewed a representative sample of posts to determine a qualitative variance of perceived tone and message.

In this study, our analysis successfully identified the subreddits that offer statistically significant intersection of terms. Our preliminary analysis demonstrates that in each instance, the intersection of at least four of our identified potential words suggests a re-presentation of hate. However, while we initial intended to divide these instances further into a context of "repeating or reporting" hateful ideologies, we determined that further parsing would be necessary to understand the nuanced sentiments that were present in each instantiation. Whether the intent is to reproduce hate or to reduce and critique such ideologies, our aim was to first document presence rather than intention. Therefore, we conclude with justifying this method as a first step to addressing harmful ideologies since the presence of hateful ideologies can inflict trauma regardless of why they were (re)produced. We thus present our findings as a promising foundational method and offer potentials for using this process to develop a more robust methodology that can be employed to curtail the harmful effects of antisemitism.

2 Related Work

In our paper, we employ methods of data mining and textual analysis in order to determine if antisemitic sentiments result from long-standing user sentiments or event-specific contexts. Building upon previous studies that seek to collect and analyze forms of hate speech our paper borrows the methods and rationale from several studies that focus on analyzing Twitter content for racist and sexist language.

In analysis of anti-Greek xenophobic sentiments on Twitter, Pontiki et al. (2020) applied techniques of data mining to parse textual data in order to determine the sentiments it presented. The authors provide a rationale for how the methodology of identifying verbal aggression (VA) can inform detection and analysis of xenophobia as modeled by a previous study of the Greek financial crisis from 2013-2016. They conclude that the xenophobic sentiments they identified in 2019 reflected "the existence of dominant perceptions that are deeply rooted in the Greek society and keep being reproduced" rather than demonstrating a direct effect of the financial crisis that was the focus of the VA analysis in the model they replicated.

Similarly, Hasanuzzaman, Dias, and Way (2017) provide a data mining model that incorporates demographic contextualization of context-embedded sentiments. The authors acknowledge that personal user experiences, as well as broader social norms, work together to contextualize the content that emerges on social media. In contrast to hyperattention on text or utterances merely contextualized by a particular domain or culture, they focus on the demographic context of users themselves in order to construct and model a process of identifying word embeddings that may represent racist or hateful sentiments. Furthermore, as their model was derived from and tested on a large corpus, the variety of content and contextual cues provides further robustness for the model they constructed. Because their model accounted for demographics rather than discrete embeddings that created such user-related contexts, the resulting methods of sentiment analysis provide an instrumental way to collect and parse user content for demographically specific sentiments of hate speech. However, they call for further implementation of such methods in coordination with discrete methods in order to develop more reliable methods of detection.

Both the above studies draw methods from a study outlining techniques for detecting features of hate speech on Twitter, a further complication is presented by Wassem and Hovy. (2016) The authors of this study used extra-linguistic features and character *n-grams* to form a dictionary from the words in their data set that documented racist and sexist features of language. While these methods of data collection could be helpful to our study in providing collection and documentation, we identify features of antisemitic language drawn from a specific domain and take into account the user demographics that Hasanuzzaman, Dias, and Way (2017) contributed to Wasseem and Hovy's (2016) methodology.

In an analysis of web-based profanity detection, Sood, Autin, and Churchill (2012) provide a critique of list-based methods for detecting profanity. The authors acknowledge that methods at their time of writing did not account for adaptations in profane language use, nor do they account for specific contexts such as domain, communities of practice, of social settings. They focus their analysis on the user comments on the news site Yahoo! Buzz, employing the Amazon Mechanical Turk (MTurk) Since the user-generated comments from the social news site are labeled by Amazon Mechanical Turk workers, they were able to analyze salient features to code for profanity. They confirm their inclinations that such methods fail to adapt to evolving language features such as misspellings, disguised terms, or various forms of authorial censorship.

In "Detecting Hate Speech on the World Wide Web," Warner and Hirschberg (2012) seek to identify hate speech defined as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation" (20). Using user-flagged content from Yahoo! News and URLs collected by the American Jewish Congress with content that advertisers found "unsuitable," the authors collected content Through and then generated feature sets of unigrams of potentially hateful terms, coded as positive or negative by the measure of a "stereotype sense," determining whether they supported a problematic sentiment or not. Using their control variable of volunteer annotators, the authors conclude that hate speech was too ambiguous for the classification system to deeply parse problematic senti-

ments, and they suggest that parsing was also significantly degraded by bigram and trigram templates.

3 Data Sets

This study offers a coordinated analysis of antisemitic speech in domain-specific corpora. To isolate a representative sample of potential antisemitic speech, we utilized "lzma" and "zstandard" packages available in Python and analyzed Reddit data from January 2018 to June 2019 (793,326,999 posts, 204.4 GB in compressed format "xz" and "zst"). This data is stored in the shared directory on GH server in the Luddy School at Indiana University (gh.luddy.indiana.edu), extracting 5,824 subreddits that presented at least one of the words the GJA defined as having historical connotations of antisemitism. We then parsed this selection further to isolate only those subreddits that offered the intersection of four or more terms in a single post.

We collected these subreddits as sites of antisemitic ideologies, regardless of authorial intent. By approaching the texts to first determine the presence of such speech before attempting to determine the cause or intent of such copresence, this study offers a way to accurately identify the occurrence of ideologies—which is a necessary first step toward applying what is suggested in the title of the GJA report: "stopping antisemitism starts with first understanding it."

As a contrast to the study by Warner and Hirschberg (2012) who constructed a list of thematic stereotypes in a localized data sample, our methodological approach was meant to respond to global trends as a way of deductively identifying co-presence of prescribed antisemitism. While this is not as localized as a localized template-based strategy, it could provide sentiment connections between domains—such as the shared use of "Globalist" as an antisemitic derogation. Also, though this phase of our research was a use case for the intersection method, such a method could be complemented with the previous study's notion of "stereotype sense."

The above bar charts demonstrate our findings. In each chart, the darkened bars display the quantity of unique occurrence of each of our key terms whereas the lighter, extended bars represent the total number of occurrences for all terms. Figure 1.1 portrays the occurrence of antisemitic words over

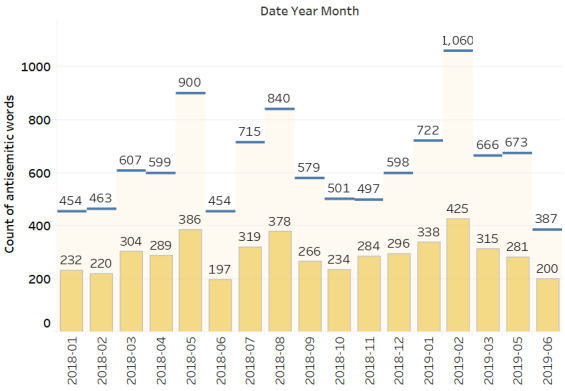

Figure 1.1

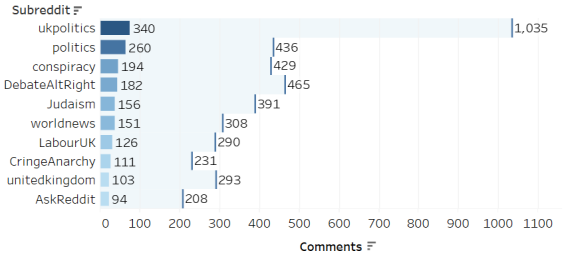

Figure 1.2

time whereas Figure 1.2 isolates the top 10 subreddits in terms of significant occurrence of antisemitic speech.

4 Methods

To administer the intersection method in our study, we identified posts that contained four or more of the words that GJA deemed to be antisemitic. We used the intersection method to identify the presence of terms T identified by the GJA report as they occurred in the subreddit comments C. We tokenized each unique comment C_i from a subreddit S to separate it and see how many comments intersect with semantic terms (occurance O). For each unique occurrence, the string would receive an additional point to its score P_s. If it reached four points, we collected it and added it to our library of potentially antisemitic strings.

$$O = \bigcap_{i=1}^{n} C_i T$$

$$P_S = length(O)$$

In order to assess the reliability of our selection decision regarding promising instances for analysis, we wanted to identify a minimum number of

terms that would give us a good sense of potential for antisemitism. To determine a minimum operating benchmark, we sampled strings at several increments of occurrences. We manually looked at the threads that we compiled using two and three intersecting terms, but we determined that a minimum of four terms occurring in a string gave the most promise for the occurrence of antisemitic ideology. We then manually analyzed a representative sample of some of those threads and found that it didn't always convey the intent of being anti-Semitic, but the ideology was almost always there. Therefore, we agreed on the heuristic for the threads we identified were either repeating or reporting antisemitism.

In the analysis of our method, we were aware of the limitations posed by context ambivalent forms of collecting content, such as *ngrams* used by Wasseem and Hovy (2016) and a template-based strategy of classification mentioned by Hasanuzzaman et al. (2017). Therefore, our approach sought to provide a non-ambivalent context from our selected list of detection terms as well as forming domain-specific context. Furthermore, our rationale for using an intersection methodology seemed justified to negate ambivalence in collection and classification due to the semantic analysis of terms by an international organization actively devoted to monitoring the nuances of antisemitism.

5 Evaluation

Because hate speech is often nuanced and ever-changing, our goal was not to determine the nature of the conversation but rather potential for offense and thus offer rich instances to implement further methods of analysis. In the following posts, we evidence how each seems to be in relative opposition to antisemitism, but the subtext could actually reinforce ill intentions. For instance, the distinction between being Jewish by blood and Jewish by practice is a common topic to our colleague who works in a Jewish outreach program, but the term "Judite" was unfamiliar in the following:

> *"I think something similar to this is going on, as far as a globalist conspiracy is concerned, but I don't think it's Jewish. I think Jews are a group used in addition to other groups. And I also just think Jews are biologically pretty smart which is why we see such a high representation of them in finance and business.*
>
> *Also we have to define what we mean by Jew. You have Jews and then you have Judites. You have Jews by blood and then Jews by religious ideology. It's a big topic." – posted 2018-01-17 00:15:44 in CapitalismVSocialism*

Additionally, negative attitudes against the political ideologies of Globalism are collapsed into nuanced antisemitic epithets—as identified by the GJA and explained in more detail elsewhere[1]. However, seeking to untangle such collapses could still be read as offensive in positioning explanations as mere exceptions to stereotypical attitudes, or as retaliatory hate speech against those espousing antisemitic sentiments. Furthermore, some of the posts we identified that were defending against antisemitism deployed nuanced hate speech such as "goyim"—a term denoting non-Jewish peoples as oppositional others. This can result in retaliatory hate speech such as in the following:

> *"i'm a 100% jewish supremacist. i support the great jew weinstein in his defense against the impure goyim women who are simply jealous of the jew success.i officially proclame that anyone that goes after my boy weinstein is a Nazi,antisemite and holocaust denier who should be punished for hate speech.Beware goyim or you shall feel the wrath of our tribe" – posted 2018-01-10 15:06:51 in movies*

Because further development of our methodology will require sentiment analysis that is domain-specific and context-embedded, we deferred such development to the next phase of our research.

6 Discussion and Conclusion

In this study, we proposed to test how the intersections of antisemitic terms might be used as a foundational step in identifying hateful speech of Reddit. We demonstrated the statistical significance of applying an approach to collect user posts for

[1]For more descriptive explanation, see: Zimmer, B. (2018, March 14, 2018). The origins of the "globalist slur," The Atlantic. Retrieved from https://www.theatlantic.com/politics/archive/2018/03/the-origins-of-the-globalist-slur/555479/

further analysis. However, our manual review of posts revealed that understanding the nuanced intentions in a given string is more difficult than isolating a minimum amount of predetermined antisemitic terms. In order to advance the aims of this method, future applications must also (1) incorporate additional context (such as thread subject and other user posts) to distinguish interrelated semantics; (2) determine multiple connotations and denotations of terms selected for intersection capture; and (3) actively revise terms to reflect the coded language of the timeframe in which content is analyzed.

Based on our research and the demand for approaching antisemitic tracing on an international scale, we suggest that the intersection method could prove to be exponentially useful when supplemented with active methods of accumulating and updating corpora of potentially hateful terms. One such method known as CLARIN supports international efforts in the humanities and social sciences, and thus an intersection methodology would be appropriate on multiple levels (Eskevich et al. , 2020).

Acknowledgments

The authors wish to thank Dr. Allen Riddell for his instructions. This work is part of the Social Media Mining (ILS-Z639) course for Fall 2020 at Indiana University, Bloomington

References

Global Jewish Advocacy (2019, November.) *Translate Hate: Stopping Antisemiticm Starts with Understanding It.* https://global.ajc.org/files/ajc/upload/AJC_Glossary.pdf

Zimmer, B. (2018, March) The origins of the "globalist slur" https://www.theatlantic.com/politics/archive/2018/03/the-origins-of-the-globalist-slur/555479/

Mohammed Hasanuzzaman, Gael Dias, and Andy Way (2017, November.) Demographic Word Embeddings for Racism Detection on Twitter. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing*, (Volume 1: Long Papers) (pp. 926-936).

Maria Pontiki, Maria Gavriilidou, Dimitris Gkoumas, and Stelios Piperidis (2020, May) Verbal Aggression as an Indicator of Xenophobic Attitudes in Greek Twitter during and after the Financial Crisis. *In Proceedings of the Workshop about Language Resources for the SSH Cloud*. (pp. 19-26).

Zeerak Waseem and Dirk Hovy (2016, June) Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *In Proceedings of the NAACL Student Research Workshop*, (pp. 88-93).

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012 Profanity use in online communities. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp./ 1481–1490). ACM.

William Warner and Julia Hirschberg 2012 Detecting hate speech on the world wide web. *In LSM*, (pp. 19–26).

Eskevich, M., Jong, F. D., Köing, A., Fišer, D., Uytvanck, D. V., and Heuvel, H. 2020 CLARIN: Distributed Language Resources and Technology in a European Infrastructure *In Proceedings of the 1st International Workshop on Language Technology Platforms*, (pp 28-34).