Seeking optimal Human/Machine collaborative practice in antisemitic terminology detection

Wendy Melillo
Journalism Division
School of Communication
American University
melillo@american.edu

Jessica Emami
Department of Sociology
College of Arts and Science
American University
jemami@american.edu

Solene Guarinos
Journalism Division
School of Communication
American University
sg9453a@american.edu

Dhanush Kikkisetti
Department of Mathematics
College of Arts and Science
American University
vk4372a@american.edu

Melanie Klein
Justice, Law and Criminology
School of Public Affairs
American University
mk4759a@american.edu

Lisa Liubovich
Data Science for Applied Public Affairs
School of Public Affairs
American University
ll3540a@american.edu

Raza Ul Mustafa
Department of Computer Science College of Arts and Science
American University
rmustafa@american.edu

Nathalie Japkowicz
Department of Computer Science
College of Arts and Science
American University
japkowic@american.edu

2024 IEEE Digital Platforms and Societal Harms (DPSH) | 979-8-3503-5130-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/DPSH60098.2024.10775254

***Abstract*—This study is concerned with the use of antisemitic language on loosely moderated extremist alt-right social media. It compares, contrasts, and combines human- and machine-based approaches for discovering antisemitic coded or non-coded terminology in such media. Coded terminology refers to the language used in closed-knit communities to communicate beliefs and attitudes without expressing them explicitly. In non-coded discourse, the views of the writer are expressed more explicitly. The detection and analysis of antisemitic terminology in social media is an important problem that can help uncover and monitor the evolution of societal attitudes towards the Jews, and eventually, towards other minority groups once we expand our study. Given the growing prevalence of antisemitic rhetoric in online public spaces, there is a greater need for academic studies to determine its presence and identify strategies to curb its influence. Scholars have noted how intolerant expressions can threaten democratic values. Antisemitic content, like other forms of hate speech, has the potential to radicalize people who may commit violence based on their prejudicial viewpoints. The purpose of this paper is to discuss three different approaches for discovering antisemitic terminology on social media. In particular, it describes an experiment in which three groups—human, software, and mixed—"competed" on the task of retrieving antisemitic terminology over a two-month period (January-February, 2024). The results were manually evaluated for relevance and frequency of occurrence in the next month period (March 2024). The study shows that each group discovered valuable terminology and that neither the human nor the machine group could replace the other. Instead, we advocate that both groups work together using their respective knowledge to optimize the discovery of new terminology on social media.**

Keywords—hate speech, coded terminology, human-machine interaction, antisemitism, far-right extremism

Disclaimer: This study describes antisemitic language that some readers might find offensive.

I. Introduction

In the last two years, American University's Signature Research Initiative project entitled "Unmasking Antisemitism" has combined human analysts with machine learning techniques to develop a more robust method for computers to detect subtle, indirect antisemitic terms and phrases used on the internet.

Our overall project focuses on a methodology that uses three interdisciplinary faculty researchers and their teams -- called the lexical study team, the population team, and the software team -- to analyze the life cycle of antisemitic terms and phrases as they are introduced in extremist alt-right social media and either cease to exist (see "skype" example below) or gain greater popularity and become used in more mainstream media. The purpose of the project is to identify more subtle, indirect antisemitic terms and phrases, identify their existence in mainstream media, and use machine learning to train content moderation algorithms to detect the more indirect language.

In this paper, we test the ability of two of our teams – the lexical study team representing the human coders and the software analysis team representing the computer methods – to work jointly in a combined approach to the problem. The lexical study team developed a coding statement and a methodology to guide the human analysis of posts, while the software team created AI programs aimed at conducting the analysis automatically. To date, the two teams' collaboration has been minimal. The purpose of this study is to move to the next stage, which consists of comparing and contrasting the effectiveness of both teams and starting the process of combining their efforts.

We initiated this study based on our assessment that the lexical study and software analysis teams have both reached the point where they have systematic ways of discovering coded antisemitic terminology, either through the existing coding methodology or a natural language processing model that can be tuned. Based on this assessment, we ask the following three questions:

- What is the effectiveness as well as the pros and cons of both approaches?
- To what extent do they overlap and diverge?
- What would be an optimal way to combine their strengths and mitigate their weaknesses?

To answer these questions, we set up a "competition" to pit three groups against each other on a shared task. Specifically, we scraped two months of posts (Jan-Feb, 2024) from alt-right extremist social media sites based on 24 antisemitic seed terms and asked the three groups to identify new coded antisemitic terminology from these posts. The groups were divided into one made of humans only—the *lexical study* group — one using AI only—the *software* group —and one combining both humans and AI—the *mixed methods* group. Once the results were submitted, the frequency of the terms retained by each group was calculated on posts scraped in the same manner during the following month (Mar 2024), and a team of human experts evaluated the results for relevance.

This proved to be a useful initial study to understand the nature of the searches performed by each group. In addition to outlining their different features, this exercise suggested better ways to interweave the two modalities to optimize their outcomes.

The remainder of the paper comprises five sections. Section II is a literature review. Section III discusses our overall methodology, also summarizing the approaches followed by the lexical study and software groups, respectively. Section IV presents the results we obtained, and a discussion follows in Section V. Section VI concludes the paper and suggests avenues for future work.

II. Literature Review

A. Hate Speech

With the rise of social media platforms, a concurrent increase in hate speech and racism online has fueled concern among government officials, policy-makers, scholars and citizens alike. Some scholars correlate the increase in hateful speech online with a recent resurgence in far-right leadership in countries like Brazil, India, the US and the UK among others. [1]

There is no universally accepted definition of hate speech. A general definition posits that hate speech is considered an attack on a person or group with the attacks particularly targeting members of minority groups. Speech that is considered "sexist, racist, xenophobic, ageist, fatphobic, or homophobic" among other types is classified as hate speech. [2] The United Nations further defines hate speech as offensive discourse targeting groups based on "inherent characteristics (such as race, religion or gender) and that may threaten social peace." [3]

While scholars still debate the effect of online hate speech on violent conflict, Ndahinda and Mugabe's study of its use on Congolese social media and the impact on anti-Tutsi discourse demonstrates how the dehumanizing content inherent in hate speech can legitimize violence against particular groups of people. [4]

B. Quantitative Approaches

Separate quantitative and qualitative approaches to studying online antisemitism each have their limitations. Quantitative methods that develop large-scale studies by examining millions of posts using a machine learning model may yield important insights based on frequency, but they cannot always detect the more subtle antisemitic expressions. Qualitative approaches using humans to identify the indirect language lack scalability given the speed and ubiquity of the Internet.

A sampling of recent work on antisemitism using quantitative methods includes Zannettou and Finkelstein et al. [5] where hundreds of millions of posts and images were collected from alt-right social media communities like 4chan's Politically Incorrect board and Gab to quantify the escalation of antisemitic rhetoric and memes on the Web. Chandra and Pailla et al. [6] applied a multimodal deep learning system to two datasets from Twitter and Gab to evaluate the efficacy of using machine learning to detect online antisemitism. Although not directly related to antisemitism, Braddock and Dillard's [7] use of meta-analysis to conduct empirical tests on the effects of narratives on beliefs, attitudes, intentions, and behaviors is relevant to understanding how antisemitic terms and phrases can be molded into radical ideologies that can prove harmful to democratic societies. Their results suggest a positive relationship between exposure to a narrative and narrative-consistent beliefs.

C. Qualitative Approaches

The social sciences literature is rife with studies on antisemitism from a socio-historical perspective. Recent work that is salient to our project includes lexical studies that focus on the relationship between online language and its influence on democratic societies. Rossini [8] argues that scholars should focus on online expressions of intolerance, which have a more detrimental effect on democratic societies than incivility. In her study of public comments on news websites and Facebook, she concludes that incivility in online political talk is associated with "meaningful discursive engagement" whereas intolerance is more likely to occur in homogenous online political discussions about minorities and civil society, which has a more harmful effect on democracy.

D. Mixed Methods Approaches

One project that is similar to our mixed methods approach of combining both qualitative and quantitative research methods to address antisemitism in online social media is worth noting. Becker and Bolton [9] use AI machine learning and qualitative content analysis to examine the frequency, content, and linguistic structure of online antisemitism in three European countries: the United Kingdom, France, and Germany. The study differs from our work by first identifying comments posted on social media following real-world incidents considered likely to stimulate antisemitic behavior and, although they acknowledge the difficulty of handling

coded terminology, they do not address its evolution the way our project does.

E. Other Observations

Finally, it's important, for this study, to mention the work of Whitney Phillips [10] in her book about online trolling culture. Phillips argues that such cultures use "highly stylized" language and behavior practices to "disrupt and upset as many people as possible." She notes that the trolling culture is an integral part of online social media, particularly in more extremist spaces where the discourse is frequently racist, sexist, and homophobic. Such language is often interspersed with calls for violence.

F. Limits of Common Research Approaches

Methodological differences among scholars who study hate speech also raise questions about the validity of trying to address hate speech through one academic approach as the work of Matamoros-Fernández and Farkas, previously cited above, suggests. The scholars found a "clear discrepancy" in the use of concepts in text-based analyses of the problem. Researchers using quantitative methods tended to use the term "hate speech" in their studies while other scholars using qualitative methods tended to search for and analyze the term "racism" in their work. "This points to a terminological divide in the field, indicating a lack of scholarly exchange between the humanities/social sciences and computer science/data science," Matamoros-Fernández and Farkas said in their study. By taking an interdisciplinary approach to our overall project, we seek to combine a sociohistorical approach with natural language processing and machine learning to address online hate speech.

III. Methodology

For this experiment to compare, contrast, and then combine the human and software teams, a shared task was created, and the parameters of the exercise were set. Finally, an evaluation protocol was established and applied to the results. Each aspect of our methodology is described in the next subsections.

A. Groups and Material

We divided the members of our teams into three groups: the lexical study group made up of three researchers; the software analysis group made up of a single researcher and the mixed methods group made up of two researchers, one with lexical analysis expertise and one with software expertise. The team members with lexical analysis consisted of faculty and student research assistants at the undergraduate, graduate, and Ph.D levels with a combined knowledge of communication, data science and Jewish studies. Selection was based on members who understood and used qualitative research methods such as content and textual analysis as well as understanding the socio-historical roots of antisemitism in the U.S. The researchers involved with software analysis on both the software analysis and the mixed teams had natural language processing and machine learning expertise as well as familiarity with software methods for combatting hate speech in general and antisemitism in particular.

The groups were given a task "to compete" on involving the discovery of "emerging terms," which in this study are defined as indirect, subtle, and covert expressions of antisemitism. The testing period took place between March 8, 2024, and March 26, 2024.

The lexical study researchers on both the lexical study and mixed methods groups used the methodology they had previously developed to identify antisemitic coded terminology.[1] Before the testing period, they collected data from the software company called Pyrra[2] and identified 24 coded antisemitic terms and phrases. That included previously recognized antisemitic terminology that was increasing in frequency, or words and expressions that had not yet appeared on other lists compiled by advocacy groups like the Anti-Defamation League. Some of the terms and phrases were directly antisemitic while others were more indirect where the exact meaning was not as obvious.

The software analysis researchers in both the software and mixed groups used several versions and combinations of the AI tools they had created, and that are described in [11] and [12]. The software group combined the approaches in [11] and [12] and submitted their results directly to the competition judges. The mixed group only used the software approach described in [11] and passed their results on to the human researcher for assessment before submission to the competition judges. In more detail, the system in [11] proceeds as follows: after some standard text clean-up, the approach extracts the terms found most relevant to the documents in the collection by calculating TF-IDF scores. These terms are then filtered according to whether they are grammatically coherent or not. A Large Language Model is then used to help establish the semantic similarity between the terms retained by the system and terms known to be antisemitic in nature (see [11] for details). The approach in [11] can be used in conjunction with that in [12] which, first, clusters the posts by topics before proceeding with terminology extraction. Many different parameters and settings were tried by the software team.

The mixed group researchers proceeded by having the machine generate candidate terms that the human researcher vetted using lexical analysis methods. Though the mixed

[1] The coding statement is available upon request.

[2] Pyrra is a company that scrapes posts from more than ten alt-right social media platforms including 4chan, Gab, and Truth Social. See: https://www.pyrratech.com/

team had a more thorough makeup than the other teams, given the time constraints it was not able to go as deeply into the lexical or software analysis as the two other teams. In particular, fewer parameters and machine settings were experimented with and only one instead of three human coders was available to conduct the lexical study on the candidates generated by the machine. This constrained the human coder who was not given any chance to extract terminology separately from the extraction conducted by the software.

B. Shared Task

The 24 seed terms used in this exercise that were identified by the lexical study team in the months prior to this study are:

1488, 6 Gorillion, Anudda Shoah, Bankster, Chaim, Coastal Elite, Cohencidence, DOTR (Day of the Rope), Fellow whites, Gas Yourself, Globohomo, Goyslop, Great War, GTKRWN (Gas the Kikes, Race War Now), Le Happy Merchant, Loxism, Oven Dodger, Pisrael, Shlomo, Skype, The Final Victory, WEF Puppet, Who Nose, Zionazi.

The data for this study was scraped from Pyrra, for a 2-month period going from January 1st to February 29th, 2024 using these 24 seed terms. This resulted in 47,186 posts distributed amongst the 24 terms. The distribution of posts per seed word is shown in Figure 1 where only the seed words that led to the retrieval of over 1% of all posts in the collection are shown.

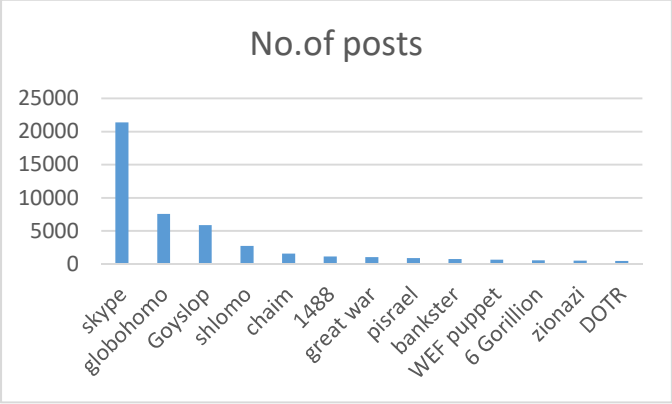

Figure 1: Frequency of posts for the 13 most frequent seed words

It is worth noting that the seed term "skype" presents an interesting example. Skype is a well-known telecommunications application similar to Zoom, but pre-dating it. Skype is also a coded antisemitic word given its similarity to the antisemitic slur "Kike" which gained popularity in 2016 when online users assigned code words to groups based on their race or ethnicity to avoid attempts by social media sites to remove offensive language [13]. For example, Yahoo was used to refer to Mexicans and Bing to Chinese people. Skype became the code word for Jews. As shown in Figure 1, "skype" is the seed term that yielded the largest number of posts, 45.31% of all posts to be exact. However, all three teams quickly realized that the great majority of posts in the "skype" data set referred to the app rather than the antisemitic slur. All teams, therefore, after spending a little bit of time on it ended up dropping that data from their analysis. This is an example of a new antisemitic term entering the life cycle and then quickly fading out over a short time period.

C. Parameters of the Exercise

All groups were initially given two weeks to conduct their analysis. Because of the researchers' different work schedules, everyone was asked to record the time they spent on their tasks. Due to personal circumstances, we extended the time when needed opting for greater accuracy over strict time control.

Although the lexical study team normally requires at least three coders and their agreements on decisions, for the interest of time and given the volume of data that needed to be processed, the task was divided amongst the three researchers who each analyzed different parts of the data. The lexical study group used the 24 emerging terms as "seed words" to manually identify other emerging terms from a sampling of 100 posts per seed term. The seed terms were manually entered in Pyrra and 100 posts per term were extracted from the search results.

The software analysis group identified emerging antisemitic terms from seed terms using the natural language processing models previously mentioned. The researchers tasked with handling the software could test different parametrizations of their approaches and combinations of the available software. For example, the software group tried several methods: one where the term extraction component was run on the entire data set; one where the data sets were divided according to the terms they stemmed from and processed separately unless the data sets were too small in which case they were combined. Though the software group was allowed to use the researcher's judgment in choosing what to submit as their final output, they were not allowed to perform detailed research concerning the terms they found in order not to veer toward the performance of the mixed group.

The mixed group was given the same parameters for the exercise as the software and the lexical study groups. Due to the need to work together and the fact that only one lexical study researcher was involved, however, the capacity of the mixed group was somewhat reduced. The mixed methods group worked by identifying emerging terms from seed terms using a natural language processing model, but the emerging terms were then verified by a human researcher and coded on whether they were antisemitic or not according to the lexical study group's coding statement.

D. Evaluation

After a 19 day period, the results were analyzed according to the following process.

Initial Filtering:

- The list of terms returned by the software team was purged of irrelevant content. This was done in a quick fashion by the three members of the lexical study group who scanned the list and noted the terms worth retaining.

Only terms judged relevant by all three reviewers (after discussions) were retained in what we call the filtered list or simply the list in the following discussion. The original software analysis list contained 815 terms. The filtered list retained only 129 of them.

- The lists of terms returned by the lexical study and mixed groups were considered relevant enough not to need the filtering required for the software results. The lexical study group returned 59 terms and the mixed team returned 40.

Frequency Analysis

- The terms returned by the lexical study and mixed groups as well as the terms present in the software analysis filtered list had their frequency in posts checked and listed in the month following the end of the competition (March 2024).
- A threshold of 20 posts using a term during the month of March was selected to indicate that a term was worth being considered.

Time Analysis

In order to obtain a thorough assessment of the three different modalities we are assessing – lexical analysis, software analysis, and mixed analysis—another criterion to consider is the time taken by the term extraction process by each modality. Setting a time limit of two weeks for the exercise is useful, but not sufficient since different researchers work for different numbers of hours each week. We therefore asked the different groups to keep track of their time (as well as of the machine compute time when relevant). It is clear that software-based approaches are more efficient than human-based ones, but we are trying to establish the extent of that efficiency and pit it against its performance with respect to human-based extraction.

Assessment Procedure

To assess the worth of the terminology returned by each of the three groups, the researchers from the lexical study group assessed the results obtained by each of the three groups according to a ranked scale. The group created the following ranked scale based on its original coding statement and by its definition of an emerging term. As mentioned before, in the coding statement, an emerging term is defined as coded, covert expressions of antisemitism. While emerging terms are intertwined with antisemitic tropes such as "Jews have too much power," emerging terms repackage historic hatred with modern themes.

- **4: Perfect**—the group got it right. It is an ***emerging***, ***antisemitic term*** as defined in the lexical study team coding statement.
- **3: Good**—it is an ***antisemitic term*** as defined in the lexical study team coding statement, but it is ***not an emerging*** term.
- **2: OK**—it is an ***antisemitic term*** as defined in the lexical study team coding statement, it may or may not be emerging, but it ***does not appear frequently enough as per the established threshold***.
- **1: Nonsense**—it is ***not an antisemitic term*** as defined in the lexical study team coding statement.

The scale can also be visualized according to the table below.

	Antisemitic	Emerging	Frequent
Perfect	**Yes**	**Yes**	**Yes**
Good	**Yes**	No	**Yes**
OK	**Yes**	*N/A*	No
Nonsense	No	*N/A*	*N/A*

In this study, we both extract antisemitic terminology and the post it appears in rather than simply classify a post as antisemitic. We do this for two reasons. First, an important component of the overall project is to study the life cycle of antisemitic terms and how that language evolves over time. One of the most common ways antisemitism continues to spread online is through the use of coded, veiled euphemisms that are not caught by content moderators. Using older antisemitic terms to find new ones can increase the likelihood of finding modified expressions of older antisemitic terms expressed in more current vernacular or other veiled expressions among like-minded people who are posting in more homogeneous social media groups. This approach builds on previous scholarly research where intolerance is more likely to occur in online discussions among more homogeneous groups as referenced in Rossini's work previously addressed in the literature review of this study. Both the linguistic expressions and the context in which they occur are of value to tracing how antisemitic language changes over time and continues to spread. Second, the context in which the term or phrase is used is also important when labeling a term or phrase antisemitic or not. This is clearly explained in the American Jewish Committee's Translate Hate Glossary which explains when a coded term should be considered antisemitic or not.[3] The researchers working on scaling the terminology and deciding whether it was coded or not worked in teams of two, and discussed areas of disagreement on all decisions, until a compromise could be reached.

IV. Results

We begin by reporting the terminology extracted along with some raw computations to get a sense of what the three teams uncovered. We then compute more sophisticated metrics to shed light on the pros and cons of each approach. Table 1 lists the extracted terminology considered perfect (4s) or good (3s) along with the group(s) that uncovered it.

[3] https://www.ajc.org/translatehate

Extracted Terminology

Rate	Terms and retrieval method: L, S, M
4: Perfect score	#FuhrerFriday (L), adl sewer/sewer rat (S,M), antisemitism strategy (S), Christkikes (L,M),controlled opposition (S), GLR (L), gorillion dollars (S), gorillion times (S), gorillion years (S), Goycattle (M), hasbara propaganda (S), Israhell (L), jew media (S), jew world (S), Jew York City (M), jewish family (S), jwo file (S), J0os or Joos (L), Jewdicial (L), Jewish Doctrine (M), JewJab (M), Jewkraine (L,M), jewtuber / Jewtube (L,M), Kike Shill (M), Moloch Worshippers (M), satanic jews (S), Satanyahu (M), Talmudvision (L,M), TKD (L), white genocide (S), world government (S), world order (S)
3: Good score	annuda/annudah/anuda shoah/shoa (S), bankster wars/wwwar/wwwars (S), fellow white (S), final victory (S), full 1488 (S), globohomo propaganda (S), great reset (S), great war (S), happy merchant (S), Hebe (L), Hymie (L), Kosher Nostra (M), Philo-Jew/Philo-Semite (M), rabbi Shlomo (S), race war (S), Sheeny (L), Shylock (L), Synagogue of Satan (M), wef puppet S), white power (S), wpww (M), Zogbot (L)

Table 1: Relevant terms returned by each team (L: Lexical, S: Software, M: Mixed) and their assessment as emergent (4s) or not (3s).

Raw Results

Table 2 lists a compilation of five results for each group: the total number of terms in the list of terms that each group identified as antisemitic; the number of terms from that list that were assessed as 4s (perfect) and 3s (good), respectively; and of those, the number of terms that were assessed as coded. For example, the lexical study group returned 59 terms. Of these, 10 were assessed as 4s and five as 3s. Two of the 10 terms assessed as 4s, and four of the terms assessed as 3s were further assessed to be coded.

Team	Number of Terms Retrieved	Relevant terms with scale value…		Number of coded terms with scale value…	
		4	**3**	**4**	**3**
Lexical	59	10	5	2	4
Software	129	15	13	11	10
Mixed	40	12	4	2	1

Table 2: Number of terms returned by each team and their assessment as relevant, emergent and/or coded.

Generally speaking, Table 2 shows that of the three groups, the software group is the one that returned the largest number of relevant, emergent and/or coded terms, but that happened at the cost of returning many irrelevant terms as well. In fact, as discussed before, the 129 terms returned by the software group represent the filtered version. Prior to filtering, the software group had returned 815 terms. This could have been reduced by requesting that it return fewer terms, but it is possible that with a smaller allowance, it would have returned fewer terms of interest. The data and mixed teams returned a similar number of terms of interest, each, but the mixed team had the advantage of returning fewer altogether, thus showing a better targeting than the other two approaches, especially since the number of relevant terms it returned is larger than that of the lexical study group (though fewer of these are coded).

Another piece of information not shown in Table 2 is presented in Table 3. It concerns the overlap of terms found by each team. The numbers in the diagonal of Table 3 show the number of terms found by each team (divided into terms rated as 4s or 3s) that were not found by the other team. Numbers in squares not on the diagonal are those that were found by both methods represented by the row and the column. For example, the table shows that the lexical study group found 10 terms rated as 4s, six of which were not found by any other group and four of which were also found by the mixed group. Of the twelve terms rated as 4s and found by the mixed group, as just discussed, four of them were also found by the lexical study group, and, in addition, one was also found by the software group. These, however, are the only overlaps found in the three lists of terms. This suggests that each team operates differently, each bringing its own value to the exercise.

	Lex4	Lex3	Soft4	Soft3	Mix4	Mix3
Lex4	6				4	
Lex3		5				
Soft4			14		1	
Soft3				13		
Mix4					7	
Mix3						4

Table 3: Amount of overlap among the three teams' results

The next table, Table 4, looks at the amount of time taken by each team. Surprisingly, the software group spent the greatest amount of human and machine time as they experimented with many different configurations of the software. This is the result of a personal choice made by the software group member and it will be interesting, in the future, to assess the extent to which this amount of time could be reduced without affecting the results.

	Time (hours)	
	Human	Machine
Lexical	28	N/A
Software	45	26
Mixed	41	6

Table 4: Time spent by each team (human and machine time) on the retrieval of antisemitic terminology

Precision and Recall Results

Precision, Recall, and F1-Score are metrics used widely in the field of information retrieval. Precision measures how accurate the process of retrieving information is. In particular,

in our case, it asks: "What proportion of the terminology returned by the approach represents a relevant antisemitic

term?" Recall asks a complementary question related to the information it could have retrieved but didn't. In our particular case, it asks: "Of all the relevant antisemitic terms that could have been returned by our approach, how many were actually returned?" The formulae for precision and recall follow:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

The F1-Score combines precision and recall by giving them equal weight. The formula for the F1-Score follows:

F1-Score = 2 * Precision * Recall /(Precision + Recall)

In all three formulae, TP, FP, and FN correspond to the true positive, false positive, and false negative entries of a confusion metric, respectively.

We consider two specific questions:

- How did each team fare on the problem of discovering ***emerging*** antisemitic terminology?
- How did each team fare on the problem of finding both ***emerging and non-emerging*** terminology?

For the first problem, TP represented the number of expressions returned by the method that were assessed as 4s on the lexical study group's scale. For the second problem, we considered both 4s and 3s as the instances representing TPs. Once the TPs were set, both FPs and FNs needed to be calculated as well. FPs were easy to establish as they were based on the following formula where PP represents the number of instances returned by the method, i.e., the Number of Terms Retrieved in Table 1:

FP = PP - TP

FNs were a bit more complex to calculate because it was first necessary to establish the list of all the antisemitic expressions in the data set. We estimated that number by taking the union of terms found by all three teams. When considering the emergent antisemitic terms only, that meant adding the numbers found in columns Lex4, Soft4, and Mix 4 of Table 2, which resulted in AT = 32. When considering emergent and non-emergent terminology, that meant adding all the numbers together, which resulted in AT = 54, where AT stands for actual antisemitic terminology. FN was then calculated by the following formula:

$$FN = AT - TP$$

which counts the number of times the approach failed to recognize antisemitic terminology.

It is important to note that our calculation of AT is an underestimate since it assumes that together, the three groups discovered all the antisemitic terminology there is to be discovered in the data set. Yet, that is not guaranteed. This means that the recall numbers shown in the following table are overestimates of the true numbers. Still, they are significant for their relative values amongst the three groups.

Precision and Recall and F1-Score results are listed in Tables 5 and 6, respectively. Table 5 reports the results obtained for emergent antisemitic terminology (4s only) and Table 6, those obtained for emergent or non-emergent terminology (4s and 3s).

	Precision	Recall	F1 Score
Lexical	0.17	0.31	0.22
Software	0.12	0.47	0.19
Mixed	0.3	0.38	0.33

Table 5: Results for coded or non-coded emergent antisemitic terminology

	Precision	Recall	F1 Score
Lexical	0.25	0.28	0.27
Software	0.22	0.52	0.31
Mixed	0.4	0.30	0.34

Table 6: Results for coded emergent or non-emergent antisemitic terminology

Both tables show that, overall, the mixed group obtains the best results. That is much more salient in Table 5 than in Table 6 where the F1 Scores are closer to one another. In both tables, however, it is worth noting that the software team obtains remarkably high recall values (though low precision values).

It is worth noting that considered in isolation, the precision, recall, and F1 scores returned in Tables 5 and 6 are very low. This, however, is not too concerning for several reasons. First, the discovery of coded hateful terminology is only one step in the process of combatting hatred. Low precision is, therefore, not too concerning given that terms wrongly assessed as hateful and/or coded would inevitably be identified as erroneous in further steps. The low recall values coupled with the small amount of overlap between groups previously reported illustrate very well the fact that man and machine need to collaborate closely and creatively to do a thorough job of identifying hateful coded terminology.

V. Discussion

Usefulness of the discovered terminology

To illustrate the value of both extracting terminology and retrieving the context in which a word or phrase appeared, we provide the following example of more antisemitic terms found when using the phrase "oven dodger" as a seed term in the search. The terms "hebe", "hymie", and "sheeny" listed in Table 1 above appeared in a February 17, 2024 post where the writer said, "You can say... kike, bagel-biter, draedel-spinner, hebe, hymie, sheeny, yid, shekel-chaser, shylock, shekel nigger, oven dodger, oven fodder, and Jew here, and nobody will care." While it can be argued that "hebe" and "hymie" are more direct antisemitic terms, "sheeny" is more indirect and subtle. In another example, a different search

using the seed term GTKRWN (Gas the Kikes, Race War Now) revealed the acronym TKD (Total Kike Destruction) as a new way to refer to GTKRWN. An October 29, 2023 post found during the testing period said, "I want GTKRWN, which apparently the kids are now calling TKD..."

Emerging terms are important to study because they show how antisemitic language and hate speech more generally evolve into new words, phrases and expressions to match the cultural zeitgeist of the times. Just focusing on existing terms that have already been identified by groups that study antisemitism is insufficient because such lists fail to capture how language changes over time. Given how trolling culture in online social media uses "highly stylized" language and practices as Phillips notes in her book, it's necessary to trace where possible the changes over time. Such work can also help train computers to more effectively detect the changes to improve content moderation practices. Practically speaking, capturing emergent terminology is very useful since that terminology can subsequently be used to scrape relevant new posts, which can then be analyzed to yield newer terminology and so on. This may allow content moderators to stay on top of the evolving discourse rather than constantly playing catch-up.

Surprising Results

Though many of the results we obtained were expected, two struck us as unexpected. First, we had expected that the results discovered by the three groups would mostly overlap, with only a few extra terms found by individual groups. Instead, the overlap is minimal, which seems to suggest that each group is useful in its own right. For example, the terms "antisemitism strategy" which implies that Jews invented the concept of antisemitism as a "strategy" to self-victimize/garner sympathy; GLR, which stands for George Lincoln Rockwell, founder of the American Nazi Party; and Satanyahu, which combine Satan and Netanyahu and were discovered by the software, lexical study, and mixed groups, respectively are all useful new terminology to have discovered.

The second surprise is the time spent by each of the groups on the exercise. Our assumption had always been that humans could not keep up pace with the rate at which social media posts were disseminated. Yet, the amount of time spent by the software group was commensurable with that spent by the human groups and both returned valuable results. Now there are some reasons why this happened. First, the task was purposely selected for its human tractability. Furthermore, the lexical study group used a sampling of 100 posts per seed word, thus ignoring a large number of potentially relevant posts. Second, the software group tried a lot of different versions of the approach and thus spent a lot of time changing the code. This will be minimized in the future. Third, and related to the first point, is the question of scalability: if we were to continue searching, say, for the month of April, the amount of time needed by the software team would be much smaller than that required by the lexical study team, given that the system has already been optimized for the task. So, we believe that this second surprise will not hold in the long run and that computer-aided search is necessary.

Pros and Cons of the three approaches

We now discuss the pros and cons of each of the three approaches in detail, keeping in mind that each of them retrieved relevant and independent terminology of interest.

The lexical study group did not use any software and yet spent the least amount of time on the exercise as just discussed. Though it is the approach that retrieved the smaller number of relevant terms, it was followed closely by the mixed approach. It also retrieved more coded terminology than the mixed method. Its F1-score is close to that of the software approach, though the lexical study approach has higher precision and lower recall.

The software approach used a lot of computer and human time and boasted the largest number of relevant terms retrieved. It also found much more coded content than the other two approaches (between three and four times more than the mixed method; and seven times more than the lexical study method). The F1-Score it returns is the lowest when considering emergent terminology only and fits between the lexical study and the mixed approaches when considering both emergent and non-emergent terminology. It has a great advantage over both other methods on recall but fares quite badly on precision. Another serious disadvantage is that it returns a lot of "garbage" since as we recall, it initially returned 815 terms that had to be filtered in the first phase by the data team who ended up keeping only 129 terms for the comparative exercise.

Of the three approaches, the mixed one seems the most balanced and, indeed, returns the best F1-Score on both the emergent only and the emergent or non-emergent tasks. On the emergent-only task, the F1-Score is noticeably higher than that of the two other approaches. The number of terms retrieved is very manageable and the time taken by the exercise was reasonable.

Limitations of our study

The title of our study stated that we were seeking an optimal Human/Machine collaborative practice. Though we have, indeed, discovered that the optimal set-up for this kind of work is a mixed team made up of human lexical analysts and software design experts, we have to recognize that we have not yet fully identified the best ways for these experts to collaborate. In addition, there are several other limitations worth mentioning. First, the test was conducted in a limited fashion: the database of posts on which the search took place spanned only two months of social media posts while the database used to validate the findings spanned only one month. Furthermore, the teams were given a short timeframe (fewer than three weeks) to sift through the two months of posts. While the software and mixed teams were able to process all the posts, the lexical analysis team had to restrict itself. It chose to use samples of 100 posts per seed word, a choice that may have limited their findings. Second, several

parameters were set to conduct the study. For example, a threshold of 20 posts was established to consider a new term as worth considering. In the future, it would be worth investigating the effect of such a choice. A systematic parameter sensitivity analysis could also be performed to analyze the effect of parametric choices on the software system.

VI. Conclusion and Future Work

This study considered the problem of discovering emergent and non-emergent, coded and non-coded antisemitic terminology in extremist social media. Three groups were set up that considered the same task of searching for relevant terminology over a two-month period, but each using a different modality. The lexical study group did not use any software for their search, relying only on human capabilities. The software group used AI-based tools and used very limited human judgment. The mixed group was allowed to use both AI-powered technology and human knowledge of the problem to fulfill the task. As expected, the mixed team outperformed the other two, although the exercise left us with two surprising results: the terms retrieved by each group overlapped only minimally, if at all, with the other groups. and the software group used much more time of human work than expected. We believe that these surprises are due to the fact that our work is still preliminary and we conclude, from this study, that the best avenue to follow, in the future, is the mixed approach that we will need to refine along the way.

In the future, we, thus, propose to pursue the mixed method and refine it in the following ways:

We would like the lexical study group to instruct the software group as to what modalities of their software are most useful so that the coder involved in this exercise can limit the time spent on fine-tuning the approach.

We suggest that the humans involved in the mixed group use the software to guide them, but also do some independent searches. Perhaps, an approach could be to not quickly filter the initial list suggested by the software as was done here, but instead, to consider it more carefully, going back to the posts the suggested content comes from. While a targeted search focused on the terminology the software considers relevant can be useful, a less targeted human search may be able to retrieve content that the AI approach doesn't consider relevant.

Once we have determined an optimal human/machine collaboration regimen, we will test our resulting approach on other kinds of hatred in extremist social media, such as Islamophobia, and hatred against Asians, BIPOC and other ethnic groups.

Acknowledgment

We gratefully acknowledge the financial support of American University's Signature Research Initiative Project.

References

[1] Matamoros-Fernández, A. & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. Television & New Media, 22(2), 205-224.

[2] Di Fátima, B. (2023). Hate speech on social media: A global approach. LabCom Books & EdiPUCE, 5.

[3] United Nations. Understanding hate speech. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[4] Ndahinda, F. M. & Mugabe, A. S. (2022). Streaming hate: Exploring the harm of anti-Banyamulenge and anti-Tutsi hate speech on Congolese social media. Journal of Genocide Research, 71.

[5] Zannettou, S., Finkelstein, J., Bradlyn, B. and Blackburn, J. (2020). A Qualitative Approach to Understanding Online Antisemitism. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media,* pp. 786-797.

[6] Chandra, M. and Pailla, D., Bhatia, H., Sanchawala, A., Gupta, M. Shrivastava, M., Kumaraguru, P. (2021). "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. *Proceedings of The 13th ACM Web Science Conference*, pp. 148-157.

[7] Braddock, K. and Dillard, J. (2016). Meta-analytic Evidence for the Persuasive Effect of Narratives on Beliefs, Attitudes, Intentions and Behaviors. *Communication Monographs*, vol. 83, pp. 446-467.

[8] Rossini, P. (2022). "Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, vol. 49, pp. 399-425.

[9] Becker, M. and Bolton, M. (2022). The Decoding Antisemitism Project –Reflections, Methods and Goals. *Journal of Contemporary Antisemitism*, vol. 5, pp. 121-126.

[10] Phillips, W. (2016). *This is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. MIT Press, 2016, pp. 2.

[11] Kikkisetti, D., Ul-Mustafa, R., Melillo, W., Corizzo, R., Boukouvalas, Z., Gill, J., & Japkowicz, N. (2024). "Coded Term Discovery for Online Hate Speech Detection", *Proceedings of the 11th IEEE International Conference on Data Science and Advanced Analytics (DSAA'2024)*.

[12] Mustafa, R.U. and Japkowicz, N. (2024). "Monitoring The Evolution Of Antisemitic Hate Speech On Extremist Social Media", *Proceedings of the 4th IEEE Conference on Digital Platforms and Societal Harms (DPSH'2024)*..

[13] Magu, R and Luo, J. (2018). Determining Code Words in Euphemistic Hate Speech Using Word Embedding Networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.