Adversary Detection

Ariel Cann, Menachem Jacobs, William Feller

August 2024

1 Introduction

While much research has been done regarding the ability of Natural Language Processing techniques to classify a message from social media as either hate-speech or acceptable on a semantic basis, and in a few of these instances, even to classify a message as antisemitic or acceptable, there remain two questions left largely unanswered:

1. Can a system capable of identifying some messages as antisemitic use these messages to learn new words or phrases that are subtly antisemitic, and thereby obviate existing systems that require large amounts of training data to detect the fact that a new slur has been invented?

2. Can the information produced by an adaptive system such as the one described be put to use classifying accounts, rather than messages, as antisemitic? Conceptually, this is a difficult problem. Does one antisemitic message mark an account as completely antisemitic? Can such a system account for people with obvious connections to antisemites on the platform?

These questions bring us to the next steps after semantic message classification: Using the classified messages to detect accounts that are overtly antisemitic, and using commonalities among these accounts to discover accounts that are provably, but not obviously, antisemitic.

To do this, we have developed a feature set which makes use of an account's subscriptions along with its own messages to provide a score, representing the confidence that a particular account can be determined to be antisemitic. This scoring of an account begins to answer Question 2, at least, in the case of overtly antisemitic accounts. To answer Question 1 (identification of new slurs and code words), we have developed a system to compile lists of words and of phrases that appear disproportionately in the messages of known antisemites as compared to the messages of a platforms overall user base (a kind of TF/IDF analysis), these words being another way to identify accounts as covertly antisemitic.

2 Scoring: What do we do when we are talking.

Early in designing the account scoring system we ran into a logical problem. If each account will be judged in part by the accounts it is subscribed to, that is, if each account's score is dependent on the scores of its subscriptions, then to begin scoring any account would require having already scored some number of prior accounts. This makes logically impossible the existence of a "first account scored."

To overcome this difficulty, each account has been given two different scores, a "secondary score," dependent only open the behavior of the account itself, and a "primary score," which combines the secondary score of an account with the secondary scores of its subscriptions.

2.1 Overt Antisemites: Secondary Scores

The secondary score is the confidence rating of a classifier that a given account is antisemitic on its own merits. This classifier uses four features compiled from the account to make this judgment:

1. Average Message Score. This is the sum of the sentiment scores for all messages posted by an account, divided by the number of messages. It is the simple average of the sentiment scores for an account.

2. Message Score per Day. This is the sum of the sentiment scores for all messages posted by an account, divided by the number of days between an account's earliest message and its latest message.

3. Rate of positivity. This is the fraction of an account's messages that the sentiment classifier supposes to be non-antisemitic.

4. Message Score by Density: This is by far the most complicated feature. It is given by the procedure below.

Let P be a number of days, equal to the total number of days between an account's earliest and latest message divided by the floor of log1.5 of the total number of days.

Let M = $\{M_1, M_2, M_3, \dots\}$ be the set of all messages in an account, order by date of post. Let T = $[T_1, T_2, T_3 \dots T_k]$ a list of time spans, each of length P, which exhaustively and non-overlappingly cover the total time. Let T(M_n) be the span which M_n is in, and let Size(T_n) represent the number of messages posted in the span T_n.

$$\text{MSD} = \frac{\sum_{n=1}^{k} M_n * Size(T(M_n))}{Total\ number\ of\ Days.}$$

2.2 Overt Antisemites: Primary Scores

The formula for the "primary score," is calculated as follows;

$$\frac{\sum_{i=0}^{n} \frac{secondary_i}{n} + 2 * suspect}{3}$$

Where n is the number of subscriptions in the account and suspect is the secondary score of the account. Twice the secondary score of the suspect is added to the average secondary score of its subscriptions, and the total is divided by three. Accounts with a high enough primary score (we are using a preliminary estimation of 0.8) are considered to be overtly antisemitic and will be used to create feature lists for uncovering covertly antisemitic accounts in a later step.

2.3 Covert Antisemites: Scoring by Commonalities

In the course of this paper, we refer to accounts as belonging to one of four partly overlapping classes. As mentioned earlier, an account may, by virtue of its messages and subscriptions, be considered an overtly antisemitic account, shortened to "Overt Account" in the rest of this paper. All accounts not Overt are considered to be "Suspicious Accounts." Suspicious accounts are further divided into two classes. Accounts not antisemitic, hereafter termed "Positive Accounts," and accounts covertly antisemitic, hereafter termed "Covert Accounts." After overt accounts have been identified, (that is, accounts with a primary score of 0.8) the following features are extracted from them, to be used in identifying covert accounts:

1. Vocabulary. These are words and phrases used with disproportionate frequency by overt accounts as compared to suspicious accounts.

2. Timing. These are dates when posts by overt accounts are disproportionately coincident as compared to suspicious accounts.

The word "Disproportionate" takes on two distinct and simultaneous meanings here. That is to say, there are two different versions of both Vocabulary and Timing being tracked. One version of the features considers disproportion to be the case where the number of occurrences in one group of accounts is absolutely greater than the number of occurrences in the other of accounts. We refer to this as "Absolute Disproportion." The other considers disproportion to be the case where the number of occurrences in one group per account, that is, the number of occurrences divided by the number of accounts in the group, is greater than the number of occurrences per account in the other. We refer to this as "Normalized" or "Comparative Disproportion."

The feature set of an account has a field for each of these, tracking the number of times an account has used some bit of vocabulary or has posted upon a date disproportionately associated with the overt accounts. In addition to these Six features, Absolute and Comparative, Words, Phrases and Dates, we add a seventh and eighth. The seventh feature is referred to as Networking. This is a count of the number of occasions upon which a suspicious account has engaged with overt accounts. The eighth feature is simply the primary score of the account.

These eight features gathered for an account are then fed into a classifier trained to the task, which produces a rating as to its confidence that the given account is antisemitic. Suspicious accounts rated as probably antisemitic are considered to be covert accounts. Covert accounts identified in this way have their secondary score increased, thereby increasing their own primary score, and the primary score of any accounts subscribed to them (and those to them, and so on).

3 Limitations

The primary limitation of this model is that it assumes that only an account's subscriptions, not its subscribers, are significant in scoring the account itself. Though there is some logical justification for this, in large part this decision was made to avoid the problems created by cycles in a subscription network, and the computational burden that detecting and solving them would entail. Therefore, if an account belongs to someone who is subtle in their antisemitic content, such that their secondary score is low, but whom a human observer would easily flag as antisemitic due to the sheer number of antisemites following them, this model will be prone to false negatives.

A second limitation has to do with scale. As many of the features are interrelated, most obviously with the secondary score of one account affecting all of its neighbors primary scores, which are used to generate account scores, the system will become more complicated more quickly than the linear growth of accounts under review. This is compounded by the need to search through all accounts within the sub-network in order to determine whether a given element exists. Although this problem can be solved to some extent with good programming practices and data-structures, these are heuristic solutions. Ultimately, the work is extremely computationally intensive at scale.

Finally, a distinctive element of tweets is the ability to retweet. Our system is not capable of taking this into account. A system that could do so would have access to an additional feature we don't have recourse to.

4 Hyper-parameters

par Related to the rather limited feature set is the question as to what hyper-parameters can be used in this model. Hyper-parameters related to secondary scores may include the following:

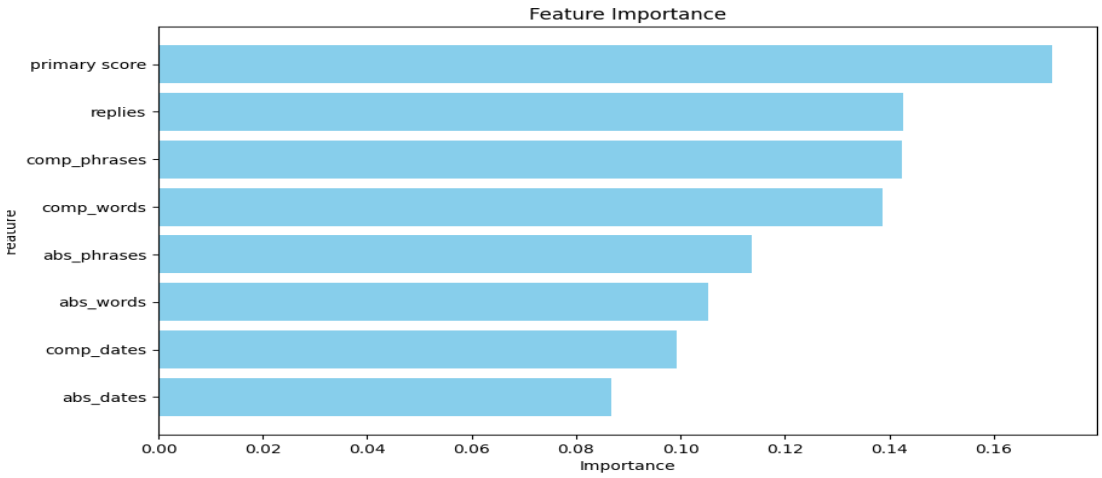
1. How many partitions to make when evaluating Message Score by Density, the number log1.5(Total Days) may be subject to change.

2. The maximum size of a phrase, as determined by the n-grams created in the search through among overtly antisemitic account messages; code phrases may be longer than bigrams.

3. The threshold regarding primary scores at which point the account is considered overtly antisemitic.

5 Data Synthesis

In order to meaningfully test our system, even in theory, we needed to control certain variables when synthesizing account data. We ensured that overt accounts subscribed to both overt accounts and suspicious accounts, but more frequently to the former, and vice-versa for suspicious accounts. Covert accounts had equal chances of subscribing to overt or suspicious accounts. Messages classified as antisemitic with a confidence higher than 0.5 (using a Random Forest Classifier by a different team) were considered overtly antisemitic, while the rest were randomly divided 66.6% into normal messages and 33.3% into covertly antisemitic messages. In order to detect code words out of context, we tested with code words appearing in all account types, but at different frequencies. Overt messages had fabricated code-words inserted at 10% of possible positions in each message, and bi-grams inserted at 2% of possible positions. Words were inserted at 5% of positions for normal messages while bi-grams were inserted at 1%. For covertly anti-Semitic messages, words were inserted at 7.5% and bi-grams at 1.5%. Messages were made more likely to be replies to subscriptions than to anyone else, and therefore anti-Semites were more likely to reply to each other than to others. Likewise with significant dates, 6% of message dates were assigned chosen significant dates for overt messages, 4% for covert messages, and 2% for normal messages, as normal accounts would also have somewhat heightened activity on dates (like October 7Th) which were significant offline, if not as much as anti-Semitic accounts. Using such conditions, the classifier we used to uncover covertly antisemitic accounts weighted the features we collected as such:



Interestingly, Primary Score is considered the most important feature, however, other features still receive significant enough weight that we believe them to be useful.

6 Conclusion

We believe that Adversary Detection is not only theoretically possible, but that we have actually done it. While Primary Score may be the most heavily weighted feature, and thus our highest indicator of covert antisemites remains the probability of overt antisemites, this may simply be indicative of reality, and the consistent weight of other features may indicate which lines of questioning are most fruitful for future research into this topic.