Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

It's a Thin Line Between Love and Hate: Using the Echo in Modeling Dynamics of Racist Online Communities

Eyal Arviv, Simo Hanouna, Oren Tsur
Software and Information Systems Engineering
Ben Gurion University of the Negev
{eyalar,hanouns}@post.bgu.ac.il, orentsur@bgu.ac.il

Abstract

The `(((echo)))` symbol – triple parentheses surrounding a name, made it to mainstream social networks in early 2016, with the intensification of the U.S. Presidential race. It was used by members of the alt-right, white supremacists and internet trolls to tag people of Jewish heritage – a modern incarnation of the infamous yellow badge (Judenstern) used in Nazi-Germany. Tracking this trending meme, its meaning, and its function has proved elusive for its semantic ambiguity (e.g., a symbol for a virtual hug). In this paper we report of the construction of an appropriate dataset allowing the reconstruction of networks of racist communities and the way they are embedded in the broader community. We combine natural language processing and structural network analysis to study communities promoting hate. In order to overcome dog-whistling and linguistic ambiguity, we propose a multimodal neural architecture based on a BERT transformer and a BiLSTM network on the tweet level, while also taking into account the users ego-network and meta features. Our multimodal neural architecture outperforms a set of strong baselines. We further show how the use of language and network structure in tandem allows the detection of the leaders of the hate communities. We further study the "intersectionality" of hate and show that the antisemitic echo correlates with hate speech that targets other minority and protected groups. Finally, we analyze the role IRA trolls assumed in this network as part of the Russian interference campaign. Our findings allow a better understanding of recent manifestations of racism and the dynamics that facilitate it.

Introduction

Hate speech proliferates in social media (Waseem and Hovy 2016; Laub 2019). While harassment may be targeted at any individual, hate speech typically references groups and targets individuals for their group identity. Women, people of color, the LGBT community, Muslims, immigrants, and Jews are among the most targeted groups. Recent studies report on a surge in Islamophobia (Sunar 2017; Akbarzadeh 2016; Osman 2017), Antisemitism (ADL 2020; Zannettou et al. 2020), xenophobia (Iwama 2018; Entorf and Lange 2019) and hate toward other groups (Dodd and Marsh 2017; Edwards and Rushin 2018; Perry et al. 2020).

Online hate speech is not merely an online inconvenience. It directly manifests itself in "real life" through shooting, bombing, stabbing, beating, and vandalizing. These violence incidents are often linked directly to online activity (Hankes and Amend 2018; Munn 2019; Malevich and Robertson 2019; Thomas 2019). A recent U.N. report on the Freedom of Religious and Belief, transmitted by the Secretary General amidst the global rise in antisemitism, asserts that "*antisemitism, if left unchecked by Governments, poses risks not only to Jews, but also to members of other minority communities. Antisemitism is toxic to democracy*" (Shaheed 2019).

Like misery, racial hate likes company. Hate is not expressed and promoted by random individuals – rather, it is the product of *communities*, often embedded in larger communities. Therefore, in order to better understand the social mechanisms involved in the promotion of online hate, we need to understand how these communities are structurally organized and how they use specific, often elusive, language to promote their cause.

In this paper we combine structural analysis of Twitter networks with textual analysis (NLP) to identify hate communities that are embedded in broader networks. Using the *echo*, an elusive antisemitic meme, as a starting point, we demonstrate how networks promoting hate can be recovered, which in turn enables us to identify key features that contribute to the promotion of hate.

Specifically, we address the following questions:

- **Disambiguation** – can we properly identify hate speech and disambiguate the uses of nuanced language and memes that are used both legitimately and in a dog-whistling manner?
- **Text and social structure** – can we leverage the network structure in order to achieve detection of hate-mongers?
- **Intersectionality of hate** – how is the antisemitic echo meme related to other forms of hate speech and other minority groups?
- **Linguistic variation** – can the evolution of the echo meme be interpreted according to linguistic theory?

The two former questions are addressed algorithmically in both unsupervised and minimally supervised way, as we propose a multi-modal neural architecture based on a BERT

Transformer and a BiLSTM for the utterance (tweet) level that feeds into another classifier that also models the user ego-network and metadata. The two latter questions, are addressed qualitatively.

Language Warning Some of the examples included in this paper contain offensive language. All examples are taken from the data.

Background and Related Work

The Echo The triple parentheses, or triple brackets, also known as the `(((echo)))`, is an antisemitic symbol that is used to highlight the names of individuals of a Jewish background (e.g., Jeffery Goldberg, Editor-in-Chief of The Atlantic), organizations owned by Jewish people (e.g., Ben & Jerry's), or organizations accused of promoting "Jewish globalist values" (e.g., the International Monetary Fund). Originally an audial meme used at the podcast *The Daily Shoah*, the meme was popularized in a textual form in the white-supremacy blog *The Right Stuff*. The echo slowly drifted from fringe websites to mainstream social platforms like Twitter, reaching a wider audience and expanding its user base. Typical examples of an antisemitic use of the echo are presented in rows 1-4 in Table 1. Tweets 1,2 were posted by regular users, referring to an individual (#1), and promoting an antisemitic trope about Jewish domination (#2). The 3rd tweet, promoting a similar antisemitic trope, was posted by a high profile organization – the official `@WikiLeaks` account, after the organization was criticised for alleged ties to the Russian Intelligence. It was retweeted hundreds of times before it was deleted within a few hours.

Members of hate communities often use specific language and symbols to convey their affiliation and promote their agenda. Unique, vague and ambiguous patterns of language may arise from community culture and are often used as a dog-whistling practice used in order to avoid detection and suspension[1]. While used as a hate symbol by some users, the echo has multiple senses, e.g., 'broadcasting', 'emphasis' or a 'virtual hug' (see Table 1 8-10). In Section we further discuss special lingo and ambiguous terms.

The recent rise in online hate speech attracts significant body of research. Broadly speaking, this body of research could be broken to two main categories, focusing on two different perspectives: (i) the algorithmic detection of hate-speech, and (ii) social analysis of the use (and users) of hate speech. In the remainder of this section we provide brief survey of relevant work.

Hate, Trolls and Online Culture The *alt-right*, short for 'the alternative right' is a term referring to a collection of organizations and individuals sharing extreme right-wing ideology that ranges from classic far-right ideology to open white-nationality and white-supremacy. While traditional Internet trolls are not promoting a specific ideology (Phillips 2015), alt-right trolls, rooted in Internet culture, seek to promote an extreme political agenda (Hawley 2017). The similarity between gamer-gate trolls and the online activity of members of the alt-right is explored in (Bezio 2018).

Hate speech is especially habitual in Gab and some forums on 4chan and Reddit (Hine et al. 2017; Nagle 2017; Zannettou et al. 2018; Grover and Mark 2019). These platforms support a community structure in an almost explicit way[2] and users adopt specific language to signal their affiliation and further enhance community bonds (Tuters and Hagen 2019; Zannettou et al. 2020). On Twitter, on the other hand, communities are formed implicitly, as individuals follow or engage with other (like minded) individuals, thus the habit of signaling affiliation through the use of specific language and memes is of increased significance. However, since Twitter is more tightly moderated than 4chan, Reddit or Gab, the use of language tend to be more nuanced.

Detection of Hate Speech The use of ambiguous words, coded language, and dog-whistling pose significant challenges to text-based detection of hate-speech (Davidson et al. 2017; Ribeiro et al. 2017). The detection of implicit forms of hate speech is addressed by (Gao, Kuppersmith, and Huang 2017), and (Magu, Joshi, and Luo 2017) detects the use of hate code words (e.g., google, skype, bing and skittle for Black, Jew, Chinese, and Muslim, respectively).

The use of demographic features such as gender and location in the detection of hate speech is explored by (Waseem and Hovy 2016), and user meta features, e.g., account age, posts per day, number of followers/friends, are used by (Ribeiro et al. 2017).

Computational methods for the detection of hate speech and abusive language range from the classic machine learning approaches such as SVM and logistic regression (Davidson et al. 2017; Waseem and Hovy 2016; Nobata et al. 2016; Magu, Joshi, and Luo 2017), to neural architectures such as RNNs and CNNs (Gambäck and Sikdar 2017; Zhang, Robinson, and Tepper 2018; Del Vigna12 et al. 2017; Park and Fung 2017), and BERT transformers (Mozafari, Farahbakhsh, and Crespi 2019; Samghabadi et al. 2020; Salminen et al. 2020). For comparative surveys of taxonomies of hate speech and abusive language, available datasets, and models see (Salminen et al. 2018), (Chen, McKeever, and Delany 2019), and (Wullach, Adler, and Minkov 2020).

The diffusion of hate in Twitter and Gab is modeled by (Ribeiro et al. 2017) and (Mathew et al. 2019), respectively. These works are close to our work as they address the user level, taking into account user meta features and network structure. However, the user meta features and network features are fixed and the textual analysis is basic. In contrast, we are concerned with the classification task rather than explicitly modeling the diffusion process. We put emphasis on the text, combining a BERT Transformer and a Bi-LSTM

[1] While some social platform, e.g., Reddit, 4chan and Gab (Zannettou et al. 2018; Lima et al. 2018) have limited or no moderation, platforms like Twitter officially prohibit hate speech.

[2] 4chan and Reddit communities are defined by the boards and subreddits they subscribe to (e.g., 4chan/pol and reddit/altright). While the design of Gab is similar to Twitter, it brands itself as the "free speech" platform, thus attracts users that are banned from other networks for promoting hate.

	LABEL	TWEET
1	HM	Don't Trigger Mr Trump (((Rosengerg))) it might cause him to fire up the ovens #OvenWorthy
2	HM	RT @USR: Andrew Breitbart was murdered by (((Globalists))). #PizzaGate
3	HM	Trybalist symbol for establishment climbers? Most of our critiques have 3 (((brackets around their names))) & have black-rim glasses. Bizarre.
4	HM	That's because Trump doesn't hate white Gentiles like ((((((((((THEY)))))))))) do.
5	R	ADL adds (((echo))) symbol to hate list
6	R	People are putting (((echoes))) around their names on Twitter - here's why
7	R	@USR alright wise one... What does ((( ))) around someone's name?
8	N	We're (((LIVE))) on the radio near you --> its #LightOnLive with <NAME>, from now till 6am on #Live919FM
9	N	@USR THIS WOMAN NEEDS A BIG HUG (((HUG)))
10	N	can u get any cooler than that ((((nope))))

Table 1: Echo tweets and their type. HM: hate-mongering; R: response to HM. N: Neutral (not hate); User names and real names were replaced by @USR and NAME, respectively. A tweet containing expressive lengthening can be seen in the fourth example of this table.

to classify users, and boost the model confidence by taking into account a weighted score assigned to other users in their network. The work of (Zannettou et al. 2020) is similar to ours in the sense that it is tracking a specific antisemitic meme ('the Happy Merchant') and address the community aspect of the phenomena. However, not only that our computational approaches are radically different, Zannettou et al. are mostly concerned with quantifying specific antisemitic trends in Gab and 4chan/pol, while our focus is the disambiguation and user classification, and the analysis of the memetics of the echo from a linguistic perspective.

Our work differs from each of the works mentioned above in at least two of the following fundamental aspects: (i) we aim at detecting hate-mongers and self organized hate communities, not only hateful posts, (ii) we harness both language (beyond keywords) and network structure in a multimodal neural network, (iii) our dataset is radically different and significantly bigger than other datasets of hate and abuse in mainstream platforms like Twitter, and (iv) our dataset was collected in an organic way by bootstrapping the ambiguous and elusive echo symbol. As such, it contains tweets posted by a diverse set of users, many of whom are not hate mongers, although they may use similar linguistic forms.

Data

Echo Corpus

A large dataset of over 18,000,000 English tweets posted by ∼7K echo users was constructed in the following manner[3]:

1. **Base Corpus** We have obtained access to a random sample of 10% of all public tweets posted in May and June 2016 – the peak use of the echo.
2. **Raw Echo Corpus** Searching the base corpus, we extracted all tweets containing the echo symbol, resulting in 803,539 tweets posted by 418,624 users. Filtering out non-English Tweets and users who used the echo less than three times we were left with ∼7K users[4].
3. **Echo Corpus** We used Twitter API to obtain the most recent tweets (up to 3.2K) of each of the users remaining in the English list[5]. This process resulted in ∼18M tweets posted by 7,073 users. Some of the accounts we found using the echo were already suspended or deleted at the time of collection, thus their tweets were not retrievable.

To the best of our knowledge, this is the first time this dataset is being analyzed computationally and on a large scale.

Data Annotation

We sampled a thousand users from the dataset, inspected their use of the echo, and manually assigned each user one of three labels: HM (Hate Monger), R (Response) for users discussing the hate symbol, and N (Neutral) for users using the symbol in non-hate contexts. Examples of tweets from each category are presented in Table 1, and descriptive statistics of the users of the different categories are presented under GOLD USERS in Table 2.

Network Statistics

Hate does not propagate in a void. Reconstructing the network of echo users enables us to identify structures, roles and interfaces that facilitate the propagation of hate-speech.

[3]Twitter Search API ignores special characters, thus querying for the echo was not feasible.

[4]The echo is found in tweets written in multiple languages, particularly in East-Asian languages of which the user based is known for heavy use of ascii art and kaomoji (McCulloch 2019).

[5]The data was collected in December 2016, amidst reports on the trending 'echo'.

	GOLD USERS				PREDICTED	
Label	HM	R	N	R+N	HM	R+N
Total #Users	170	55	775	830	1136	5927
Total #Tweets	339K	141K	2M	2.15M	2.26M	15.44M
Avg. #Days Active	999±783	1910±973	1558±853	1582±866	1080±894	1511±876
Avg. Tweets/day	11±19	7±9	19±37	18±36	15±31	32±96
Avg. #Friends	674±1445	741±1103	972±2136	957±2084	783±1527	1224±5527
Avg. #Followers	1022±2619	1067±2070	1941±5991	1884±5817	3848±60925	4432±85490
Avg. %Replies	37±24	27±21	27±23	27±23	41±26	26±23
Avg. %Retweets	34±24	26±24	24±22	24±22	31±25	24±22
Avg. %URL	73±21	58±28	57±26	57±26	74±21	56±26
Avg. %Hashtags	16±14	22±23	20±23	20±23	18±16	19±24

Table 2: Account statistics derived from the annotated data (left) and predicted classes (right). Standard deviation is marked with ±. Average days accounts are active, tweets per day, friends and followers are based on available account meta data. Average replies, retweets, URLs and hashtags ratios are based on tweets in the Echo corpus.

Assuming that different types of engagement reflect different types of relations, we consider three different network semantics: mention-based, reply-based and retweet-based. In order to reduce noise we consider an edge only if its weight is higher than some threshold $\delta \geq 3$. The mention-based[6] network presented in Figure 1 contains 3977 singletons (not presented in the figure), 2226 connected components (269 weak, 1993 strong), and a total of 3,092 nodes and 12,622 edges. Figures 2a and 2b present only the nodes annotated as part of the gold standard, each node is colored by its label. The tendency of hate users to form tight communities is evident by the dominant cluster of red nodes that form the largest connected component (LCC). A detailed comparison between network statistics of the full network and the LCC can be found in Table 3 (top two rows).

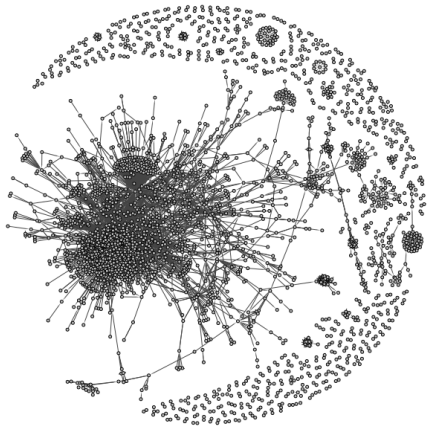


Figure 1: Mentions network of echo users. Layout: force-directed. Minimum edge weight: 3.

[6]Retweet and reply network are significantly sparser, but exhibit a similar structure in terms of communities. In the remainder of the paper we report results based on this network.

Unsupervised Detection of Racist Users

While the analysis of the properties of the social network may shed light on the emergence of racist communities, some patterns may be missed due to data sparsity and the constraints imposed on data collection. We therefore opt for unsupervised content-based methods in order to discover disconnected individuals and clusters of like-minded racists.

Experimental Setting In order to achieve an abstract representation of topics and semantics we represented the text in two ways: word embeddings (Bojanowski et al. 2017) and topic models (Wallach, Mimno, and McCallum 2009). All tweets by a user were concatenated to one long text and users were represented in three different ways: (i) an average of their embeddings (EMBD), (ii) their single most salient topic (TM^S), and (iii) the full topic distribution vector for each user (TM^F). Clustering is done with the classic k-mean algorithm, assuming two settings: (i) three clusters, corresponding to the HM, R and N classes and, (ii) two clusters, collapsing the R and N classes to a single class of ¬HM.

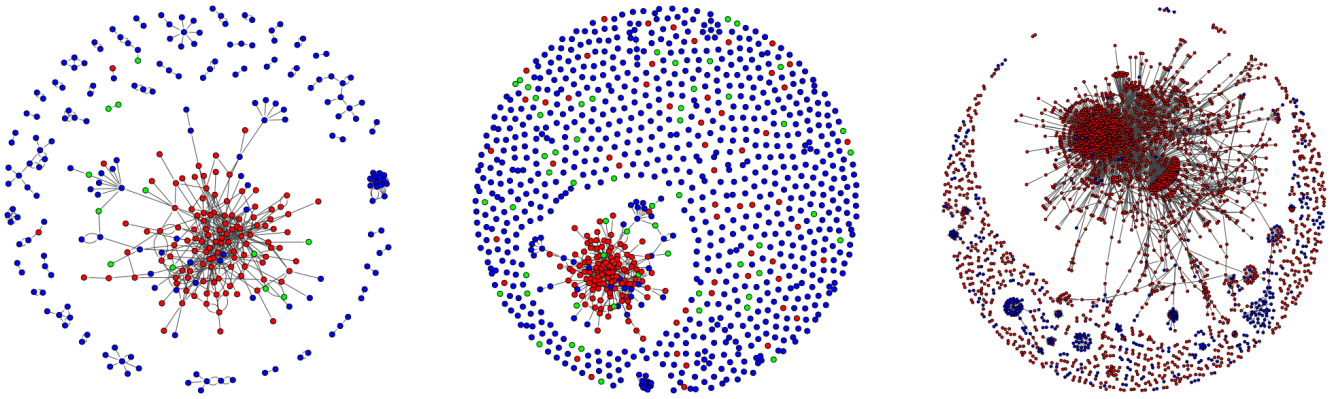
Clustering Results The Rand Index (Rand 1971) is used to evaluate cluster quality against the gold standard set. All methods and settings achieved decent clustering results (see Table 4). Best results were obtained using the full distribution of topics ($k = 30$). Figure 2c presents the user cluster assignments (color) in the full network (singletons removed). Both the Rand Index (RI) results and the graphic visualization suggest a strong correlation between the network structure and the language used. These results are in line with previous studies of hate-speech in other platforms such as Gab and 4chan (Ribeiro et al. 2017; Zannettou et al. 2020).

Multi-Modal Neural Architecture

Given that hate speech does not propagate in a void (see the previous section and related work), we propose a multi-modal neural architecture that takes into account the text of a user, as well as the texts of other users in her ego network. The main motivation for this approach is that multiple weak

Graph	#Nodes	#Edges	Density	Diameter	#Triangles	Max #triangles	#Strong CC	#Weak CC
Full	3092	12622	0.0013	20	24261	1988	1993	269
LCC	553	5215	0.0171	19	11114	1352	n/a	n/a
HM	730	5783	0.0109	11	11188	1728	387	34
R+N	2362	5018	0.0009	24	8918	669	1710	362

Table 3: Network (without singletons) features computed on the full mention network of echo users, and on its largest connected component (LCC), HM, and R+N subnetworks. We report on the following features for each network: Number of nodes, number of edges, density (computed without loops), diameter (within connected components), number of triangles, maximum number of triangles for a single node, and number of strongly and weakly connected components. Minimum edge weight: 3.



(a) Gold labels (no singletons) (b) Gold labels (with singletons) (c) Predicted labels (no singletons, R+N)

Figure 2: Mention based networks of echo users. Nodes colored by label – HM in red, R in green, and N in blue. Figures 2a and 2b contain only the annotated subset of nodes. Figure 2c is the same network presented in Figure 1, the R and N classes are collapsed.

MODEL	RI^2	RI^3
EMBD	0.687	0.687
TM^S	0.801	0.672
TM^F	**0.811**	**0.743**

Table 4: Rand Index of clusters vs. Gold Standard set. Clusters are computed based on user's textual embeddings (EMBD), most salient topic representation (TM^S), and full topic distribution (TM^F). RI computed for two (RI^2) and three (RI^3) class/cluster settings. Embedding dimension: 300. K: 30.

signals from user u's "neighborhood" can be used to fine-tune the signal produced by user u herself. This approach is common in sociology and demographic polling (Johnston 1974; Latané 1981; Sampson 1988) and we expect that it will be especially beneficial in the cases in which obscure, vague or ambiguous language is used.

Post-Level Module (PLM) The basic unit for classification is a single post (tweet). We fine-tune a BERT transformer (Devlin et al. 2018) on the annotated dataset. Fine tuning is done after adding a bi-directional-LSTM with global max pooling, a dense, and a dropout layers. The architecture of the post level module is illustrated in the orange box in the center of Figure 3.

User Network Module The post level module is used to process three distinct streams of tweets: (i) tweets of the user we wish to classify (user u), (ii) tweets of the users following u, and (iii) tweets of the users u is following. The full architecture is illustrated in Figure 3. In this work, a user v that mentioned $u \geq \delta = 3$ times is considered to be a follower of u, however, directed relations can be defined by other engagement patterns. The outputs of each of the three streams are processed slightly differently. While the results of the PLM of the user in question (u) are concatenated, the PLM results of each of her followers and followees are averaged on the user level (separately), thus each follower or followee contributes a single value to the final vector. All the PLM outputs are concatenated to a single vector that is composed of the all PLM predictions for tweets of u (blue vector), a concatenation of all averaged scores for each of the followers and followees (yellow and green vectors, respectively). This vector is further concatenated with a vector of network features (red vector) of user u, e.g., in-degree, out-degree, betweeness, number of triangles u is part of etc. The concatenated vector could be fed to any classification model. We experimented with a three-layer FFNN, Gradient Boosted Machine (GBM) algorithms, and Logistic Regression.

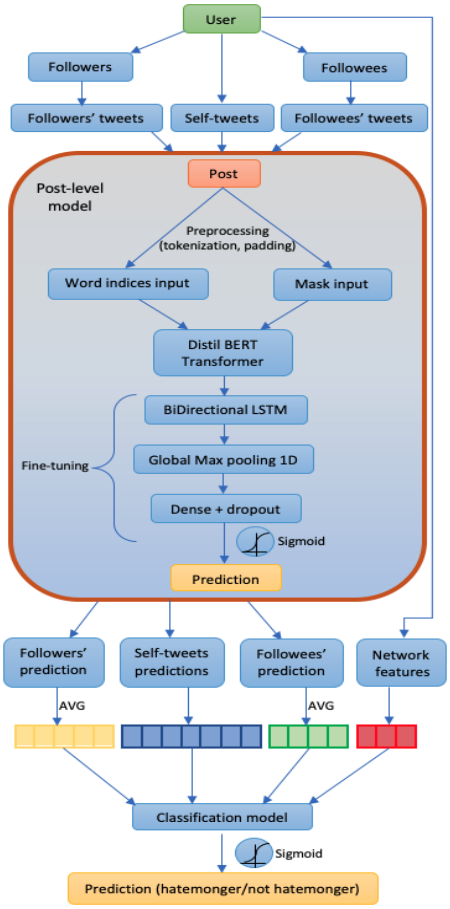


Figure 3: Multi-Modal neural architecture. The neural model process texts in different streams and treatments based on the structure of the social network.

MODEL	Precision	Recall	F1	AUC
U	0.692	0.607	0.613	0.925
U+$\overrightarrow{FU}$	0.659	0.607	0.633	0.935
U+$\overrightarrow{UF}$+$\overrightarrow{FU}$+N	0.873	0.666	0.755	0.959
U+N	0.822	**0.725**	0.77	0.958
U+$\overrightarrow{UF}$	0.898	0.686	0.777	0.955
U+$\overrightarrow{UF}$+N	**0.923**	0.705	**0.8**	0.959

Table 5: Ablation results achieved by the multi-modal neural network. U: tweets of a single user u; $\overrightarrow{UF}$: tweets of users followed by u; $\overrightarrow{FU}$: tweets of the followers of u; N: u's network features.

Multi-Modal Neural Results Using our proposed multi-modal architecture we achieve an F-score of 0.8. Table 5 presents ablation results achieved using different components of the multi-modal architecture. Best results were achieved in the setting that includes the user's tweets, the tweets of the users she follow and the network features (U+$\overrightarrow{UF}$+N). It is worth noting that both network structure features and texts of followees provide a significant improvement over the baseline. These results are inline with our hypothesis that hate-mongers operate as a group and with the network coloring reported in the previous section.

Interestingly, accounting for tweets posted by the followers of u harms the classification. This result suggests an interesting dynamic between hate-mongers and mainstream users – while extremists tend to mostly engage with other extremists, some mainstream users refer and engage with extremists. The reasons for this "assymetry" are beyond the scope of this work, however we hypothesize that some mainstream users are referring to and engaging with extremists either out of curiosity or in an effort to point to the phenomena. This hypothesis will be tested in future work.

Analysis and Discussion

The result reported in Sections and demonstrate the strong connection between the network topology and the language used in different subcommunities. Moreover, using a multi-modal neural architecture we demonstrated that processing texts, while taking the network structure into account improves results significantly, especially in the case of vague or ambiguous language. In the remainder of this paper we further discuss various aspects related to the use of language, the network structure and the activity of hate groups.

Network Structure and Predicted Labels

Clustering results were reported in Section . While node colors in Figure 2c are decided by cluster assignment, similar results are obtained when node colors are decided by the multi-modal neural architecture (Section). Most singletons are neutral users, in line with the trend presented in Figure 2b. Hate-mongers, on the other hand, make the bulk of the large component and more likely be part of a connected component. This tendency is striking as hate-mongers tend to have significantly less friends and followers (see Table 2). The discrepancy between their connectedness in the echo-induced network and their global degree suggests that the echo is more infectious as a hate-symbol/meme than in its other senses (a hug, broadcasting, etc.) – highlighting the communal aspect in the adoption of hate. This is in line with previous work reporting that radical content travels faster and further in the network (Mathew et al. 2019). These observations also provide a different angle on the notion of the 'lone wolf' discussed in (Ribeiro et al. 2017) – on the one hand, hate mongers are highly active and organized, while on the other hand, their in and out degrees are significantly smaller than those of mainstream users. It is interesting to see that some responders and neutral users are also at the core of the large components. We manually examined some of them, observing that responders often attract response from the hate-mongers. A typical exchange is presented in Figure 4.

Hate Leaders

Using network centrality measures, we can find leaders and promoters of racism and hate. Using the three network semantics (reply, mention and retweet) and seven centrality measures (in/out/total degree, betweenness, eigenvector,

Figure 4: A hate-monger responds to Jefferey Goldberg, Editor-in-Chief of The Atlantic, explaining the meaning of the echo. Note the Nazi salute used as the profile picture of the HM, and the user name – a reference to the antisemitic conspiracy trope of the Protocols of the Elders of Zion.

User name	Suspended	Predicted	Manual Label
ThaRightStuff	✓	HM	✓
ramzpaul	✗	HM	✓
PaulTown	✓	HM	✓
Third_Position	✗	HM	✓
DrDavidDuke	✗	HM	✓
SeventhSonTRS	✓	HM	✓
TrumpHat	✓	HM	✓
TheeCurrentYear	✓	HM	✓

Table 6: Accounts with the highest generalized centrality. The not check mark (✗) in the suspended column refers to accounts that are temporarily suspended (at the time of the query) as it violated the Twitter Media Policy.

closeness and page-rank) we rank users by the number of times they appear in the k-core of most central users. We call this ranking *generalized centrality*. A number of users with the highest generalized centrality are listed in Table 6, some of which are known leaders of white supremacy movements, e.g., David Duke (@*DrDavidDuke*), the former Grand Wizard of the KKK, and Mike Enoch (@*TheRightStuff*), founder of The Right Stuff media network and The Daily Shoah podcast. The centrality of these notorious figures serves as a sanity check for the generalized centrality measure. To further verify the role of the central nodes we observe their label as predicted by the neural network model. As can be seen in Table 6 all of the users presented are predicted as HM. Moreover, as a proxy, we queried Twitter for their account, finding that all of them were suspended – an indication for high profile, malicious, often racist, activity. No N or R users were found to have high generalized centrality rank.

The absence of other known leaders of the alt-right, e.g., Mike Cernovich and Richard Spencer, is somewhat surprising. We attribute it to the constraints imposed in the curation of the raw echo corpus (see Section). These users may have not used the echo (in the 10% sample) during the two month span during which the base corpus was curated. However, their centrality to the network is evident by the large number of times they are mentioned or being retweeted by the nodes flagged as hate-mongers, compared to the minimal trace they leave in the users of the R and N groups. Spencer, for example, is mentioned 5565 times, retweeted 1716 times, and replied to 1611 times by the HM group. These numbers are comparable to the mention/retweet/reply counts of the users with the top generalized centrality. These findings suggest that we managed to accurately reconstruct the network of hate mongers, in spite of the limitations and constraints imposed by Twitter API and other access issues.

Linguistic Variations: Orthography and Semantics

We observe variations in the orthography and the semantics of the echo symbol. These variations are typically the result of the canonization of a word or a term within a certain speaker community, hence providing another perspective on adaptation of linguistic forms by a wider community.

Abstraction and Semantic Drift Starting as an abstract symbol, the echo was used to mark concrete named entities – *people* of Jewish heritage. It further evolved to mark *abstract entities* such as `(((bankers)))` and `(((globalists)))`, echoing ancient antisemitic tropes. The use of the echo to mark abstract *concepts* such as `(((narrative)))` or suggestive pronouns like `(((they)))` and `(((who)))` reflects another stage in the semantic evolution of the symbol. Finally, anecdotal evidence demonstrate that the antisemitic symbol is being repurposed to target other minority groups, e.g., `(((illegal mexicans)))`, (Tuters and Hagen 2019).

Expressive Lengthening Expressive lengthening, common in online informal writing, is the habit of adding characters to words in order to enhance the message or the sentiment conveyed in it. Typical examples are 'aaaaaaaaargh', 'lollll', and 'sweeeeet!!!!' (McCulloch 2019). We observe expressive lengthening of the echo as hate-mongers try to underscore their hate, e.g., `((((((bankers))))))` and `(((((jooooooos[7])))))`. Another interesting orthographic phenomena is the use of a reverse echo to declare an 'Aryan' affiliation, e.g., users declaring themselves `))))goyim_godess((((` and `)))anti-semitic(((`.

The Intersectionality of Hate

When used as a hate-symbol, the echo is mostly used in an antisemitic context (see previous subsection for exceptions). Although one would expect a dataset constructed around the echo to contain mostly antisemitic hate speech, we do observe the "intersectionality of hate" which allows us to explore the attitude of hate groups toward other minorities and protected groups.

[7] A derogatory term for Jews, used for its (expressively-lengthened) homophony.

Looking at the words and hashtags used most frequently by each of the groups, we observe a general racial pattern, going well beyond the antisemitic use of the echo. Users flagged as hate-mongers by our algorithm, are twenty times more likely to use the term Zionist as a general slur; talk about whiteness and white genocide; use derogatory terms like kike[8], cuck[9], and skittle[10]. In addition, these users are more likely to refer to Arabs, Muslims and immigrants in more explicit derogatory ways. For example, Muslims are addressed as muzzies[11] and the hashtag `#rapefugees` is used to depict refugees as rapists.

The following tweet, posted by an HM account provides an illuminating example for the "intersectionality" of hate: `Poland refuses #rapefugees and is now on the verge of civil war. (((Who))) could be behind this? #WhiteGenocide`. Notice the dual strand of hate in the tweet: labeling Muslim refugees as rapists, along with the abstract use of the echo to hint that the influx of the "rapefugees" is a Jewish plot to destabilize western countries as part of a war on the "white race".

Comparing the popular hashtags among the HM and N groups we find that the HM group is trending with #pizzagate, #minorityPrivilige, #WhiteGenocide, #altright, #tcot[12], #AmericaFirst, #GamerGate, #FeelTheBern, #MAGA #Brexit and #rapefugees, while the N echo users tend to use the hashtags #job, #sex, #LIVE, #broadcasting, #party and #NowPlaying, all associated with other meanings of the echo symbol – a visual resemblance of an engulfing hug or a radio tower.

It is interesting to note that the R users seem to exhibit a stronger interest in politics, compared to their N counterparts. The most frequently used hashtags of the R group are: #localbuzz, #Facebook, #SocialMedia, #antisemitism, #DemDebate, #VPDebate, #Israel, #LonelyConservative, and #NeverTrump, as well as #MAGA and #Trump2016. We attribute this tendency to an inherent selection bias – the responders are those who care more about political agenda and therefore try to have a better grasp of its extreme fringes. These findings also support the split of the non hate-mongers users to two different clusters with unique features instead of a single large cluster that combines both groups.

We wish to stress that while the HM users of the echo tend to enthusiastically support a separatist right-wing agenda, not all conservative users or supporters of Trump or of the Brexit are hate-mongers. We also wish to point out that the perceived nicety of the R+N users, demonstrated by the heavy use of positive words is somewhat misleading. This may be a side-effect caused by the manner in which the corpus was constructed, since the meanings and the contexts in which the echo symbol is used are polarized. While the vast majority of the tweets in the corpus do not contain the echo at all, all users in the data did use this unique symbol, often as a very strong sentiment/stance marker.

[8]Ethnic slur for Jews.

[9]A weak and submissive person. Similar to the classic 'pussy'. Often used to describe minorities and "intellectuals".

[10]Originally a small fruit-flavoured candy, repurposed as a derogatory term.

[11]A religious slur referring to Muslims.

[12]A reference to the "Top Conservatives On Twitter".

	GOLD USRS		ALL USRS	
Label	HM	R+N	HM	R+N
#Users	170	830	1136	5937
#Users mentioning IRA	88	67	623	379
#User retweeting IRA	81	53	529	312
#Unique IRA mentioned	24	25	63	76
#Unique IRA retweeted	19	20	45	46
#Total IRA mentions	196	102	1375	595
#Total IRA retweets	167	79	1088	479

Table 7: Engagement of echo users with IRA accounts. #Users mentioning/retweeting IRA: the number of echo users mentioning/retweeting an IRA user. #Unique IRA mentioned/retweeted: the number of unique IRA accounts mentioned by echo users. #Total IRA mentions/retweets: the total number of mentions/retweets of IRA users by echo users.

Links to Russian Trolls

Recent studies suggest that foreign activity on social media was strategically used in an attempt to further radicalize groups that already have an inclination to extremism (Jamieson 2018; Addawood et al. 2019). We conclude this paper with a brief examination of foreign involvement with alt-right communities.

The Internet Research Agency (IRA) is a Russian troll-farm linked to the Russian intelligence, according to a declassified report by the United States Office of the Director of National Intelligence (2017), and the Special Counsel report on the Investigation into Russian Interference (Mueller 2019). A list of 3,814 account handles, linked to the IRA was identified and released by Twitter. In the ten-week period preceding the 2016 election these accounts posted 175,993 Tweets, approximately 8.4% of which were election-related (Twitter 2018). None of the IRA trolls used the echo in the %10 sample of two months covered in the base corpus. However, we find their impressions in the network. Analysis of the data reveals that hate-mongers (HM) are eight to nine times more likely to mention or retweet an IRA user than their R+N counterparts. Looking only at users that actively engage with IRA accounts, a hate-monger engages with an IRA account in a higher rate, see Table 7 for more details. While a detailed analysis of these efforts are beyond the scope of this paper, our computational results support the qualitative analysis of foreign meddling in local politics (Jamieson 2018).

Conclusion

Antisemitism is only one manifestation of racism. Using a large and unique corpus constructed around an ambiguous antisemitic meme, we showed how networks of hate-mongers can be reconstructed. Analyzing content and the network structure in tandem provides significant insights on the promotion of hate, beyond antisemitism, the central figures dominating the network, the engagement between hate-

mongers and other users and the utilization of this network for international political warfare. Future work includes a temporal analysis of the formation of the network as well as a finer analysis of the types of hate promoted by the network.

References

Addawood, A.; Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Linguistic Cues to Deception: Identifying Political Trolls on Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 15–25.

ADL. 2020. Antisemitic Incidents Hit All-Time High in 2019. https://www.adl.org.il/en/news/antisemitic-incidents-hit-all-time-high-in-2019/. Accessed: 2020-08-12.

Akbarzadeh, S. 2016. The Muslim Question in Australia: Islamophobia and Muslim Alienation. *Journal of Muslim Minority Affairs* 36(3): 323–333.

Bezio, K. M. 2018. Ctrl-Alt-Del: GamerGate as a precursor to the rise of the alt-right. *Leadership* 14(5): 556–566.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135–146.

Chen, H.; McKeever, S.; and Delany, S. J. 2019. The Use of Deep Learning Distributed Representations in the Identification of Abusive Text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 125–133.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Del Vigna12, F.; Cimino23, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 86–95.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dodd, V.; and Marsh, S. 2017. Anti-Muslim hate crimes increase fivefold since London Bridge attacks. *The Guardian* 7.

Edwards, G. S.; and Rushin, S. 2018. The effect of President Trump's election on hate crimes. *Available at SSRN 3102652* .

Entorf, H.; and Lange, M. 2019. Refugees welcome? Understanding the regional heterogeneity of anti-foreigner hate crimes in Germany. *Understanding the Regional Heterogeneity of Anti-Foreigner Hate Crimes in Germany (January 30, 2019). ZEW-Centre for European Economic Research Discussion Paper* (19-005).

Gambäck, B.; and Sikdar, U. K. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, 85–90.

Gao, L.; Kuppersmith, A.; and Huang, R. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 774–782. Taipei, Taiwan: Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1078.

Grover, T.; and Mark, G. 2019. Detecting Potential Warning Behaviors of Ideological Radicalization in an Alt-Right Subreddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 193–204.

Hankes, K.; and Amend, A. 2018. The Alt-Right is Killing People. https://www.splcenter.org/20180205/alt-right-killing-people. Accessed: 2019-08-13.

Hawley, G. 2017. *Making sense of the alt-right*. Columbia University Press.

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.

Iwama, J. A. 2018. Understanding hate crimes against immigrants: C onsiderations for future research. *Sociology compass* 12(3): e12565.

Jamieson, K. H. 2018. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press.

Johnston, R. J. 1974. Local effects in voting at a local election. *Annals of the Association of American Geographers* 64(3): 418–429.

Latané, B. 1981. The psychology of social impact. *American psychologist* 36(4): 343.

Laub, Z. 2019. Hate Speech on Social Media: Global Comparisons. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons. Accessed: 2019-08-15.

Lima, L.; Reis, J. C.; Melo, P.; Murai, F.; Araujo, L.; Vikatos, P.; and Benevenuto, F. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 515–522. IEEE.

Magu, R.; Joshi, K.; and Luo, J. 2017. Detecting the hate code on social media. In *Eleventh International AAAI Conference on Web and Social Media*.

Malevich, S.; and Robertson, T. 2019. Violence begetting violence: An examination of extremist content on deep Web social networks. https://firstmonday.org/ojs/index.php/fm/article/download/10421/9403. Accessed: 2020-07-13.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, 173–182.

McCulloch, G. 2019. *Because Internet: Understanding the new rules of language*. Riverhead Books.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, 928–940. Springer.

Mueller, R. S. 2019. Report on the investigation into Russian interference in the 2016 presidential election. *US Dept. of Justice. Washington, DC* .

Munn, L. 2019. Alt-right pipeline: Individual journeys to extremism online. https://firstmonday.org/ojs/index.php/fm/article/download/10108/7920. Accessed: 2020-08-13.

Nagle, A. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, 145–153.

of the Director of National Intelligence, O. 2017. Assessing Russian activities and intentions in recent US elections. *Unclassified Version* .

Osman, M. N. B. M. 2017. Retraction: Understanding Islamophobia in Asia: The Cases of Myanmar and Malaysia. *Islamophobia Studies Journal* 4(1): 17–36.

Park, J. H.; and Fung, P. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206* .

Perry, B.; Akca, D.; Karakus, F.; Bastug, M. F.; et al. 2020. Planting Hate Speech to Harvest Hatred: How Does Political Hate Speech Fuel Hate Crimes in Turkey? *International Journal for Crime, Justice and Social Democracy* 9(2).

Phillips, W. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66(336): 846–850.

Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A.; and Meira Jr, W. 2017. "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. *arXiv preprint arXiv:1801.00317* .

Salminen, J.; Almerekhi, H.; Milenkovic, M.; Jung, S.-g.; An, J.; Kwak, H.; and Jansen, B. J. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *ICWSM*, 330–339.

Salminen, J.; Hopf, M.; Chowdhury, S. A.; Jung, S.-g.; Almerekhi, H.; and Jansen, B. J. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10(1): 1.

Samghabadi, N. S.; Patwa, P.; Srinivas, P.; Mukherjee, P.; Das, A.; and Solorio, T. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 126–131.

Sampson, R. J. 1988. Local friendship ties and community attachment in mass society: A multilevel systemic model. *American sociological review* 766–779.

Shaheed, A. 2019. Elimination of all forms of religious intolerance. Technical Report A/74/358, The United Nations, Secretary General.

Sunar, L. 2017. The long history of Islam as a collective "other" of the west and the rise of Islamophobia in the US after Trump. *Insight Turkey* 19(3): 35–52.

Thomas, E. 2019. ASPI explains: 8chan. https://www.aspistrategist.org.au/aspi-explains-8chan/. Accessed: 2019-08-13.

Tuters, M.; and Hagen, S. 2019. (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society* 1461444819888746.

Twitter. 2018. Update on Twitter's review of the 2016 US election. *Twitter Public Policy Blog. Retrieved January* 5: 2020. URL https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html.

Wallach, H. M.; Mimno, D.; and McCallum, A. 2009. Rethinking LDA: Why Priors Matter. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, 1973–1981. USA: Curran Associates Inc. ISBN 978-1-61567-911-9. URL http://dl.acm.org/citation.cfm?id=2984093.2984314.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wullach, T.; Adler, A.; and Minkov, E. 2020. Towards Hate Speech Detection at Large via Deep Generative Modeling. *arXiv preprint arXiv:2005.06370* .

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringini, G.; and Blackburn, J. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, 1007–1014. International World Wide Web Conferences Steering Committee.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, 745–760. Springer.