Dynamic Topic Modeling Reveals Variations in Online Hate Narratives

Richard Sear[1(✉)], Nicholas Johnson Restrepo[2], Yonatan Lupu[1], and Neil F. Johnson[1]

[1] The Dynamic Online Networks Lab, The George Washington University, Washington, DC 20052, USA
{searri,neiljohnson}@gwu.edu
[2] ClustrX, LLC, Washington, DC 20007, USA

Abstract. Online hate speech can precipitate and also follow real-world violence, such as the U.S. Capitol attack on January 6, 2021. However, the current volume of content and the wide variety of extremist narratives raise major challenges for social media companies in terms of tracking and mitigating the activity of hate groups and broader extremist movements. This is further complicated by the fact that hate groups and extremists can leverage multiple platforms in tandem in order to adapt and circumvent content moderation within any given platform (e.g. Facebook). We show how the computational approach of dynamic Latent Dirichlet Allocation (LDA) may be applied to analyze similarities and differences between online content that is shared across social media platforms by extremist communities, including Facebook, Gab, Telegram, and VK between January and April 2021. We also discuss characteristics revealed by unsupervised machine learning about how hate groups leverage sites to organize, recruit, and coordinate within and across such online platforms.

Keywords: Latent Dirichlet Allocation · Online hate · Hate speech · Machine learning · Topic modeling

1 Introduction

Online hate speech is a very worrying societal problem that is attracting significant attention not only among academics, but also among policy makers because of its highly negative impact on victims [1–5]. Arguments continue to rage around the trade-off between the need to moderate such content and to regulate or punish social media companies that do not comply, versus the need to protect online users' free speech. The presence of hate speech raises a plethora of issues, perhaps most importantly that it can precipitate offline acts of violence. Better-moderated social media platforms such as Facebook have been stuck in a fight against the spread and proliferation of hate speech for years, with efforts increasing in early 2021 after the riot at the U.S. Capitol on January 6. Despite efforts to curtail it, hate speech continues to be a problem. Its resilience is partly a result of the adaptive, multi-platform network that carries hate speech throughout the internet between both moderated and unmoderated platforms. A better understanding of

K. Arai (Ed.): SAI 2022, LNNS 507, pp. 564–578, 2022.
https://doi.org/10.1007/978-3-031-10464-0_38

this network, and the narratives it carries, is important for academics and policymakers looking to gain a better picture of the online battlefield across which online hate evolves. In particular, an automated approach could help social media companies to better police their own platforms in light of the sheer volume of new content that appears on each one daily. Indeed, Facebook already uses artificial intelligence to help it with moderation [6]. In short, online social media platforms with built-in community features are known to be popular fora in which producers of hate speech congregate. Unfortunately, social media companies have an uphill battle containing it due to the enormous amount of fresh material combined with its frequent virality across multiple social networks.

Our study here is prompted by the following questions: (A) how are different social media platforms used to spread hate narratives? (B) Can an automated technique be developed in order to overcome the practical problem that human moderators cannot sift through such enormous amounts of content quickly enough every day across multiple platforms? The procedure used in this study is by no means a complete solution to these issues, but it provides a useful framework to be built upon in future work. We show here that a machine learning model like Latent Dirichlet Allocation (LDA) can provide useful insights into the ways that hate groups utilize different social media platforms for different purposes, both in the spreading of narratives and their attempts to coordinate and organize.

Our study does not use Twitter data since Twitter tends to be used more as a "broadcast" medium, whereas narratives tend to be nurtured on platforms that have community spaces specifically built around fostering discussion (e.g. Facebook's "Page"). Twitter is in the early stages of developing such community spaces, but has not yet made the feature widely available [7]. In the present methodology, we obtain the material from community content that is publicly available on Facebook Pages, VKontakte, Telegram, and Gab. All pages or groups used in this study were categorized by our team of subject matter experts (SMEs) as being "hateful" according to well-established criteria as discussed in Sect. 2. We stress that our analysis does not require individual or personal information to gain useful insights, similarly to how understanding conversations in a crowded environment does not need information about the individual people who make up the crowd. Details of our approach are provided in Sect. 2 of this paper. Though our study would benefit from further improvement and refinement, it represents one of the first attempts at a highly automated yet transparent model of hate speech analysis across multiple social media platforms.

2 Data and Machine Learning Methods

We start by briefly describing the online ecosystem in which hate manages to thrive, and how the online audience aggregates itself within this ecosystem. The global social media universe comprises several billion users who operate within and often across multiple social media platforms. Most of these platforms have an in-built community feature that allows online users to aggregate around a topic of interest. Each platform uses their own term to describe such online communities; for example, Telegram uses "Group" or "Channel" whereas Facebook uses "Page" or "Group" [8]. Typically, these communities feature relatively benign narratives around sports or lifestyle choices, but

some generate or focus on more extreme content which can be regarded as 'hateful'. This subset of hateful communities and their narratives can survive a long time if the platform has lower levels of moderation. Any such in-built community can feature links (hyperlinks such as URLs) into other communities whose content is of interest to them, within the same social media platform and between different ones. This can help these communities keep their members away from moderator pressure [4, 5, p. 202].

Hence, the online ecosystem and its audience comprise a highly complex network of interconnected communities within and across platforms, through which hate narratives can evolve and move. Between 2019 and 2021, Facebook developed new content moderation policies designed to counter violent extremism and reduce hate speech [9, 10]. By contrast, Gab and Telegram have largely grown their user-base by positioning themselves as unmoderated (or less moderated) free-speech alternatives to major platforms like Facebook and Twitter [11, 12]. VKontakte (VK) is a social network with many similar features to Facebook, but based and hosted in Russia. While VK is subject to more content moderation policies than unmoderated alternatives like Gab, past research has shown that American and European white nationalists have 'migrated' to there after being banned from Facebook and Discord [13].

In this study we look across multiple social networks to capture and measure the publicly available text in posts that were shared in hateful communities. To label a community as 'hateful', two SMEs who focus on right-wing extremism manually reviewed each community's most recent 25 posts. When the opinions of the SMEs coincided that two of these posts exhibited hate against protected groups referenced in the "hate crimes" description from the FBI, then we labeled that community as hateful and we included it in this study [14]. Reviewers also drew on the text of Mann's discussion of violence that is ethnically and racially motivated as "cleansing nation-statism through paramilitarism" [15]. As a result of these definitions, the hateful communities included in this study include organized, well-known real-world hate groups like the KKK, as well as decentralized movements like certain Boogaloo groups. Our study uses only English text as identified by Google's Compact Language Detector. However, our methodology and implementation can easily be extended to other languages. Our collection of hateful communities was carried out irrespective of their geographical location. All posts used in the study were created between January 1 and April 30, 2021 (inclusive).

We perform standard preprocessing on the text to remove emojis, URLs, and stopwords. Notably, we leave in domains by converting them to recognizable tokens with the following procedure: "domain.com" becomes "domain__com". Within the LDA model, such a token will be treated like any other word. We do this so that if there is a useful signal related to social media posts' domain usage, it remains visible upon manual inspection of the output topics for LDA. Additionally, during this preprocessing phase, we unroll contractions (i.e. "don't" becomes "do not") and lemmatize and stem words using the Natural Language Toolkit.[1] The goal of this preprocessing is to reduce the "noise" present in the text; generic articles and commonly-used words are not good indicators of topic, and therefore the LDA models will achieve a better fit without them. We base this off a similar preprocessing setup in previously-published work [16].

[1] https://www.nltk.org/.

We processed the text content by aggregating it for each platform (Telegram, Facebook, Gab, and VK). We then analyzed it using the machine learning tool LDA, which is an unsupervised learning algorithm [17]. This algorithm detects the emergence and evolution of topics by regarding documents as distributions of topics and topics as distributions of words. It learns how to fit these distributions to the dataset during training. We then employ a dynamical version of LDA, which also accounts for the timestamp when the post was created, to extract the evolution of the emergent topics over time [18]. We employ the Gensim implementation for both standard and dynamic LDA.[2] This is a completely unsupervised process: all we need to input is the "number of topics" (*n_topics*), which is a parameter that designates how many groups the model should cluster text into.

Having carried out this process, we then use C_V coherence as an evaluation technique (see [19] for details). There are many types of coherence score which provide a quantitative method for measuring the alignment of words within an identified topic and can be used as a "goodness of fit" measurement for the topic modeling process. This C_V coherence quantity is generated by a separate algorithm that analyzes the set of topics (coherence *is not* specific to LDA). Coherence analyzes the entire vocabulary of words in a corpus, ranked according to the word distribution in each topic. The C_V coherence score for a single model is obtained by calculating the arithmetic mean of the scores obtained for each topic. Specifically, C_V is calculated using a sliding window, one-set segmentation of the top words. This comprises collections of probability measures on how often top words in topics co-occur with each other in the corpus. The C_V formula also incorporates cosine similarity as an indirect confirmation measure for the context vectors generated by the one-set segmentation. A full description and explanation of C_V is given elsewhere [19]. Our manual review of the top words in each topic's word distribution reveals that they do indeed relate to separate conversation topics.

Sophisticated automation could be used to address the problem of troublesome content in the sea of new material appearing every day on social media platforms; specifically, the combination that we present here of both standard and dynamic LDA approaches. Many standard LDA models can be trained and then their topic keywords quantified using C_V to determine the best number of topics discussed in particular platforms (i.e. the highest value of C_V). We can then seed this parameter into a dynamic LDA model that over a longer time period can automatically track the evolution of topics in terms of their highest-probability keywords. In the next section, we illustrate the output of this method [16], where we train multiple standard LDA models and then average their coherence scores to determine an optimal fit. All code used in these experiments is open-sourced and documented at the following repository: https://github.com/gwdonlab/topic-modeling. It can be used to run similar experiments on arbitrary text datasets.

[2] https://radimrehurek.com/gensim/.

3 Results and Discussion

Here we show the results of data collection and analysis. We split the data into nine two-week time frames to provide a reasonable balance between the following competing issues: having a large enough time frame such that there is sufficient data within each to get a good fit for the topic model, while also having a sufficiently small time frame to robustly identify the evolution of topics over time. Table 1 shows the quantity of data in our study.

We first train multiple standard LDA models to determine the best number of topics for each platform. After training 10 standard LDA models for each value of *n_topics*, we evaluate the average C_V coherence score, producing the coherence plot shown in Fig. 1. We then look at the coherences (C_V) to identify the value of the best fit *n_topics* per platform. This turns out to be 12 for Telegram, 9 for Facebook, 25 for VK, and 8 for Gab. Specifically, we do this by finding the peak in the average coherence scores which typically precedes their decay for large values of *n_topics*. We note that the platform Telegram has the most available data by far: this likely explains why the coherence scores for models trained on this data are so high relative to other platforms.

Table 1. Data quantities. Each date indicates the start date of its two-week time frame. Even though the amount of posts in each time frame is not uniformly distributed, we believe each has enough data for the models to achieve a good fit.

	Facebook	Gab	Telegram	VK
1/1	8,689	5,659	114,488	692
1/15	9,493	2,458	188,108	237
1/29	7,985	20,022	99,747	1,109
2/12	8,207	15,104	104,142	1,095
2/26	3,778	3,290	90,436	824
3/12	3,722	13,202	78,006	731
3/26	6,357	12,696	65,807	742
4/9	3,936	14,070	62,504	816
4/23	2,343	10,120	32,688	540
Total	54,510	96,621	835,926	6,786

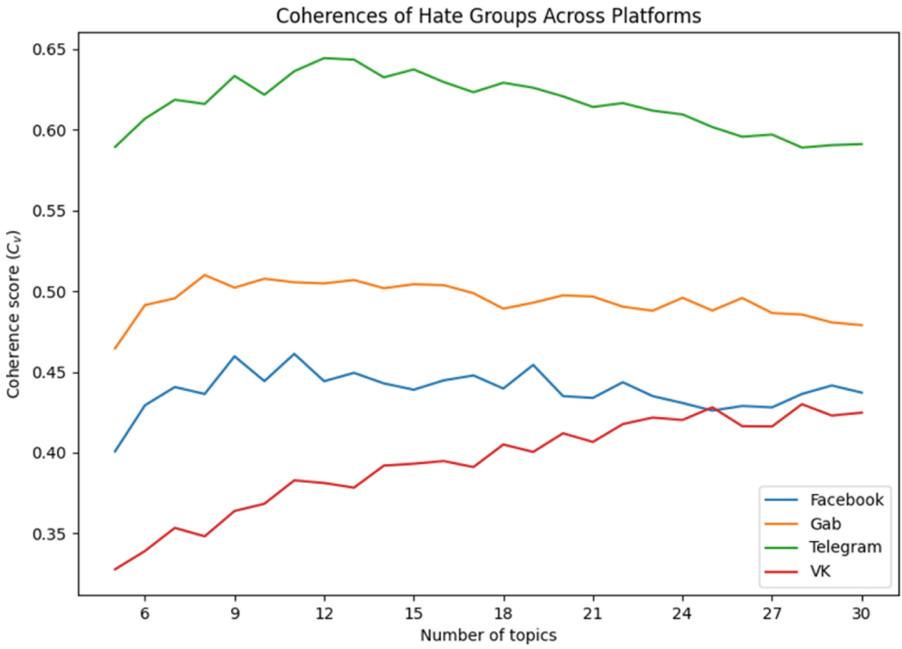

Fig. 1. For different numbers of topics (horizontal axis), the average coherence score is shown (vertical axis) for standard LDA models used to analyze the content of hateful communities within four separate social media platforms.

Figures 2, 3, 4 and 5 show the resulting coherence scores for each platform, disaggregated by topic, after performing dynamic LDA using the aforementioned *n_topics* values. Due to the implementation of dynamic LDA and our own computing restraints, we only train one dynamic LDA model for each platform. This is why the prior step of training standard LDA models to determine an optimal *n_topics* value is important as an attempt to avoid over- or underfitting.

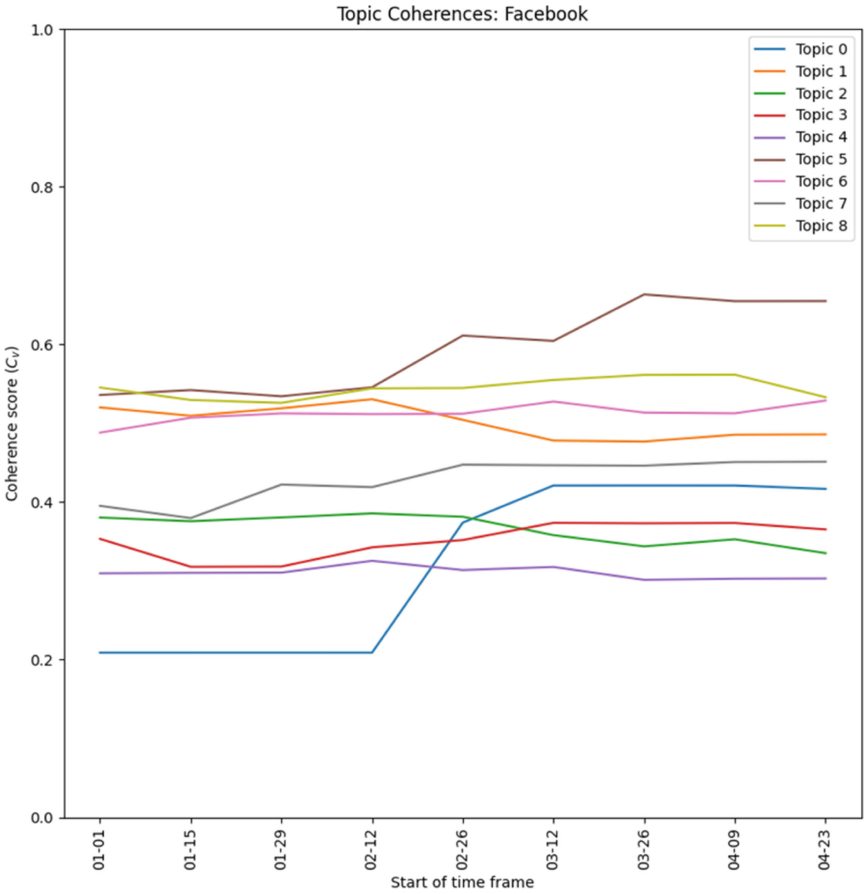

Fig. 2. Individual topics' coherence scores within a 9-topic Facebook dynamic LDA model

Of particular note is discussion of the U.S. 2020 Presidential Election on several platforms. In itself, this is perhaps not surprising (even though the study period ranges well past January 2021) given the longevity that this election had in the U.S. and to some extent world media. However, our topic modeling reveals the variations between platforms in which this event was discussed. The topics relevant to the election were Topic 10 on Telegram and Topic 5 on Gab. The keywords and their probabilities for these topics are shown in Figs. 6 and 7. Posts which contained these topics tended to discuss events related to individual states' recount efforts and generally the "stop the steal" narrative; this is also evident from analysis of the topics' keyword evolution through all time frames (the word "military" appears in the topics during mid-March). Telegram's Topic 10 was the most coherent of all topics anywhere: this suggests that Telegram acted as the primary platform where this narrative was prominently featured.

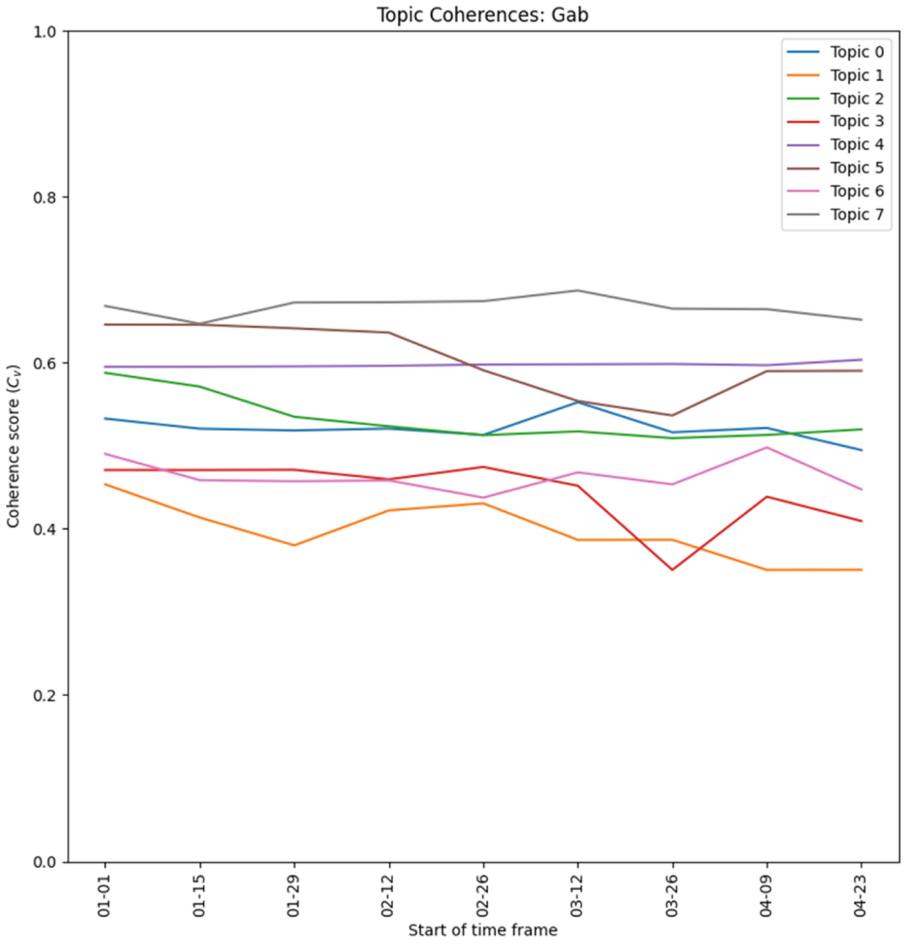

Fig. 3. Individual topics' coherence scores within an 8-topic Gab dynamic LDA model

Facebook, on the other hand, had a far lower amount of such election content: no particular topic featured keywords related to the "stop the steal" narrative or, generally, the 2020 election. Our data suggests that during our study period, fewer English-speaking white nationalists/white supremacists were active on Facebook. This is likely because of a 2019 policy introduced by Facebook concerning hate speech and violent extremism, together with increased scrutiny within the U.S. For example, there was a major deplatforming event in the summer of 2020. Our data comes from clusters of users that identify as white nationalist; that is, the communities that persisted on Facebook concentrated towards "softer," more peripheral hate narratives like white motherhood, white beauty, children's defense, and political topics like immigration. These communities have survived on Facebook because they make a point of avoiding explicit hate. By contrast,

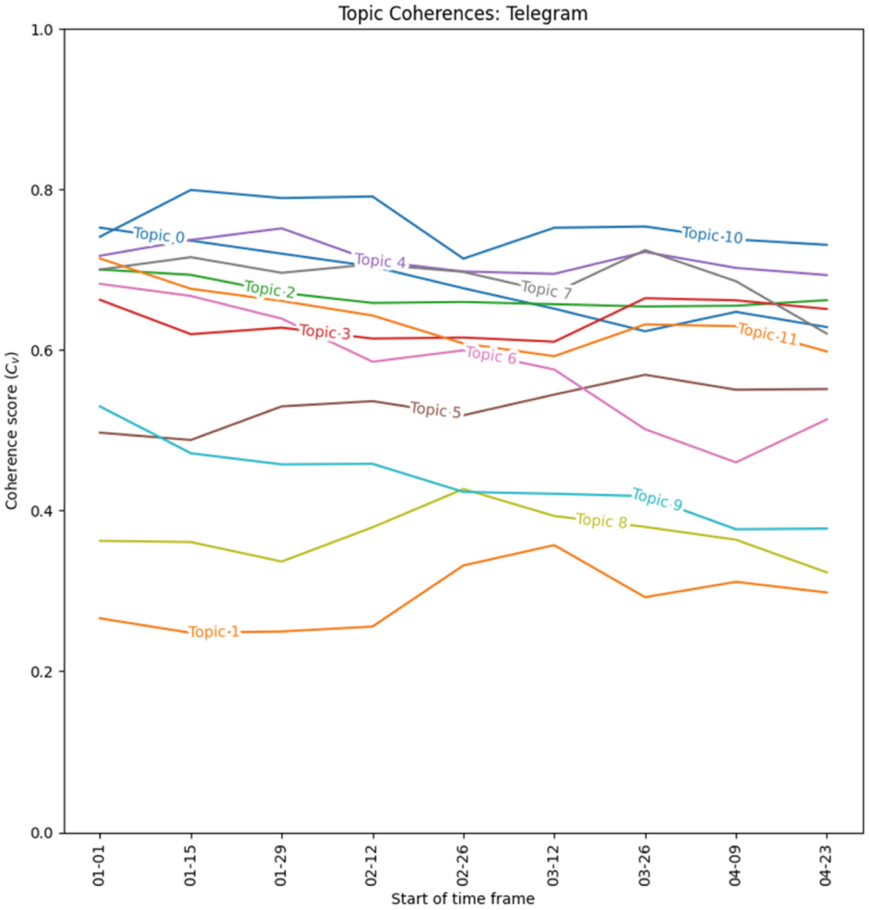

Fig. 4. Individual topics' coherence scores within a 12-topic Telegram dynamic LDA model

communities on the less moderated platforms were free to blend these "soft-hate" topics with more explicit narratives including (but not limited to) "stop the steal." This self-censorship is likely the reason the 2020 election is not as prominent among topics discovered in Facebook groups.

Interestingly, increases and decreases in coherence score can also prove useful in analyzing when communities are increasing or decreasing their interest in some broad narrative or set of conversation topics – and hence aggregating towards or fragmenting away from these things. On Telegram, for example, the most significant decrease in coherence score over our study period is shown by Topic 6. Topic 6 features discussions around getting banned or censored as well as mentions of other platforms like Parler and Twitter. There are peaks in its coherence score in January and February, which coincides with the aftermath of the January 6 Capitol riot when more mainstream, better-moderated

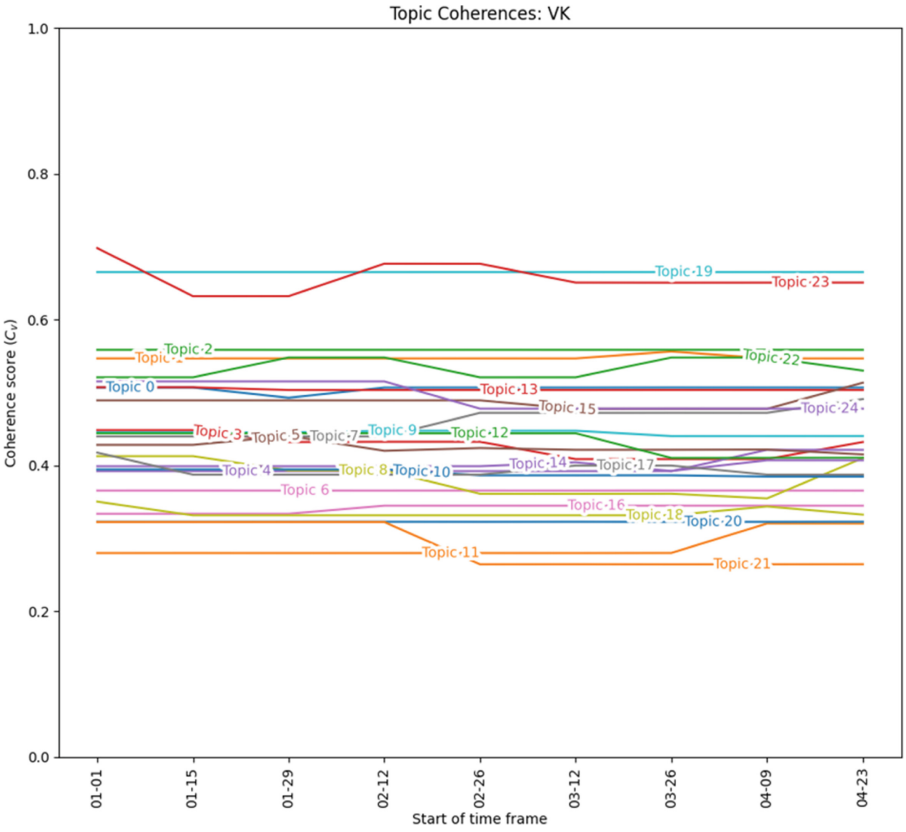

Fig. 5. Individual topics' coherence scores within a 25-topic VK dynamic LDA model

platforms like Facebook or Twitter, and web hosts like Amazon were removing communities [20]. After people were banned from these mainstream, moderated platforms, many of them migrated to Telegram [21]. The evolution of this topic demonstrates how dynamic LDA can be leveraged to detect coordination within and across platforms at the macroscopic movement level. Notably, the coherence score then decreases over March and April as users settle into their new platforms.

Finally, it is noteworthy that multimedia content is a key part of the narratives in the less-moderated platforms; specifically, videos on external websites which can help reinforce hateful narratives being expressed. On Gab, Telegram, and VK, our LDA approach found a topic that included the "youtube__com" signal, indicating links into YouTube. Topic 3 on Gab also included frequent use of video platforms Rumble and BitChute, indicating the wide variety of platforms employed to host these narratives, as well as the frequent linkage between them.

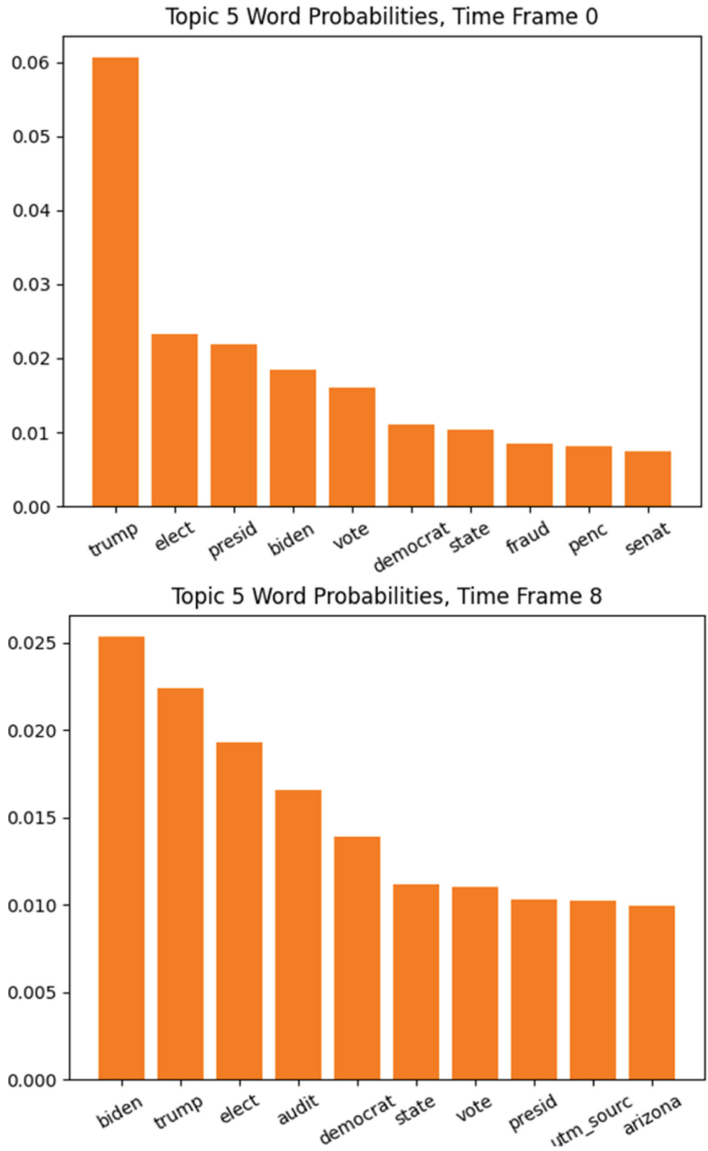

Fig. 6. Word probabilities for Gab LDA, topic 5, during the first and last time frames

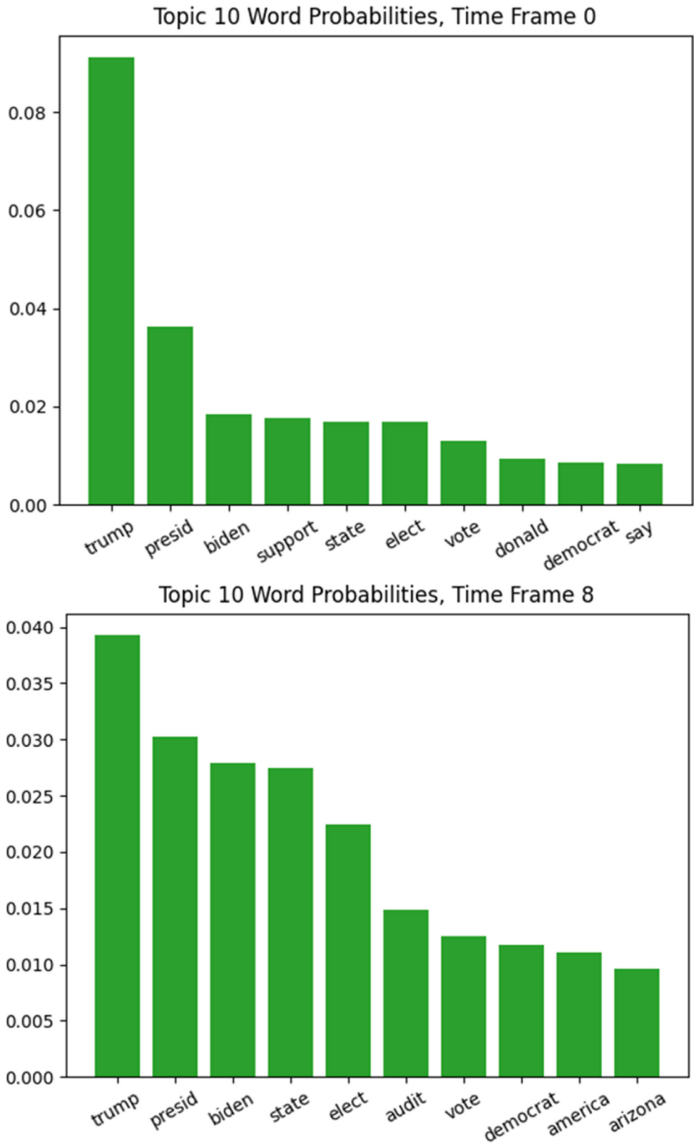

Fig. 7. Word probabilities for Telegram LDA, topic 10, during the first and last time frames

4 Limitations of the Study

Of course, much work remains to be done. It would be interesting to directly address the question of external agents or entities. Specifically, it would be useful to try to gauge how much influence such forces exert on these networks [22]. We note, however, that troll or bot-like behavior tends to be weeded out by self-policing within these online communities. We also know that more granular analysis of the types of content could prove fruitful, as well as incorporating more platforms. Shorter time frames would allow analysts to study with greater precision the ways in which these narratives evolve. Ideally, this analysis will go beyond just the use of LDA algorithms and analysis of pure text. This would be of interest since multimedia posts are very common. Further research is also required to derive actionable results for social media moderators and policymakers. Another open question is whether the structure of the network itself could aid the analysis of these narratives, or whether the topic modeling presented here could aid network analysis.

5 Conclusion

We have shown that application of simple unsupervised topic model architecture like LDA can provide significant insights into the online hate ecosystem, in particular the style of narratives users share to these communities and the ways different platforms are employed. Our methodology and machinery can potentially be used at scale to help moderation efforts across platforms and hence reduce the spread of hateful material. Specifically, we showed that a machine learning algorithm (LDA) can identify word distributions within posts from historically hateful online communities which are both plausible as distinct conversation topics and useful for gaining insights into the structure of narratives in these communities. Algorithms like LDA can not only handle huge quantities of data, but deliver results quickly. These techniques are significantly less costly than needing to rely on human labeling.

Acknowledgments. We acknowledge Rhys Leahy and Nicolás Velásquez for their help finding and downloading the data used in this study.

We are grateful for funding for this research from the U.S. Air Force Office of Scientific Research under award numbers FA9550-20-1-0382 and FA9550-20-1-0383. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

1. Velásquez, N., et al.: Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Sci. Rep. **11**(1), 11549 (2021). https://doi.org/10.1038/s41598-021-89467-y
2. Hate crime: abuse, hate and extremism online – Home Affairs Committee – House of Commons. https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm. Accessed 5 October 2021

3. Cullors, P.: Online hate is a deadly threat. When will tech companies finally take it seriously? CNN. https://www.cnn.com/2018/11/01/opinions/social-media-hate-speech-cullors/index.html. Accessed 5 October 2021
4. The year in hate and extremism 2020. Southern Poverty Law Center. https://www.splcenter.org/news/2021/02/01/year-hate-2020. Accessed 5 October 2021
5. The Daily 202: hate crimes are a much bigger problem than even the new FBI statistics show. Washington Post. [Online]. https://www.washingtonpost.com/news/powerpost/paloma/daily-202/2018/11/14/daily-202-hate-crimes-are-a-much-bigger-problem-than-even-the-new-fbi-statistics-show/5beba5bd1b326b39290547e2/. Accessed 5 October 2021
6. Vincent, J.: Facebook is now using AI to sort content for quicker moderation. The Verge, Nov. 13, 2020. https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation. Accessed 27 September 2021
7. Communities: talk about your thing with people who get you. https://blog.twitter.com/en_us/topics/product/2021/testing-communities. Accessed 27 September 2021
8. Facebook. https://www.facebook.com/policies_center/pages_groups_events. Accessed 3 September 2021
9. Removing new types of harmful networks. About Facebook, Sep. 16, 2021. https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/. Accessed 30 September 2021
10. Combating hate and extremism. About Facebook, Sep. 17, 2019. https://about.fb.com/news/2019/09/combating-hate-and-extremism/. Accessed 30 September 2021
11. Roose, K.: On Gab, an extremist-friendly site, Pittsburgh shooting suspect aired his hatred in full. The New York Times, Oct. 28, 2018. [Online]. https://www.nytimes.com/2018/10/28/us/gab-robert-bowers-pittsburgh-synagogue-shootings.html. Accessed 30 September 2021
12. Schwirtz, M.: Telegram, pro-democracy tool, struggles over new fans from far right. The New York Times, Jan. 26, 2021. [Online]. https://www.nytimes.com/2021/01/26/world/europe/telegram-app-far-right.html. Accessed 30 September 2021
13. Johnson, N.F., et al.: Hidden resilience and adaptive dynamics of the global online hate ecology. Nature **573**(7773), 261–265 (2019). https://doi.org/10.1038/s41586-019-1494-7
14. Hate Crimes. Federal Bureau of Investigation. https://www.fbi.gov/investigate/civil-rights/hate-crimes. Accessed 30 September 2021
15. Grand, A.D.: Michael Mann, Fascists. J. Mod. Hist. **78**(2), 473–475 (2006). https://doi.org/10.1086/505814
16. Sear, R.F., et al.: Quantifying COVID-19 content in the online health opinion war using machine learning. IEEE Access **8**, 91886–91893 (2020). https://doi.org/10.1109/ACCESS.2020.2993967
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
18. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning – ICML '06, Pittsburgh, Pennsylvania, pp. 113–120 (2006). https://doi.org/10.1145/1143844.1143859
19. Syed, S., Spruit M.: Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), October 2017, pp. 165–174. https://doi.org/10.1109/DSAA.2017.61
20. Trump and his allies are banned from these platforms. The Washington Post. https://www.washingtonpost.com/technology/2021/01/11/trump-banned-social-media/. Accessed 30 September 2021

21. Far-right groups move online conversations from social media to chat apps – and out of view of law enforcement. Washington Post. [Online]. https://www.washingtonpost.com/technology/2021/01/15/parler-telegram-chat-apps/. Accessed 30 September 2021
22. Broniatowski, D.A., et al.: Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am. J. Public Health **108**(10), 1378–1384 (2018). https://doi.org/10.2105/AJPH.2018.304567