Advances in Artificial Intelligence and Machine Learning; Research 2 (1) 257-272 Received 18-01-22; Accepted 08-02-22; Published 15-02-22

Dynamic Latent Dirichlet Allocation Tracks Evolution of Online Hate Topics

Richard F. Sear, Rhys Leahy
The Dynamic Online Networks Lab,
George Washington University, Washington D.C. 20052 USA
Nicholas J. Restrepo
The Dynamic Online Networks Lab,
George Washington University, Washington D.C. 20052 USA
ClustrX LLC, Washington D.C. 20007 USA
Yonatan Lupu
The Dynamic Online Networks Lab,
George Washington University, Washington D.C. 20052 USA
Department of Political Science,
George Washington University, Washington D.C. 20052 USA
Neil F. Johnson neiljohnson@gwu.edu
The Dynamic Online Networks Lab,
George Washington University, Washington D.C. 20052 USA
Department of Physics,
George Washington University, Washington D.C. 20052 USA

Corresponding Author: Neil F. Johnson.

Abstract

Not only can online hate content spread easily between social media platforms, but its focus can also evolve over time. Machine learning and other artificial intelligence (AI) tools could play a key role in helping human moderators understand how such hate topics are evolving online. Latent Dirichlet Allocation (LDA) has been shown to be able to identify hate topics from a corpus of text associated with online communities that promote hate. However, applying LDA to each day's data is impractical since the inferred topic list from the optimization can change abruptly from day to day, even though the underlying text and hence topics do not typically change this quickly. Hence, LDA is not well suited to capture the way in which hate topics evolve and morph. Here we solve this problem by showing that a dynamic version of LDA can help capture this evolution of topics surrounding online hate. Specifically, we show how standard and dynamical LDA models can be used in conjunction to analyze the topics over time emerging from extremist communities across multiple moderated and unmoderated social media platforms. Our dataset comprises material that we have gathered from hate-related communities on Facebook, Telegram, and Gab during the time period January-April 2021. We demonstrate the ability of dynamic LDA to shed light on how hate groups use different platforms in order to propagate their cause and interests across the online multiverse of social media platforms.

Keywords: Online hate, Machine Learning, Latent Dirichlet Allocation, Topic Modeling.

Citation: Richard F. Sear et al. Dynamic Latent Dirichlet Allocation Tracks Evolution of Online Hate Topics. Advances in Artificial Intelligence and Machine Learning. 2022;2(1):17.

1. INTRODUCTION

It is an outstanding problem for governments and social media platforms to know how to mitigate online hate. Such hate is known to have significant impact on its victims: from psychological harms because of racism or anti-women rhetoric, for example, but also driving some victims to harm themselves offline [1–9]. The social science literature is full of studies giving opinions about how best to prevent its spread. At the same time, social media companies do not want to blanket-ban large swathes of language because of potential legal claims over freedom of speech.

Irrespective of the best approach, a key problem facing all mitigation approaches – and one which social media companies struggle with every day – is how to identify how the main themes of hate are evolving. This is important because many moderation schemes involve taking a static list of hateful words or phrases as a 'redlist' lookup TABLE, in order to judge whether future postings should be banned or not. The obvious problem is that these topics can evolve and substitute terminology adopted in order to prevent new hateful language from matching the redlist and hence catching moderators' attention. The same problem would face any machine learning or AI tool that does not embrace the reality that topics in hate can evolve, either naturally or on purpose, and that this evolution may be daily or slower. This is important because it is known that Facebook does use such AI to inform its moderation process [10].

The main contribution of this paper is to show a modification of a standard machine learning tool, specifically a dynamic version of Latent Dirichlet Allocation (LDA) can be used together with standard LDA [11], in order to track the evolution of topics that evolve in online hate. While the same techniques could also be applied to non-hate material, we analyze here online hate because it is often nuanced and nudged by its proponents in order to avoid attracting moderator attention – hence a significant evolution of hate-related topics over time is to be expected. In addition to laying out a new use for dynamic LDA, our findings from the LDA analysis help shed new light on what enables hate groups to organize and then coordinate. They also shed light on differences and similarities between the content that distinct social media platforms share, and hence the extent to which these platforms each play a niche role. More broadly, our paper contributes to the existing literature by helping provide a better understanding of how online hate evolves and hence how it might be better controlled or even eradicated.

FIGURE 1 provides a schematic of our methodology in this study. The structure of the rest of the paper is as follows. In Sec. II we explain where our data comes from, namely the online social media platforms on which there are built-in communities (e.g. Facebook Pages). These in-built communities are known to create a highly attractive space for promoting and developing hate speech and recruiting new followers. For this reason we do not use Twitter, since its version of community spaces is under development and not yet widely available [12]. In Sec. III, we introduce LDA and dynamic LDA. We then explain the methodology of our dynamic LDA study. In Sec. IV we present our findings. Sec. V contains limitations of our study. Sec. VI presents our conclusions and a brief overview of future work that leads on from this study.

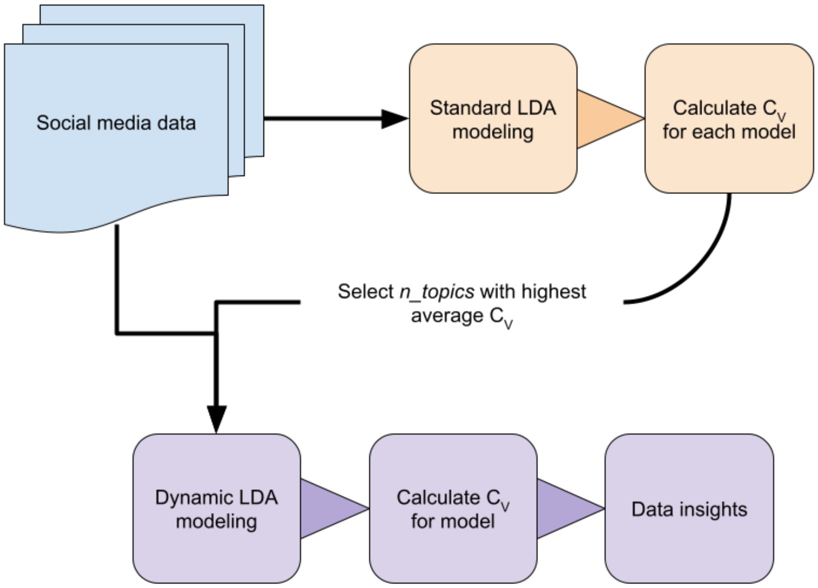
Social media data
Standard LDA modeling
Calculate C_V for each model
Select *n_topics* with highest average C_V
Dynamic LDA modeling
Calculate C_V for model
Data insights

Figure 1: Flowchart of our methodology in this study

2. RELATED WORK

2.1 Online Hate

There are many studies of online problematic material, including hate as well as misinformation. We refer to References 1-25 for a sampling of the discussions [1-25], surrounding the complex issue of online material and moderation together with potential analysis and mitigation tools, including from government policy and legal perspectives as well as for different platforms [21-25]. In particular, Ref. 24 contains a systematic review and critique of the topic of hate speech on social media. It makes the point that social media collectively has provided a space for hate to thrive. This includes the use of added meme weaponization as well as the use of fake personal details [24]. Meanwhile, Ref. 25 showed that Reddit tends to allow the emergence of toxic subcultures, while right racist influencers often emerge on YouTube [24]. Furthermore, it has been concluded that Twitter allows for coordinated harassment [26]. Even the seemingly innocent use of emojis and GIFs can contribute to the spread of hate [24]. Operationally, despite decades of technical, social, and legal efforts to do so, controlling the spread of this content has proved elusive. Critics tend to blame the largest social media companies, such as Facebook, for failing to eradicate it. However, this misses the point that social media platforms are generally separate universes, i.e. operationally and commercially independent, and are often situated in independent legal jurisdictions (e.g. Facebook in U.S. and VKontakte in Russia). Hence any unilateral action by a single company is necessarily limited to reducing such malicious matter down to a tiny fraction of isolated incidences within its own universe. Criticizing the largest companies also ignores the fact that there are a multitude of smaller social media platforms being created all the time, thanks to freely available software that enables decentralized setups across servers and locations. While the current study is not intended to address this far-reaching issue, the procedure we introduce is agnostic of platform and can be used to understand the evolution of hate topics within and across a multiplatform setting.

2.2 Dynamic Latent Dirichlet Allocation

The model we employ in this study, LDA, is a widely-used unsupervised machine learning model. Its dynamic version (referred to here as "dynamic LDA" and in our code's implementation as "sequential LDA"), introduced in [17], is less frequently employed in the literature. One study applies the model to the domain of expert identification to tackle a similar issue as we face in the domain of online hate speech: the language used in the domain tends to evolve more rapidly than the original standard LDA model was designed to handle [27]. Other published research proposes more advanced or more efficient dynamic topic model architectures [28, 29]. However, to our knowledge, no other studies have applied the widely available Gensim implementation of a dynamic topic model to the domain of online hate speech.

3. COLLECTION OF SOCIAL MEDIA HATE CONTENT

Here we describe the data collection process. This corpus of data provides the input to the machine learning algorithms. Though this data collection stage does not itself involve machine learning, it might in the future. Hence, we lay it out in detail in the hope that future work can help provide a reliable yet fully automated and more rapid process by using machine learning instead of manual human work.

The first step of our data collection involves obtaining a list of in-built communities that feature hate content, drawn from across social media platforms. These in-built communities are referred to by different names according to the platform: e.g. Facebook Page, Telegram Channel, and Gab Group. The second step of our data collection involves collecting all the posts from the communities in this list. We stress that all this data is publicly available and that an in-built community is different from a personal account like a personal Facebook page. We do not access any such personal pages or accounts, even though they might also be publicly available. Specifically, the Facebook Pages, Groups, and Events Terms & Policies page notes that "Content posted to a Page is public and can be viewed by everyone who can see the Page" [13]. We also note that the self-weeding tendency within such in-built communities tends to capture and remove fake profiles and bots. Hence, we can be reasonably confident that we are looking at the hate-related activity of many humans online.

To perform the first step of obtaining a list of hate-related, in-built communities from across social media platforms, our subject matter experts started with a seed list with obvious hateful content and then looked at what communities they then linked to. To establish what qualified as hateful content, they used a well-established set of criteria that has itself been previously published and scrutinized by other experts [3]. The process was as follows: two subject-matter experts who had been trained for several years on analyzing right-wing extremism started by manually reviewing the most recent 20 posts in each community. When both reviewers agreed that two of the most recent 20 posts showed hate and bias against the protected classes listed in the FBI classification of hate crimes, this in-built community was labeled as a hate-related community for this study [14]. To help clarify the meaning of hate, the subject matter experts also profited from Michael Mann's discussion of racially and ethnically motivated violence in fascism as being the pursuit of "cleansing nation-statism through paramilitarism" [15]. Applying these definitions, they obtained a list of communities that included organized hate groups such as the KKK, in addition to more decentralized movements such as Boogaloo groups. For simplicity, the present study is focused around content in English. To automatically download the posts into a tabular format, we use API access for each platform.

After obtaining the posts, we used Google's Compact Language Detector in order to label each post's language. We note that other languages can be treated using the same methodology as we lay out here. We did not limit our study to any particular geographical region. The posts that we included all appeared in the period January 1 to April 30, 2021. We also used links within the communities to extend our list of hate-related in-built communities. This snowball-like process for collecting hate-related in-built communities yielded a larger list which was then classified as well.

By including a range of social media platforms, from well-regulated and widely-used (Facebook) to smaller, less regulated and less used platforms, we achieved a broad spectrum of hate content – with words and language ranging from subtle to blatant. It is known that Facebook introduced new moderation policies in the period 2019-2021, to try and reduce hate speech [6, 7]. The opposite seemed to hold for Telegram and Gab, which seemed instead to purposely avoid such moderation. In this way, Telegram and Gab were able to appeal to users interested in unmoderated free-speech and hence grow a corresponding user-base.

We then implement the second step of the data collection process. Specifically, we captured publicly available posts that were within these hate-related communities in our final list. This gave us a corpus of text by platform, from their respective hate-related communities. This text for a given platform provided the input to our machine learning (Fig. 1).

4. METHODOLOGY OF OUR MACHINE LEARNING ANALYSIS

The content from the hate-related in-built communities was bundled separately for each social media platform (Facebook, Gab, and Telegram) for the machine learning stage. Our methodology for the machine learning portion of the study is outlined in FIGURE 1 and uses both standard LDA and dynamic LDA. The advantage of dynamic LDA modeling is that it takes into account the timing of input documents. However, it is cumbersome and inefficient to run in bulk, hence we could not rely exclusively on dynamic LDA models. Therefore, we adopted a compromise methodology where we first employ the highly efficient standard LDA over the entire dataset. We train 10 randomly initialized models for each "*n_topics*" (discussed later), calculate the coherence score for each, and then average those scores to obtain an average coherence score corresponding to each value of *n_topics*. The highest value of *n_topics* is used for the input to the dynamic LDA model. This procedure therefore yields a dynamic LDA model best fit to a dataset without the time-consuming process of trying multiple *n_topics* values for many dynamic LDA models.

We now explain the detailed steps in this methodology (Fig. 1) and its rationale. The machine learning tool Latent Dirichlet Allocation (LDA) [16] is a powerful way to deduce topics in text. LDA works by modeling documents as distributions of topics, and in turn topics as distributions of words. LDA is therefore a generative statistical model that treats documents as mixtures of a small number of topics, and each topic is made up of a collection of words. We refer to Ref. 11 for a summary. The training stage involves these distributions being adapted in order to fit the dataset. Carrying out LDA on each timestep's data would produce a far too abrupt shift in topics. This is because LDA involves a separate maximization (optimization) process every time it is applied. It can therefore end up choosing a particular set of topics that looks very different for each timestep, even though there were suboptimal choices that by contrast persisted for many timesteps. Hence, we could not rely entirely on LDA and instead also employ a more dynamical version – dynamic LDA – which incorporates information about when a post appeared. This gives it the ability to follow the evolution in choice of words in particular topics as they change in time [17]. Like the regular non-dynamic LDA, the dynamic LDA method is completely unsupervised. The only input that we need to supply to the LDA and dynamic LDA models, apart from the text, is the "number of topics" parameter (referred to earlier as *n_topics*) which tells the algorithm the number of sets into which to cluster text. For convenience, we employed the Gensim implementation for both the LDA and dynamic LDA, which is freely available at https://radimrehurek.com/gensim/.

To assess how "coherent" the topics were that the standard and dynamic LDA algorithm identified – i.e. how clearly the topic stood out as stand-alone from the perspective of the algorithm – we use the measure called the 'coherence score' [18]. This is a popular choice for measuring quantitatively how well words are aligned within an identified topic, and acts as an effective "goodness of fit" evaluation technique. It exploits normalized pointwise mutual information and the cosine similarity and is made up from collections of probability measures that determine how frequently top words in topics co-occur with each other in instances where the topics appear. We evaluate the coherence score for a set of topics output from the standard and dynamic LDA, using a separate algorithm that is dedicated to evaluating the coherence score. The overall coherence score for a single model is the arithmetic mean over its separate coherence scores per topic. Of course, there are many other possible choices of metric; however, we find the C_V coherence score particularly intuitive, in addition to its strong performance in other studies [18]. In the results shown in the next section and the FIGURES, coherence is therefore labeled C_V. It comprises collections of probability measures on how often top words in topics co-occur with each other in examples of the topics. It is interesting and reassuring that visual inspection of the word distributions that emerge for each topic output by the dynamic LDA, and having reasonably high coherence scores, do make sense as distinct conversation topics.

We now give more details of the specific implementation and examples. Before training the machine learning models, we carried out various steps to clean the content using a similar approach to other works in the literature. We have previously used this preprocessing procedure in an LDA analysis of coronavirus-related narratives in anti-vaccine communities [30].:

1. We remove mentions of URL shorteners, e.g. "bit.ly" which are pieces of text output by the APIs on some platforms.
2. It turns out that many posts link to external websites which could be an interesting component of the conversations. Hence instead of removing them completely, we replaced the domain pieces ".gov", ".com", and ".org" by "__gov", "__com", and "__org" respectively. Doing this

ensured that they would not be removed by later preprocessing. Hence "whitehouse.gov" was turned into the token "whitehouse__gov".

3. We unwrapped contractions such as "don't" to become "do not."
4. We then ran the posts through Gensim's *simple_preprocess* function. This has the effect of tokenizing the post on spaces and removing tokens that are very short, i.e. 1 or 2 characters. It also removes punctuation and numeric characters.
5. We removed tokens such as "the" that are in Gensim's list of stopwords, since these are not good indicators of topics.
6. We lemmatize tokens using the WordNetLemmatizer from the Natural Language Toolkit (NLTK). This serves to convert all words to singular form and/or present tense.
7. We stemmed tokens using SnowballStemmer from NLTK. This serves to remove affixes on words.
8. We removed any remaining URL fragments (other than domain) that remained after stemming, e.g. "www" and "http".

Steps 5-7 help make the comparison of words fair during the LDA training process. If a particular word acts as a strong indicator of a topic, this signal does not get lost simply because it appears in many different forms. For example, "runs," "running," and "run" all refer to the same concept, but without being converted to a standard form (Step 6) and converted to their root form (Step 7), they could register as three different signals and become lost as background noise. This preprocessing utilizes the set of words in NLTK's pretrained vocabulary, meaning that any word not in the vocabulary remains unchanged.

Following this preprocessing, the LDA and then the dynamic LDA model (Fig. 1) can be trained on this cleaned data. The dynamic LDA is also given the metadata concerning the time frame within which each post appeared. Though more dynamic LDA models could be trained, doing so is computationally expensive and so for the illustrative purpose of this paper, we simply trained one dynamic LDA model per *n_topics* parameter. For the standard LDA step (Fig. 1) 10 models were chosen per value of *n_topics*. The parameter range that we chose for *n_topics* ranged from 5-30. Then the C_V coherence algorithm and the coherence scores were averaged for each number of topics. Multiple trials were run for each number of topics to make sure that the coherence for that number of topics represents what the model tended to find generally and hence was not simply being dominated by a particular overfit run. A plot of the average C_V score per *n_topics* is shown in FIGURE 2.

Using these coherence scores, we determine that the best fit *n_topics* parameter per platform is 9 for Facebook, 8 for Gab, and 12 for Telegram. We determine this by finding a peak in the average coherence scores (typically followed by a series of slowly decreasing scores). This is expected behavior for testing several values of *n_topics*. Telegram has by far the most available data, which likely explains the high coherence scores for models trained on this data; these models were best able to find topic distributions that fit the posts well.

We refer to https://github.com/gwdonlab/topic-modeling for the codebase used in our study. Similar experiments can be carried out on any text dataset using this library.

5. FINDINGS

The results shown in this section come from our analysis, using the methodology of Sec. III (Fig. 1), of the development of conversations from January 1, 2021 until the end of April, 2021 (the last day we have data available for all platforms). This was a particularly relevant period of study, since it accompanied the widespread rollout of vaccines which was met with significant resistance, together with lockdowns and mask wearing. It also encompassed political events related to the 2020 U.S. Presidential election, namely the attack on the U.S. Capitol and continued online proliferation of false narratives about the election being “stolen”. All of these helped fuel online hate.

We split this data into nine two-week timeframes. Using two-week frames gave a good trade-off between having enough data within each timeframe for the topic model to get a good fit and having a small enough timeframe in order to understand the evolution over time of the topics. The quantity of data in our study as measured in number of posts, is shown in TABLE 1.

Table 1: Data quantities. Each date indicates the start date of its two-week time frame.

	Facebook	**Gab**	**Telegram**
1-Jan	8,689	5,659	114,488
15-Jan	9,493	2,458	188,108
29-Jan	7,985	20,022	99,747
12-Feb	8,207	15,104	104,142
26-Feb	3,778	3,290	90,436
12-Mar	3,722	13,202	78,006
26-Mar	6,357	12,696	65,807
9-Apr	3,936	14,070	62,504
23-Apr	2,343	10,120	32,688
Total	**54,510**	**96,621**	**835,926**

Following the process outlined in FIGURE 1 and Section II, we trained many LDA models and calculated their C_V coherence scores.

We then use the optimal *n_topics* parameter values for each platform in order to train the dynamic LDA models at two-week intervals over the study period. We then calculated the C_V coherence scores for the dynamic LDA models in each timeframe. FIGURES 3-5 show these coherence scores broken down by topic, for each platform.

Despite the fact that all the processing so far has involved machine learning tools that are not specifically designed to treat hate content online, the results produced (Fig. 3-5) can be used to extract new insights about the hate ecosystem. These results hence help demonstrate the potential value of machine learning for this complex societal problem area. Topics emerge on several platforms associated with the U.S. 2020 Presidential Election, which is not surprising, though we note that our period of study extends well past the inauguration in January 2021. But there are also substantial differences between platforms. On Gab and Telegram, the topics relevant to the election were Topic 5 and Topic 10 respectively. Analysis of the words in these topics (see FIGURE 6 for Topic 10) reveals that posts containing these topics were focused on events related to the “stop the steal” narrative and individual states’ recount efforts. We find a similar message from the evolution of

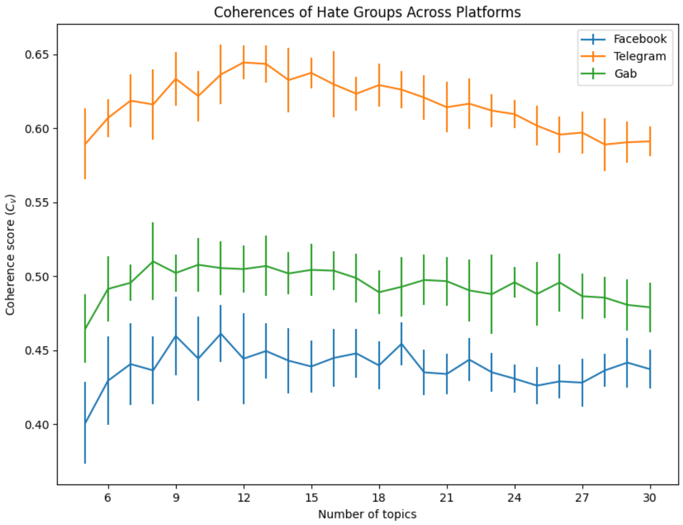

Figure 2: Average coherence score for different numbers of topics and for separate platforms. The topics were obtained from the standard LDA model, whose input was the content extracted from hate-related in-built communities on different platforms.

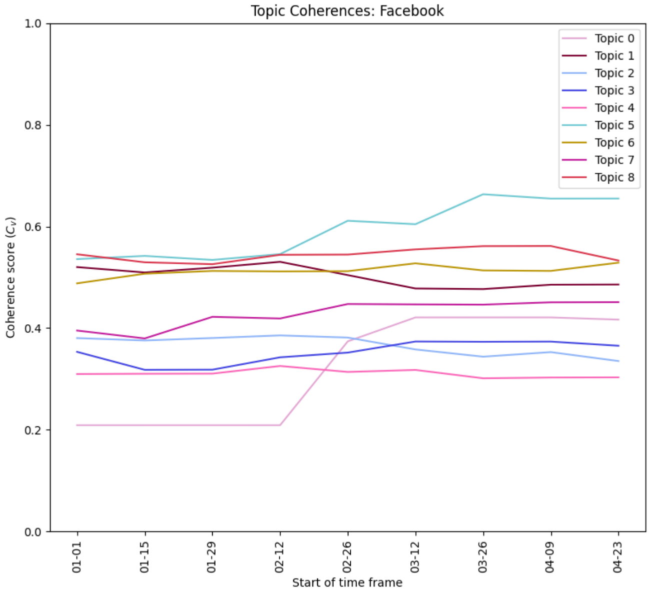

Figure 3: Coherence scores for the topics uncovered by the 9-topic dynamic LDA model. The dynamic LDA model input was the content extracted from hate-related in-built communities on Facebook.

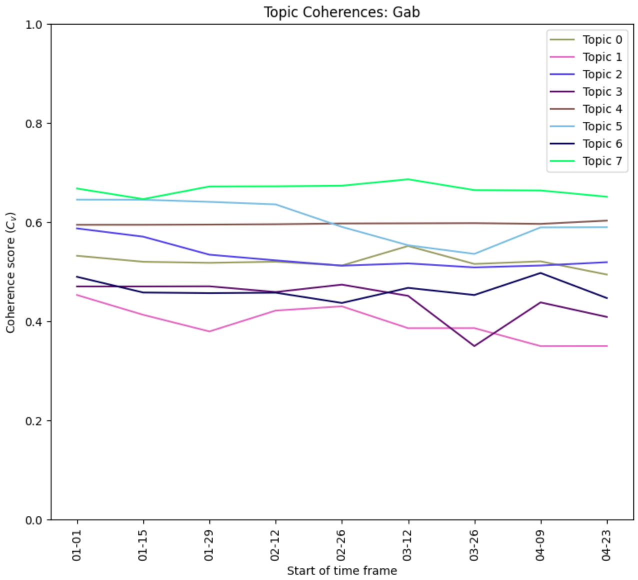

Figure 4: Coherence scores for Gab topics discovered by an 8-topic dynamic LDA. The dynamic LDA model input was the content extracted from hate-related in-built communities on Gab.

the topics' keywords. For example, the word "military" appears during mid-March. Also, our findings show that the most coherent of all topics anywhere, was Telegram's Topic 10 (FIGURE 6). This suggests that Telegram was the primary platform on which this narrative was able to establish traction. FIGURE 6 also shows how this Topic 10 became broader over time, in that the dominance of the probabilities of the top words (e.g. Trump) decreased over time.

By contrast, there was much less discussion on Facebook of the election in this context. Moreover, keywords related to the "stop the steal" narrative and the 2020 election in general did not appear in any particular topic. This is perhaps understandable because of Facebook's 2019 policy concerning hate speech and violent extremism [7]. In addition, fewer English-speaking white nationalists/white supremacists were active on Facebook during our study period because of increased scrutiny in the U.S. through 2020. Indeed, there was a major deplatforming event in the summer of 2020. Instead, the persistent groups on Facebook during this period were more focused on peripheral or "soft-hate" narratives concerning children's defense, white beauty, white motherhood and political topics like immigration. By avoiding explicit hate, these groups have managed to survive for long periods of time on Facebook. In contrast, there is significant mixing on other platforms of these "soft-hate" topics with explicit hate or conspiracy narratives like "stop the steal", because clusters on the other (unmoderated) platforms have free reign to do so and hence explore new narratives that push the boundaries. We speculate that such self-censorship is why the 2020 election does not feature prominently among the topics that our analysis uncovered in these Facebook communities.

Another finding from our analysis is that increases and decreases in coherence score can play a useful role in detecting when online communities on a given platform are coalescing around – or

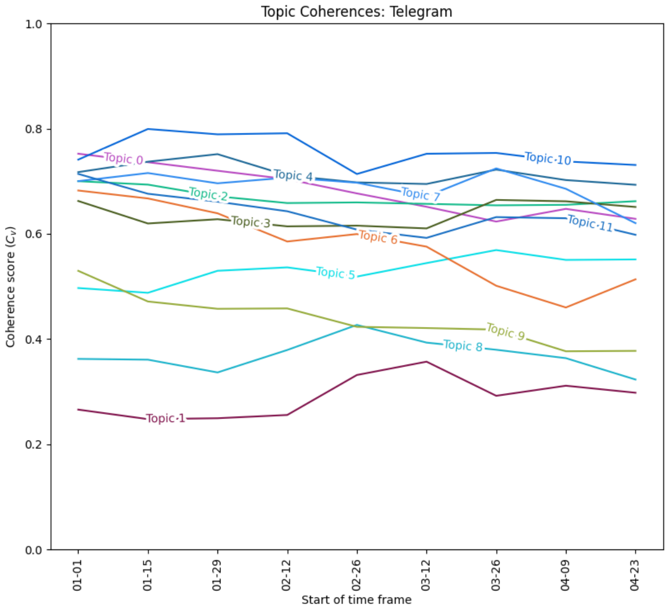

Figure 5: Coherence scores for Telegram topics discovered by a 12-topic dynamic LDA. The dynamic LDA model input was the content extracted from hate-related in-built communities on Telegram.

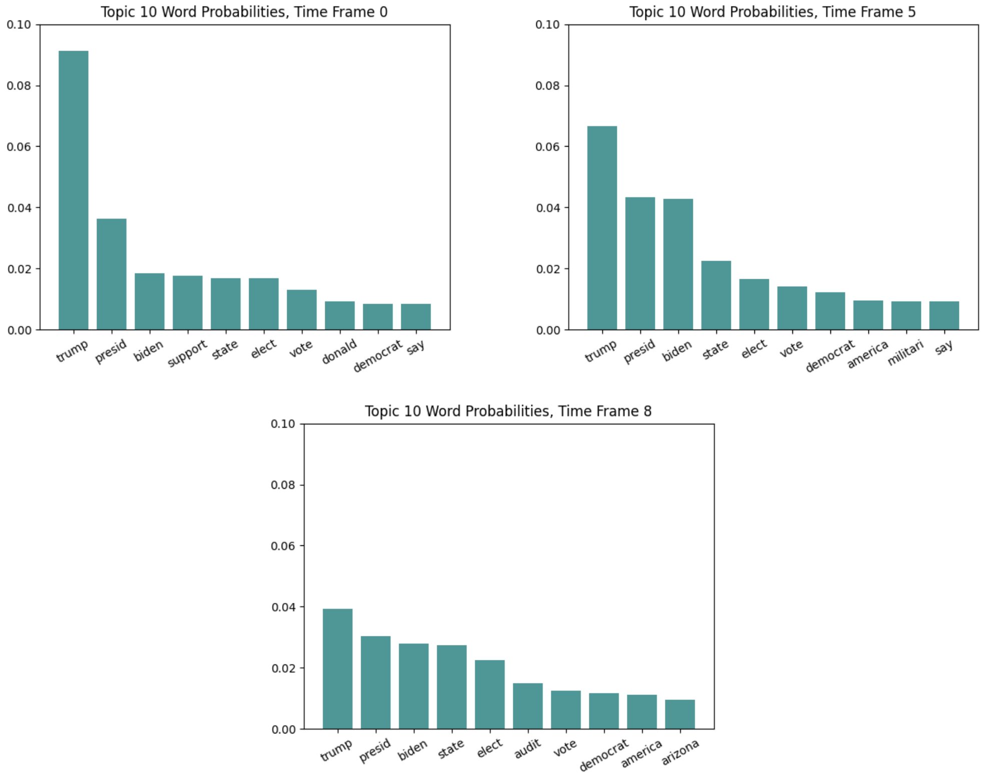

Figure 6: Keywords and associated probabilities from dynamic LDA analysis of Telegram, Topic 10. This shows the evolution of words used within this topic.

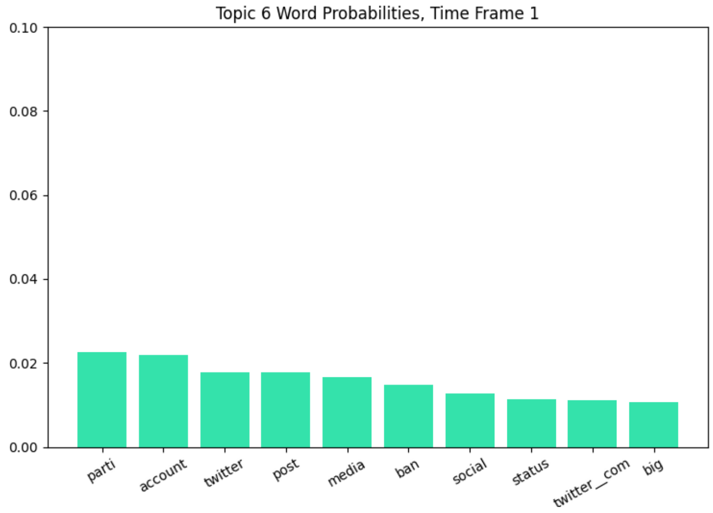

Figure 7: Example of Telegram Topic 6 keywords relating to Twitter and deplatforming efforts

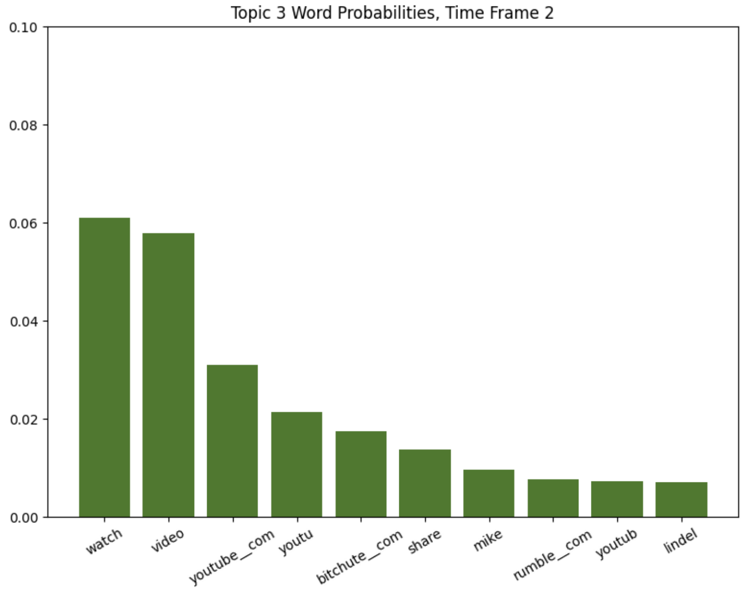

Figure 8: Example of Gab Topic 3 keywords referencing external video-hosting platforms

diverging away from – particular sets of topics or some larger narrative. An example of this comes from Topic 6 on Telegram, which is the one that shows the most significant decrease over the period of the study. This topic is dominated by references to other platforms like Twitter and Parler, and also includes discussions around getting censored or banned (see FIGURE 7). There are peaks in the coherence score for the topic in January and February: this happens to coincide with the moment when mainstream moderated platforms like Facebook and Twitter, and even web hosts like

Amazon, were actively removing users, groups, and entire apps and websites that were involved in coordinating the January 6 Capitol riot [19]. There was significant migration of many of these to Telegram [20] after they were banned from these mainstream, moderated platforms.

Seeing how this topic emerges in time on Telegram shows explicitly how dynamic LDA could be used to detect coordination at the movement level within and across platforms. It is also interesting to see that in March and April, the coherence score then decreases, presumably as individuals settle into their new platforms.

Last, it is also interesting to see that a significant portion of the content in the two unmoderated social media platforms that we study, uses multimedia content. Specifically, this includes particular videos hosted on external websites that support the hateful narratives being expressed at the time. The LDA was able to find a topic on Telegram and Gab that featured the "youtube__com" signal, hence demonstrating specific links to YouTube videos. On Gab, Topic 3 included frequent references to the video platforms Rumble and Bitchute (see FIGURE 8). Hence our LDA analysis has managed to reveal that the wide variety of platforms available to host hateful content are successfully held together by frequent inter-platform links.

6. LIMITATIONS OF OUR STUDY

Our study has various limitations, each of which represents an opportunity for future research. First, we cannot categorically discount the impact of external agents or entities in dictating these topics. However, these social media communities do tend to police themselves for bot-like or troll behavior. Moreover, our results still stand irrespective of how the topic rose to prominence. Intriguingly, it might eventually be possible to use the coherence score of topics detected by our methodology of LDA machine learning (Fig. 1) to see if organic texts can be distinguished from engineered ones. We leave that to a future study. Second, further analysis of the details of the content could be interesting, for example, by extending this to more platforms and other languages. Third, it would be informative to explore shorter timeframes so that more granularity would emerge concerning how the narratives and topics evolve. Fourth, we should explore content beyond just text and LDA, e.g. multimedia posts whose hateful content exists in images or videos, perhaps without containing hateful text. Finally, it is worth exploring how these results can be turned into actionable interventions for policy makers. An idea in this direction is to see when a new hate-related topic suddenly emerges and then gains in its coherence score rapidly.

7. CONCLUSIONS AND FUTURE WORK

It is well-known that online hate content not only spreads easily between social media platforms, but its focus can also evolve over time. We have demonstrated how machine learning tools could play a key role in helping human moderators understand how such hate topics are evolving online. Specifically, we showed that a dynamic version of Latent Dirichlet Allocation (LDA) can be used to capture the way in which hate topics evolve and morph. We used it to analyze the topics over time emerging from extremist communities across multiple moderated and unmoderated social media platforms. Our dataset comprised material that we gathered from hate-related communities on

Facebook, Telegram, and Gab during the time period January-April 2021. We demonstrated how this dynamic LDA sheds light on how hate groups use different platforms in order to propagate their cause and interests across the online multiverse of social media platforms.

There are many opportunities for future work. First, we would like to explore if it is possible to use the coherence score of topics detected by our methodology of LDA machine learning (Fig. 1) to see if organic texts can be distinguished from synthetic ones. Second, we would like to carry out more detailed analysis of the content by extending the study to more platforms and other languages. Third, we want to explore shorter timeframes so that more granularity would emerge concerning how the narratives and topics evolve. Fourth, we note that while the topic analysis presented here is fully automated, the process of adding context to the discovered keywords must still be carried out manually. A more powerful machine learning model with knowledge of word embeddings and broader context within the corpora of the social media platforms, could help to automate this process as well. Finally, we would like to understand better how these results can be turned into actionable interventions for policy makers. Though much work still needs to be done, this study is an early technical framework for a fully automated but interpreTABLE understanding of multi-platform hate speech narratives.

8. ACKNOWLEDGMENT

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.

9. CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

References

[1] Johnson NF, Leahy R, Restrepo NJ, Velasquez N, Zheng M, et al. Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology. Nature. 2019;573:261.

[2] Johnson NF, M Zheng, Y Vorobyeva, A Gabriel, H Qi, et al. New Online Ecology of Adversarial Aggregates: ISIS and Beyond. Science. 2016;352:1459.

[3] Velásquez N, Leahy R, Restrepo N Johnson, Lupu Y, Sear R, et al. Online Hate Network Spreads Malicious COVID-19 Content Outside the Control of Individual Social Media Platforms. Sci. Rep. 2021;11:11549.

[4] https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm

[5] https://www.cnn.com/2018/11/01/opinions/social-media-hate-speech-cullors/index.html

[6] https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks/

[7] https://about.fb.com/news/2019/09/combating-hate-and-extremism/

[8] https://www.nytimes.com/2018/10/28/us/gab-robert-bowers-pittsburgh-synagogue-shootings.html

[9] https://www.nytimes.com/2021/01/26/world/europe/telegram-app-far-right.html

[10] https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation

[11] https://arxiv.org/abs/1909.01436

[12] https://blog.twitter.com/en_us/topics/product/2021/testing-communities

[13] https://www.facebook.com/policies_center/pages_groups_events

[14] https://www.fbi.gov/investigate/civil-rights/hate-crimes

[15] Grand AD. Michael Mann, Fascists. J. Mod. Hist. 2006;78: 473.

[16] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993.

[17] Blei DM, Lafferty JD. Dynamic Topic Models. In Proceedings of the 23rd international conference on Machine learning - ICML ’06, Pittsburgh, Pennsylvania, 2006;113-120.

[18] 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). 2017:165–174.

[19] https://www.washingtonpost.com/technology/2021/01/11/trump-banned-social-media/

[20] https://www.washingtonpost.com/technology/2021/01/15/parler-telegram-chat-apps/.

[21] https://www.parliament.uk/documents/commons-committees/home-affairs/Correspondence-17-19/Hate-crime-abuse-hate-and%20extremism-online-Government-Response-to-Fourteenth-Committee-Report-16-17.pdf

[22] Gagliardone I, Gal D, Alves T, Martinez G. Countering online hate speech. Unesco Publishing, 2015.

[23] Banks J. Regulating hate speech online. International Review of Law, Computers & Technology. 2010;24:233-239,

[24] Matamoros-Fernández A, Farkas J. Racism, Hate Speech, and Social Media: A Systematic Review and Critique, Television & New Media. 2021;22:205-224.

[25] Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J. You Can’t Stay Here: The Efficacy of Reddit’s 2015 Ban Examined Through Hate Speech. Proc. ACM Human-Computer Interact., 2017;1:1-22.

[26] Shepherd A, Sanders C, Doyle M, Shaw J. Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. BMC Psychiatry. 2015;15:29.

[27] Chi R, Wu B, Wang L. Expert Identification Based on Dynamic LDA Topic Model. 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 2018: 881-888,

[28] Wang C, Blei D, Heckerman D. Continuous Time Dynamic Topic Models, in Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland. 2008:579-586.

[29] Bhadury A, Chen J, Zhu J, Liu S. Scaling up Dynamic Topic Models, in Proceedings of the 25th International Conference on World Wide Web, Montréal, Québec, Canada. 2016: 381-390.

[30] Sear RF, Velásquez N, Leahy R, Restrepo NJ, Oud SE, et al., Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. IEEE Access. 2020;8:91886–91893.