Topic Modeling of StormFront Forum Posts

Grigorii Nazarko
Dept. of Business and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
email: grinazarko (at) gmail.com

Richard Frank
School of Criminology
Simon Fraser University
Burnaby, Canada
email: rfrank (at) sfu.ca

Magnus Westerlund
Dept. of Business and Analytics
Arcada University of Applied Sciences
Helsinki, Finland
email: magnus.westerlund (at) arcada.fi

***Abstract*—Radicalized communities are actively using the internet to preach their ideas within society. Hence, it is crucial to research the content of such communities to understand their agenda and potentially take actions. The discussion is spread out in different forums on the Internet, but in this study, we used natural language processing technologies for revealing topics discussed in the oldest right-wing forum StormFront. As a result, we found the co-occurrence of discussed topics and real-world events, which means that the method can be used to track the agenda and changes in the community.**

Keywords-StormFront; topic model; information extraction; NLP; LDA

I. Introduction

Although most of the population uses the internet in a benign manner, there are groups, like political extremists of any conviction, that (ab)use the internet in order to facilitate the organization of violent events. In addition to simple coordination, extremists have been found to use the internet for all manner of benefits to their group, such as information provisioning, financing, networking, recruitment and information gathering [1]. Recently, there has also been evidence of some violent political extremists and terrorists being online immediately prior to their attacks [2], such as the Christchurch, New Zealand shooter in 2019 who, before live-streaming his attack on two mosques, posted his manifesto and shooting intentions on the imageboard "8chan" [3]. Further studies have also shown that, especially in extremist communities, online education, communication, and training are much more prevalent [4]. For example, of the extremists who plotted to attack government targets (as opposed to civilian targets) 83% "displayed traits of online learning", and in the context of violent extremism using Improvised Explosive Devices (IED), those who plotted to use the IEDs were 3.34 times more likely, while those who actually used IEDs were 4.57 times more likely, to use online sources for their information [4]. Clearly, extremists in general, make extensive use of the internet for their preparations.

As empirical research is suggesting that online resources are increasingly playing important roles in mobilizing extremists to violence, and that groups and movements mobilize themselves online, law enforcement agencies are increasingly recognizing the need to better understand these online mobilization efforts [5]. Online indicators of malicious behaviour are being developed by multiple law enforcement agencies (for an example, see [6]) to try to detect imminent dangers, by looking for specific actions by users, such as the posting of martyrdom videos or statements, attempting to mobilize others, or engaging in ideological discussions.

These indicators of imminent dangers are highly qualitative and difficult to detect. They are usually simple posts on online discussion forums or social media. One way to detect such dangers is to look for shifts in patterns in discussion. For example, a sudden shift from election to a discussion around the topic of "attacks" and "shooting" could be interesting. These sudden shifts in topics of discussion could indicate that the community is suddenly aware of a new trend or topic, which could be of interest to law enforcement. Thus, by monitoring online discussion forums, establishing baseline measures, and then comparing sudden changes in topics of interest, it should be possible to detect emerging threats. This is already in action in other domains, such as Google Trends [7] or Twitter Hashtag Trends [8] being used to model the emergence of news. In this paper, we evaluate the viability of using a similar approach using Natural Language Processing (NLP) to detect emerging trends within the online right-wing discussion forum StormFront with the eventual goal of creating a near-real-time monitoring system capable of identifying trends in discussion forums, as they emerge, giving law enforcement and security organizations the awareness capability to act as something is unfolding, rather than forensically after-the-fact.

In Section 2, we present a review of previous approaches to researching StormFront and its agenda. Section 3 introduces the method used to achieve the goal of the study. Section 4 provides a discussion on data exploration and data preparation. In Section 5 we review results of the study and finally in Section 6 we conclude the paper.

II. Literature review

Communities like StormFront can be analysed for the purpose of extracting information that help to describe them, to understand their motivations and progression. A traditional avenue of research is the manual gathering and processing of information. Dentice [9] performed a manual approach conducting interviews with representatives of different groups related to right-wing communities and by manual exploration of web resources, including the StormFront forum. The research provided an understanding of right-wing communities in Canada. Another manual approach analysed discussions from StormFront to reveal the understanding of the forum members' behaviour [10].

 ISBN: 978-1-61208-847-1

Manual approaches provide interesting and verifiable results, but the main drawback is the poor scalability in terms of high frequency data and the large volume of posts. Resource allocation provides a limit to manual research for analysing social media post en mass and may hinder the analysis of communities were the number of posts may be in the millions. One way to overcome this is to perform processing based on machine learning of the community data.

In one study, the effects of intergroup conflicts on StormFront and the behaviour of members were analyzed using NLP techniques (including sentiment analysis and word frequency) [11]. Another study described right-wing communities' behaviour using sentiment analysis on data from the StormFront forum [12]. Yet another approach used NLP and support vector machine to classify the StormFront posts depending on the rhetorical change in the posts [13]. However, the results of these studies provide a general description and understanding of right-wing communities; they do not provide the specifics of discussion topics and their temporal shift. We address this gap in our study of right-wing posts and detail a topic modeling approach and trend analysis for StromFront forum discussions.

III. Method

Topic modelling is an NLP task to discover abstract topics in a collection of documents. It assumes that each document in a set of documents (corpus) contains one or more topics that humans can understand; usually, those topics can be described by a set of interrelated keywords.

Latent Semantic Analysis [14] is a statistical approach that studies semantic distributions, relations between sets of documents and words that form concepts. The assumption is that closeness in meaning is related to closeness in textual distance. Latent Dirichlet Allocation (LDA) [15] is a topic modelling method, which is an extension of Probabilistic Latent Semantic Analysis (PLSA) [16]. LDA uses Dirichlet priors for the document-topic and word-topic distributions, which creates an improved generalisability compared to PLSA, i.e., it performs better on new data.

LDA and PLSA are both based on the assumption that each document in a corpus is a mixture over K topics, each topic, in turn, is a mixture over all the words from the dictionary.

LDA is a generative model. Generative models work by first determining the input and output distributions and then uses these distributions to generate additional synthetic input data. Then, the algorithm determines if the generated data represents the given data with an acceptable level of confidence. If the confidence level is not acceptable, then the model generates new data. This process runs in a loop.

In our case, the model generates bag-of-words distribution using topics per document distribution Θ and words per topic distribution φ. Θ and φ, in turn, are generated from a Dirichlet distribution with the parameters α and β.

Figure 1 shows the graphical representation of this process [15], where Z is topic-word and W word-document identities for a particular word in a document. M and N are a number of documents and a number of words in a particular document, respectively.

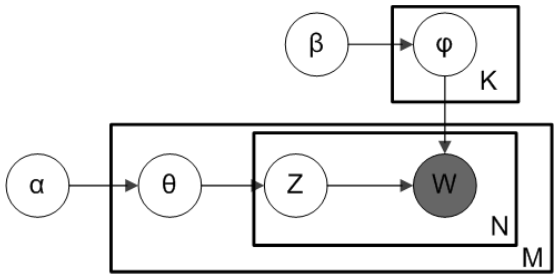

Figure 1. Graphical representation (plate notation) for LDA [15].

The next step is the estimation, whether the generative process has generated data close enough to the given data. There are various ways of performing this task. The LDA implementation, which we used, does it with the variational Bayes algorithm.

LDA takes a "bag of words" text representation as an input, which is a drawback of LDA because a large vocabulary grows the dimensionality, and as a result the computational performance of the method. LDA takes at minimum one hyperparameter, the number of topics. The optimal number of topics can be determined with perplexity as a metric. Perplexity is a measurement of how good a probability distribution predicts a sample.

IV. Data

The goal of this paper is to identify emerging trends within online right-wing communities to detect possible new events and/or problems. A number of different right-wing online forums exist online, and for this paper a specific right-wing site, the online community StormFront, was chosen.

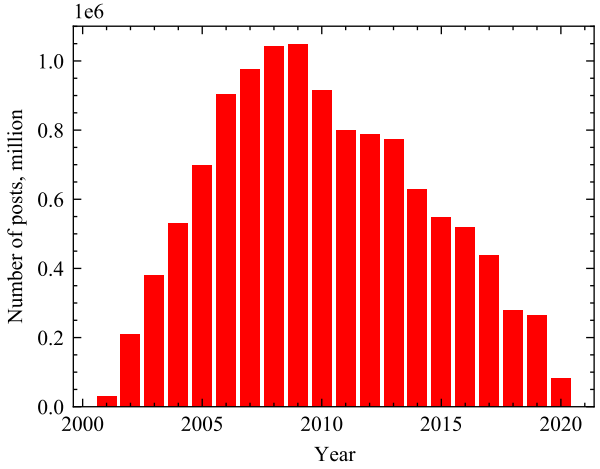

Figure 2. Distribution of posts by year.

The forum is the first dedicated to, and one of the most popular, right-wing online discussion forums [17]. The posts from the forum were collected using The Dark Crawler [18],

TABLE I
TOP TEN LANGUAGES IN THE FILTERED DATASET

Language	Number of posts
English	2,036,430
Italian	25,020
French	15,275
Spanish	14,192
Serbian	12,309
Portuguese	12,187
Croatian	5,091
German	3,187
Russian	2,111
Serbo-Croatian	943

and contain approximately 12 million posts spread over 1 million threads by almost 360,000 users. The full dataset consists of posts from the Stormfront forum from the 28th of August of 2001 to the 29th of April 2020. The dataset captured contains all thread and post information, containing post-date, author of the post, post content, thread and sub-forum information. Figure 2 illustrates the distribution of posts by year.

In this study, only data from 2015 onwards were considered. All posts containing only links, emojis, and those with less than five words were excluded since they do not tend to represent any topic but instead provide sentiments. Furthermore, the dataset contained posts created in several languages. Various languages in a corpus can confuse almost all the vocabulary-based NLP models because each language significantly enlarges a corpus's vocabulary. Therefore for each post from the resulting dataset, the language was identified using the fasttext model [19] [20]. Table I shows the top ten languages and corresponding numbers of posts written in each language. Only posts in English were used, resulting in a final total post count of 2,036,430.

V. RESULTS

A. Determining the number of topics

One of the hyperparameters for Latent Dirichlet Allocation is the number of topics. To find the best value, the perplexity metric was used. Perplexity for a topic modelling model shows an ability to generalise the results, that is, it shows how a model performs if it were to be applied to data, which was not used for training. For the experiment, the data were split into training and testing data sets, 75% and 25% respectively. Then LDA was trained on the training data, and applied to the test data with the following numbers of topics 5, 10, 30, 40, 50, 70, 100 and 200, we calculated perplexity for each of these values on the test set. The lowest perplexity value was for 40 topics, and thus was the value that was used in the following analysis.

B. Topics between January 2015 and April 2020

Once the model is trained and the hyperparameter value was selected, an exploratory analysis was performed to get preliminary insights from the matrix of the probability distribution of topics for each document. Each value in this matrix is the probability that the document belongs to the topic. Figure 3 demonstrates the average probability for each topic. The top three values of the average probability are for topics 30, 33 and 38, meaning that more posts tended to belong to these topics. Consequently, these topics are discussed more often than others in the forum. Table II shows the top fifteen keywords for these topics.

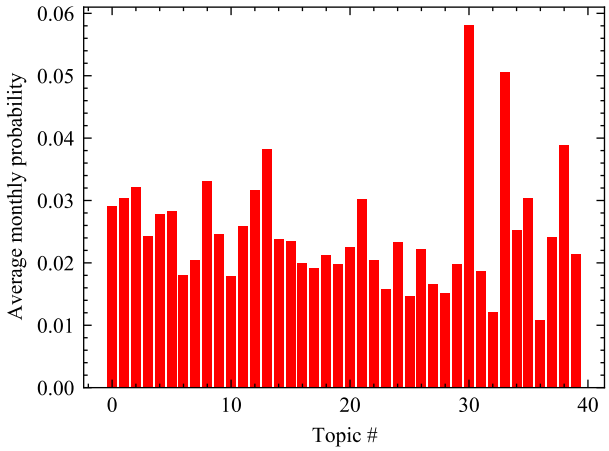

Figure 3. Average monthly topic probabilities.

TABLE II
TOP FIFTEEN KEYWORDS FOR TOPICS 30, 33 AND 38

Topic	Keywords
30	guy thing didn yeah lol thought negro crap guess hell white gay sick big funny
33	forum stormfront site find posting link comments didn google doesn internet article made lot comment
38	political party movement nationalist left support hope years real wn public members media politics change

The general picture about topic distribution showed how topics are distributed against each other, we wanted to identify individual topics and determine if they can be used to detect events, which happened from 2015 to 2020. Thus, monthly average probabilities were calculated for each topic. It was found that, as a way of presenting the level of discussion of a particular topic, if the average probability of a topic is higher in one period than in another then that topic was discussed more during that period. In other words, posts had a higher chance of belonging to this topic than to others; consequently it means that the topic appeared more frequently in this period. Figure 4 shows the comparison monthly average probabilities of the most discussed topic (number 30) and the least discussed (number 36). As shown on the plots, one line is higher than the other in all the points, meaning that topic 30 was always discussed more than topic 36 since 2015. Words on the top of each plot are the top keywords for the topic.

For example, a randomly selected post from topic 30 stated:

 ISBN: 978-1-61208-847-1

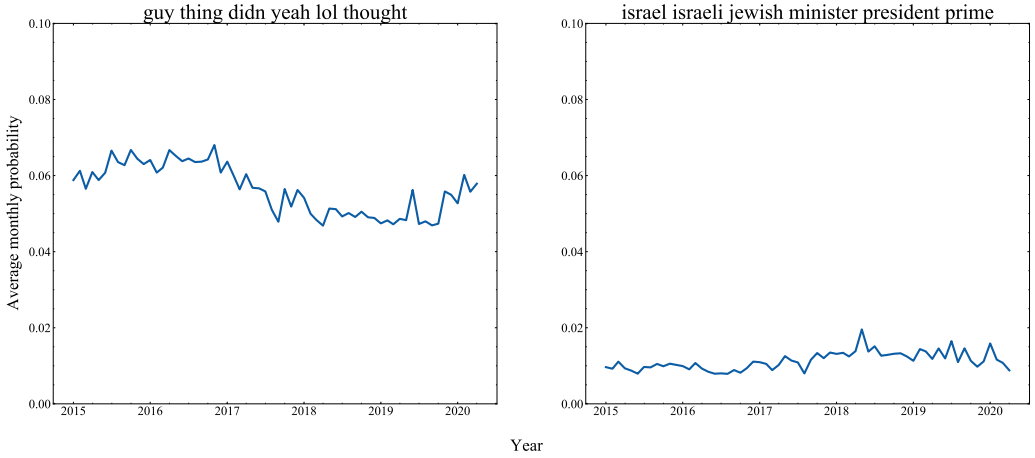

Figure 4. Average monthly topic probabilities for topics 30 and 36.

I was on my way back from the gym today when I could see two negro boys walking towards me. I knew, but absolutely knew, they were going to say something to me. As soon as they get within five feet of me they start grunting and looking at my shirt and reading the letters on it out loud. What is it with Negro kids? Whites never seem to do that kind of thing. It's like the Negro can't let anyone walk in peace. They HAVE to be noticed or affirmed. Even if it means embarrassing a complete stranger.

While a randomly selected post from topic 36 stated:

Quote: Bolivia plans to ask for a public meeting of the United Nations Security Council after U.S. President Donald Trump announces on Wednesday that the United States recognizes occupied Al-Quds (Jerusalem) as the capital of the Zionist entity and will move its embassy there. Bolivian U.N. Ambassador Sacha Sergio Llorentty Soliz said it would be a "reckless and a dangerous decision that goes against international law, the resolutions of the Security Council, it also weakens any effort for peace in the region and also upsets the whole region." Bolivia to Seek UN Security Council Meeting on Al-Quds (Jerusalem) Status – Al-Manar TV Lebanon

The probabilities of some of the topics have spikes, which indicate that it is likely that some disruption occurred around that time, reflecting the events in the real world. For example, Figures 5 and 6 show spikes in the second half of 2016, and it is likely that both topics are related to the 2016 United States presidential election [21], meaning that in the second half of 2016 the election-related these topics were being discussed a lot. This is a completely expected result, which validates the methods presented here.

Topic 16 contained a series of health-related keywords (ex: drug, medical, disease). Throughout 2015-2020, this topic was relatively consistent, and a relatively minor topic. However, at the beginning of 2020, discussions around this topic increased five-fold (Figure 7), exactly coinciding with the emergence of the COVID-19 pandemic [22].

In August 2017, the Stormfront website was banned by

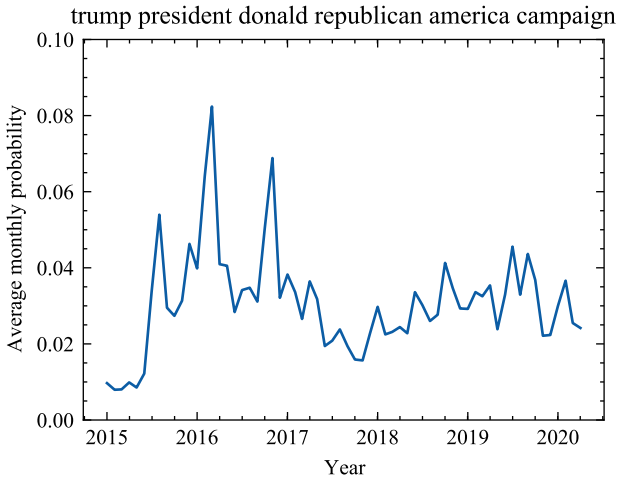

Figure 5. Average monthly topic probabilities for topic 2

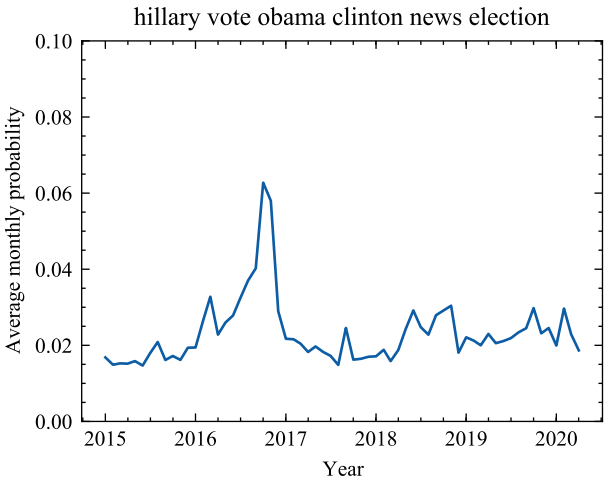

Figure 6. Average monthly topic probabilities for topic 24

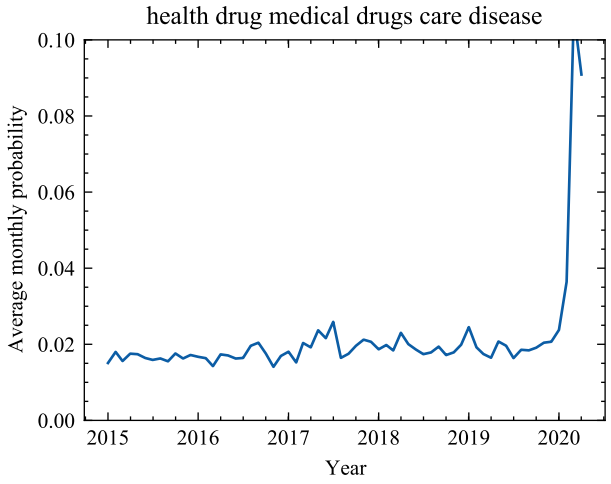

Figure 7. Average monthly topic probabilities for topic 16

its registrar. Figure 8 demonstrates the forum reaction to the event, where a very strong spike can be seen for the topic related to discussions around keywords “forum, Stormfront, site” in August 2017 [23].

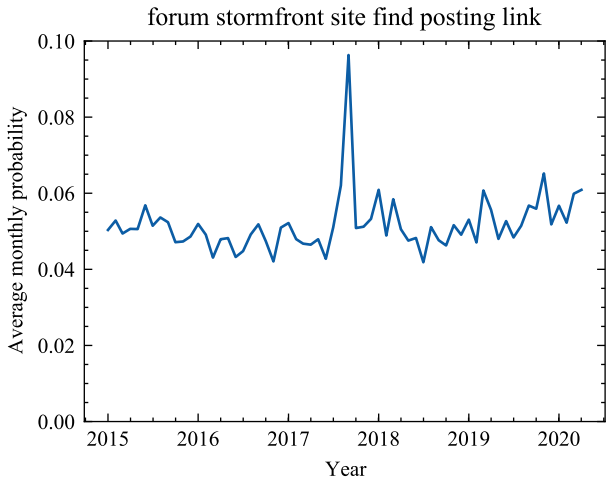

Figure 8. Average monthly topic probabilities for topic 33

Several other topics also had spikes in August 2017; therefore, they were analysed in more detail to check if there is a relationship among them. Figures 9 and 10 show the weekly data for 2017 for two topics; both of which increased in August 2017. The spike for the topic with keywords ”car shot dead fight” happens around week 33 (corresponding to the week of August 13, 2017) and most probably reflects discussions around Barcelona’s crowd rammings, which occurred August 17, 2017 [24]. Another topic with keywords ”hate speech media university” has posts related to the Unite the Right rally in Charlottesville, Virginia in August 2017 [25] and the Annual Stormfront Smoky Mountain Summit in September 2017. All the spikes in August-September 2017 have different natures, and thus it was concluded that the spikes are not related to each other.

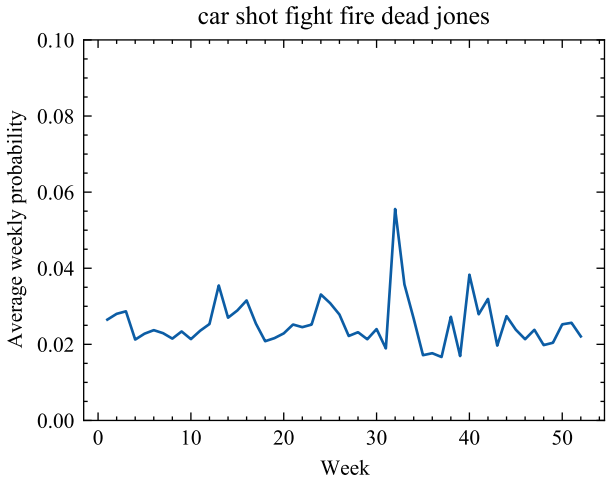

Figure 9. Average weekly topic probabilities for topic 3 in 2017

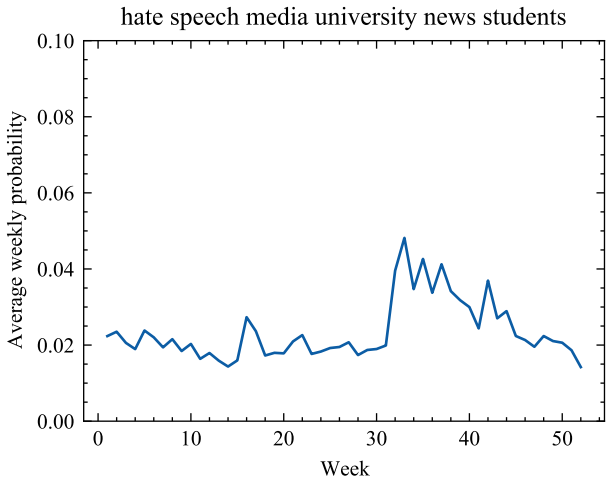

Figure 10. Average weekly topic probabilities for topic 19 in 2017

C. Weekly topics in 2019

The analysis of long periods is essential because it shows the general patterns, but the detection of smaller-scale events would allow security organizations to potentially detect quickly emerging threats or problems. Therefore data from January 2019 to May 2020 (non-inclusive) was selected, and the same trained model was applied. When all the topics were returned, the weekly average probabilities for each topic were calculated.

Figure 11 demonstrates the topic’s results with keywords ”military syria isis iran security” the peak in January 2020 coincides with a United States drone strike near Baghdad International Airport, resulting in an assassination of the major general Qasem Soleimani [26].

Figure 12 shows the topic with top keywords ”trump president donald republican america campaign”, the sudden

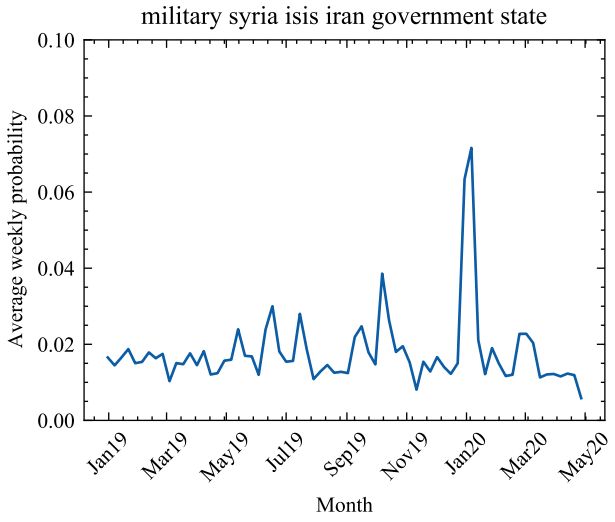

Figure 11. Average weekly topic probabilities for topic 23 from Jan 2019 to Apr 2020

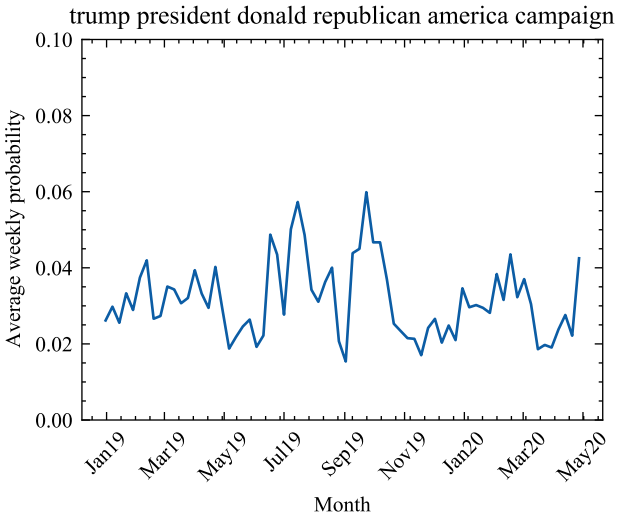

Figure 12. Average weekly topic probabilities for topic 2 from Jan 2019 to Apr 2020

increase in June 2019 is likely related with Trump's launch of his 2020 re-election campaign in Orlando, Florida [27]. Moreover, two peaks, at the beginning of July 2019 and end of September, coincide with his July 4th speech and with the 74th Session of the United Nations General Assembly, respectively.

VI. Conclusion

The models' and further analysis' results demonstrate that there is a correlation between real events and topics discussed on the forum; however, correlation does not imply causation, and thus we do not claim that the topics within the discussions lead to actions. We can see what specific topics have peaks at the time when something happens.

The distribution of topics (the vector of average probabilities of each topic within a certain period) can represent a snapshot of the community agenda, which in turn shows a general picture of the community members' attention. The model and further analysis of its results can show how the community reacts to the events - if it is a brief blimp, quickly reverting to average, then it disappeared from the agenda pretty quickly, see Figure 8, or if there is a sudden surge gradually decreasing to average after the event then it was an impactful event of great interest within the community, see Figure 5.

However, the proposed work has several areas of improvement. First of all, it is based on a bag-of-words representation of text, it works well for English because it is an analytical language, but if the intention is to use the model with more synthetic languages, for example Finnish, then changes are required, which help to reduce vocabulary of the corpus (for example, lemmatisation). Further, the model should be retrained from time to time to stay relevant, and finally, the model depends on some random processes so results can differ from run to run.

In conclusion, we can claim that anyone who analyses large-scale communities like StormFront - researchers, law enforcement, and security organisations - can use the proposed methods to detect changes in discussion topics, and detect new topics of interest.

References

[1] M. Conway, "Terrorist 'use' of the internet and fighting back," Information & Security: An International Journal, vol. 19, 01 2006, pp. 9–30.

[2] White homicide worldwide. Southern Poverty Law Center. 400 Washington Avenue, Montgomery, AL 36104. Apr. 1, 2014. [Online]. Available: https://www.splcenter.org/20140331/white-homicide-worldwide [retrieved: March, 2021]

[3] F. Bajak. Online providers knock 8chan offline after mass shooting. ABC News. Aug. 6, 2019. [Online]. Available: https://abcnews.go.com/Technology/wireStory/security-cut-off-cesspool-hate-8chan-forum-64778026 [retrieved: March, 2021]

[4] P. Gill, E. Corner, A. Thornton, and M. Conway, What are the roles of the Internet in terrorism? Measuring online behaviours of convicted UK terrorists. VOX Pol, 2015.

[5] P. Gill, E. Corner, M. Conway, A. Thornton, M. Bloom et al., "Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes," Criminology & Public Policy, vol. 16, no. 1, 2017, pp. 99–117.

[6] Homegrown violent extremist mobilization indicators 2019 edition. Federal Bureau of Investigation. Washington, DC: Office of the Director of National Intelligence. [Online]. Available: https://www.dni.gov/index.php/nctc-newsroom/nctc-resources/item/1945-homegrown-violent-extremist-mobilization-indicators-2019 [retrieved: March, 2021]

[7] H. Choi and H. Varian, "Predicting the present with google trends," Economic record, vol. 88, 2012, pp. 2–9.

[8] R. Lu and Q. Yang, "Trend analysis of news topics on twitter," International Journal of Machine Learning and Computing, vol. 2, no. 3, 2012, p. 327.

[9] B. Perry and R. Scrivens, "A climate for hate? an exploration of the right-wing extremist landscape in canada," Critical Criminology, vol. 26, no. 2, 2018, pp. 169–187.

[10] D. Dentice, "'so much for darwin': an analysis of stormfront discussions on race." Journal of Hate Studies, vol. 15, no. 1, 2019, pp. 133–156.

[11] A.-M. Bliuc, J. Betts, M. Vergani, M. Iqbal, and K. Dunn, "Collective identity changes in far-right online communities: The role of offline intergroup conflict," New media & society, vol. 21, no. 8, 2019, pp. 1770–1786.

[12] R. Scrivens, "Exploring radical right-wing posting behaviors online," Deviant Behavior, 2020, pp. 1–15.

 ISBN: 978-1-61208-847-1

[13] R. Omizo, “Machine learning approaches to understanding white supremacy online,” Rhet Ops: Rhetoric and Information Warfare (Eds. Jim Ridolfo and William Hart-Davidson), 2019, pp. 142–157.
[14] T. K. Landauer and S. T. Dumais, “A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” Psychological review, vol. 104, no. 2, 1997, p. 211.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, 2003, pp. 993–1022.
[16] T. Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50–57.
[17] G. Bolaffi, R. Bracalenti, P. Braham, and S. Gindro, Dictionary of Race, Ethnicity & Culture. SAGE Publications, 2003, p. 254.
[18] R. Scrivens, T. Gaudette, G. Davies, and R. Frank, “Searching for extremist content online using the dark crawler and sentiment analysis,” in Methods of Criminology and Criminal Justice Research. Emerald Publishing Limited, 2019.
[19] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” arXiv preprint arXiv:1607.01759, 2016.
[20] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou et al., “Fasttext.zip: Compressing text classification models,” arXiv preprint arXiv:1612.03651, 2016.
[21] D. C. Beckwith. United states presidential election of 2016. Encyclopedia Britannica. Nov. 11, 2019. [Online]. Available: https://www.britannica.com/topic/United-States-presidential-election-of-2016 [retrieved: March, 2021]
[22] D. B. Taylor. A timeline of the coronavirus pandemic. The New York Times Company. Jan. 10, 2021. [Online]. Available: https://www.nytimes.com/article/coronavirus-timeline.html [retrieved: March, 2021]
[23] J. Biggs. Another neo-nazi site, stormfront, is shut down. Verizon Media, Techcrunch. Aug. 28, 2017. [Online]. Available: https://techcrunch.com/2017/08/28/another-neo-nazi-site-stormfront-is-shut-down/ [retrieved: March, 2021]
[24] A. Smith, A. Jamieson, K. Simmons, and K. Itasaka. Spain terror: American among 14 killed in van and car attacks. NBC News. Aug. 18, 2017. [Online]. Available: https://www.nbcnews.com/news/world/van-hits-pedestrians-barcelona-n793506 [retrieved: March, 2021]
[25] H. Spencer and S. G. Stolberg. White nationalists march on university of virginia. The New York Times Company. Aug. 11, 2017. [Online]. Available: https://www.nytimes.com/2017/08/11/us/white-nationalists-rally-charlottesville-virginia.html [retrieved: March, 2021]
[26] M. Crowley, F. Hassan, and E. Schmitt. U.s. strike in iraq kills qassim suleimani, commander of iranian forces. The New York Times Company. Jan. 2, 2020. [Online]. Available: https://www.nytimes.com/2020/01/02/world/middleeast/qassem-soleimani-iraq-iran-attack.html [retrieved: March, 2021]
[27] M. Haberman, A. Karni, and M. D. Shear. Trump, at rally in florida, kicks off his 2020 re-election bid. The New York Times Company. Jun. 18, 2019. [Online]. Available: https://www.nytimes.com/2019/06/18/us/politics/donald-trump-rally-orlando.html [retrieved: March, 2021]

 ISBN: 978-1-61208-847-1