Topic Modeling of StormFront Forum Posts

Grigorii Nazarko, Richard Frank, Magnus Westerlund

Authors

Grigorii Nazarko

Master student, program "Big Data Analytics", Arcada University of Applied Sciences, Finland

Dr. Richard Frank

Associate Professor in the School of Criminology at Simon Fraser University (SFU), Canada and Director of the International CyberCrime Research Centre (ICCRC)

Dr. Magnus Westerlund

Deputy Head Of Department, Business and Analytics, Arcada University of Applied Sciences, Finland

Agenda

- Motivation & Previous studies
- Data & Method
- Results & Conclusion

Motivation

Radicalised communities use the internet to preach their ideas within society.

- Christchurch mosque shootings in 2019, live-streaming on Facebook and manifesto on "8chan"
- Terrorists who plotted to attack government targets 83% "displayed traits of online learning"
- Users of improvised explosive devices much more likely found the information on online resources than on offline ones

This point leads us to think that radicalised communities' online resources should be monitored to prevent violence.

Motivation

As an indicator of monitoring of those online resources, we suggest the aggregated set of community thoughts, represented as agenda.

- Sudden change in agenda can represent changes in a community's mood and support the decisions of law enforcement actions
- Agenda can be valuable by itself as a presentation of what occupies minds of community's members

In this study, we present the ability to use Natural Language Processing to research the agenda and its changes of StormFront - the oldest online right-wing discussion forum.

Previous studies

Previous studies of StormFrom's agenda monitoring mostly offered two approaches:

- Manual research, which provides thorough results, but limited in terms of resources
- Sentiment analysis, which is valuable for sentiment research, but does not provide much information about agenda's topics

We argue that topic modelling can be used to represent the agenda of StormFront, because this is an automated approach, which provides the information about discussed topics.

Method

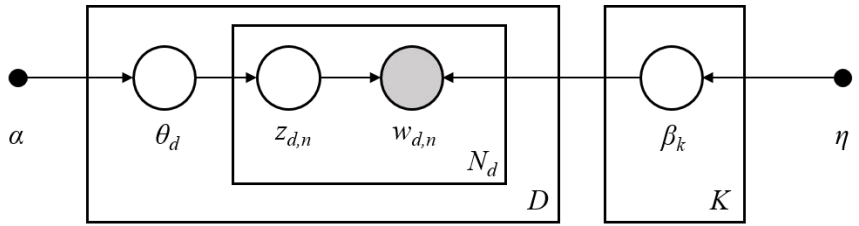
We used Latent Dirichlet Allocation (LDA) for topic modelling. It has several features:

- LDA takes a bag-of-words representation of documents, which is one of the drawbacks of the model. Bag-of-words can be heavy-tailed, for example, if corpus consists of several languages or one synthetic language.
- LDA returns the probabilities of each document to belong to each topic. This feature was precious because the average of these probabilities can represent the agenda for a certain period

To find the number of topics we used perplexity as a metric.

Method

LDA is a generative model. It generates a distribution and identifies if generated distribution represents input. Here is a plate notation of LDA process:



D, K and N - numbers of documents, topics and words, respectively.

$\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ - priors from Dirichlet distribution

$\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ - parameters for Dirichlet distributions

Z and W - generated topic-document and document-words distributions

Data

The initial dataset was collected using The Dark Crawler and contained all the posts from StormFront since 2000. Then, we cleaned and filtered the dataset, so eventually, it contained:

- Posts from the beginning of 2015 to the end of April 2020, as the most recent and relevant ones.
- Posts longer than five words, because short posts tend to provide sentiments rather than represent any topics.
- Posts in English, which is 95% of all the posts, because other languages confuse the model, which is very sensitive to the vocabulary size

The final total number of posts is 2,036,430

Results

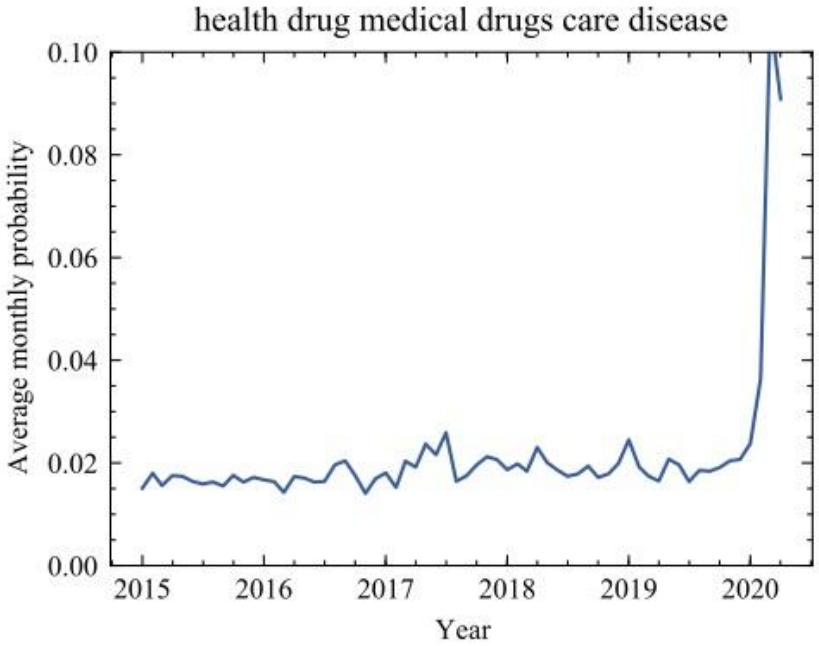
The results were presented as topic distribution for each document and keywords for each topic.

Example of the keywords for the most discussed topics:

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

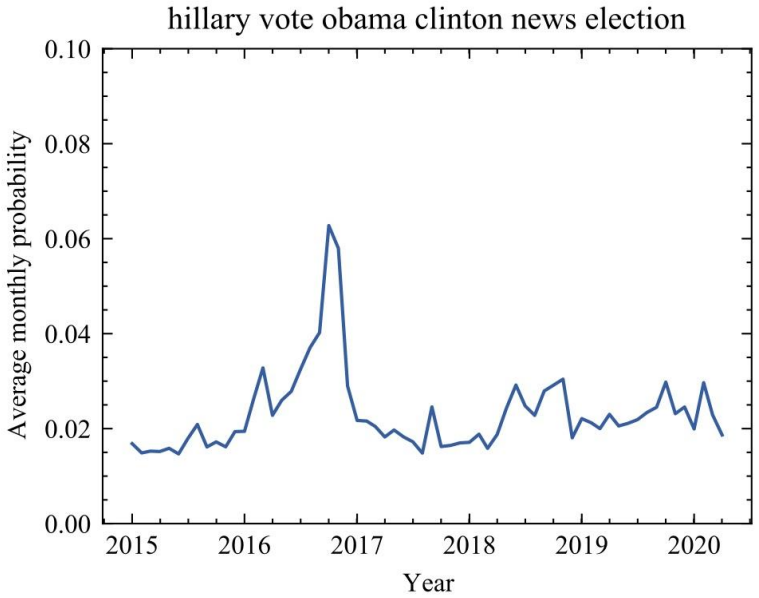
Results

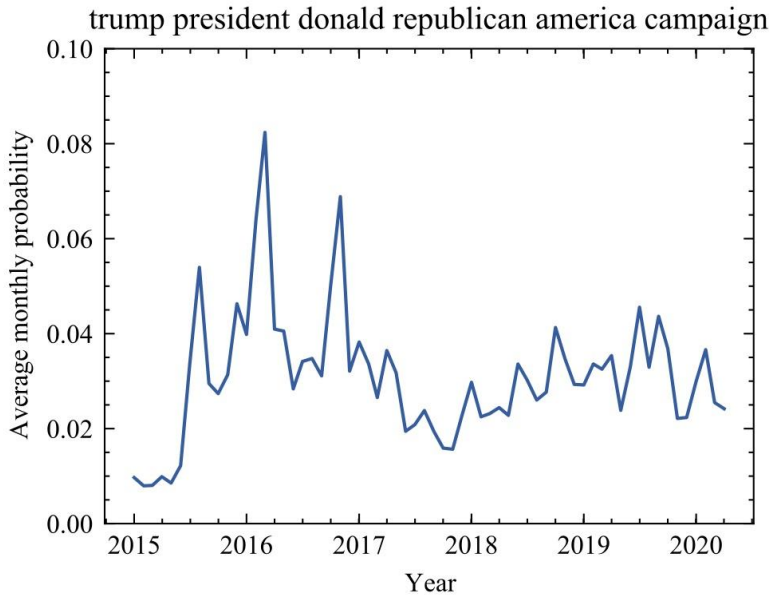
For each topic we calculated average probabilities across posts to belong to this topic for certain periods. Then we matched these probabilities to real-world events to validate the method. As an example, this chart shows monthly average probabilities for a topic with keywords "health drug medical drugs care disease", we can connect sudden growth at the beginning of 2020 to the COVID-19 pandemic.



Results

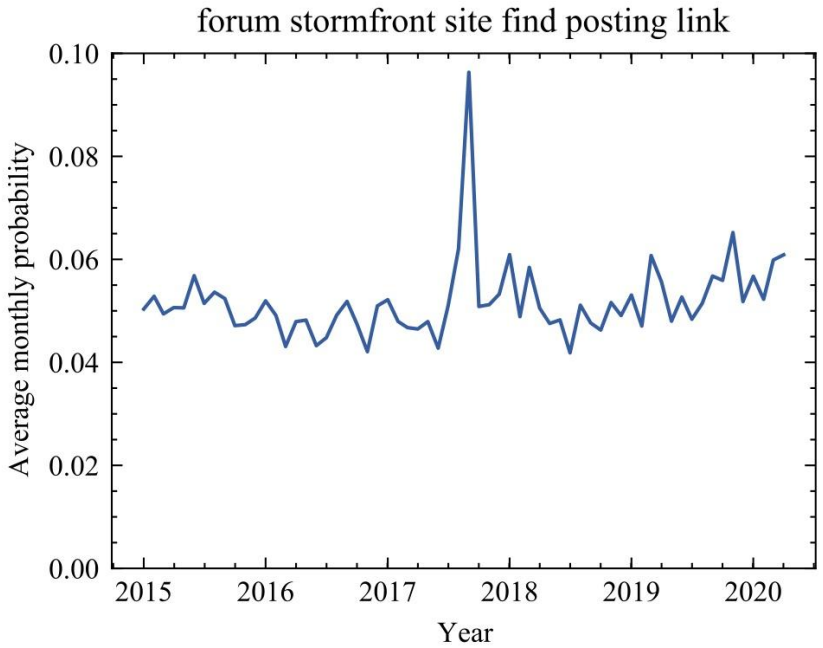
Topics with keywords related to the US election have spikes in the second half of 2016, which coincides with the 2016 United States presidential election.





Results

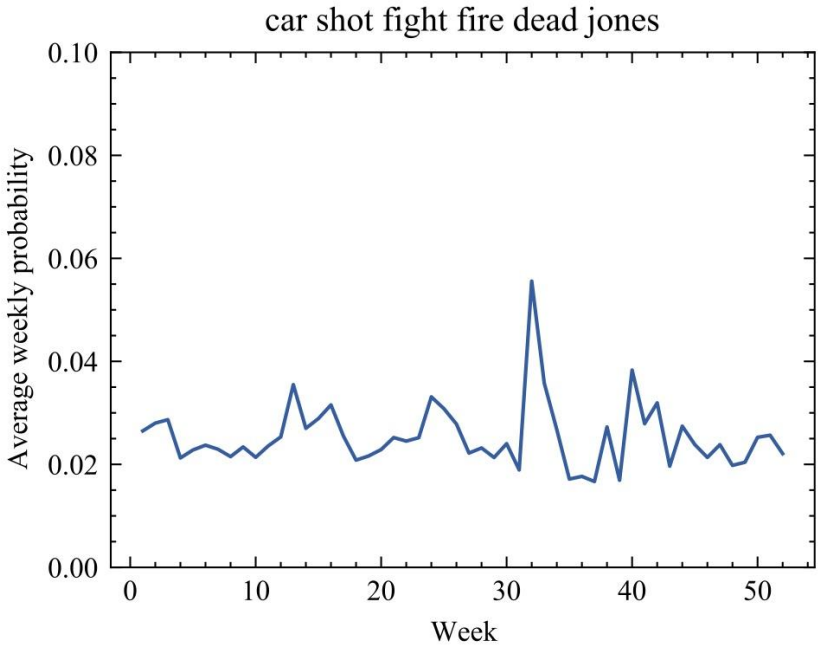
Another example is a growth of discussions on the topic with keywords "forum stormfront site find posting link" in the middle of 2017. This growth can be related to the ban of StormFront by the domain registrar in August 2017.



Results

In the middle of 2017 several other topics had spikes as well; for investigation, we calculated weekly average probabilities for those topics for 2017.

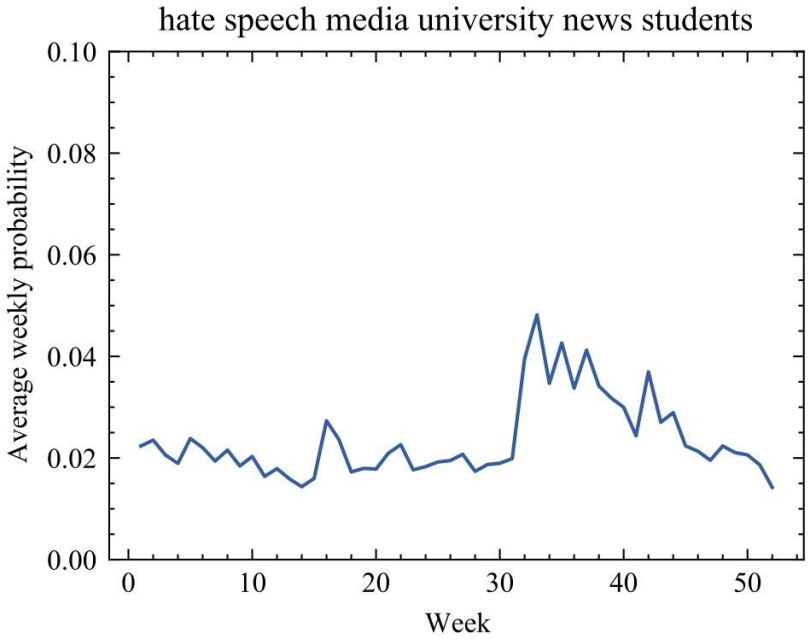
One of those topics is presented in the figure. The spike is most probably related to Barcelona's crowd rammings in August 2017 (week 33).



Results

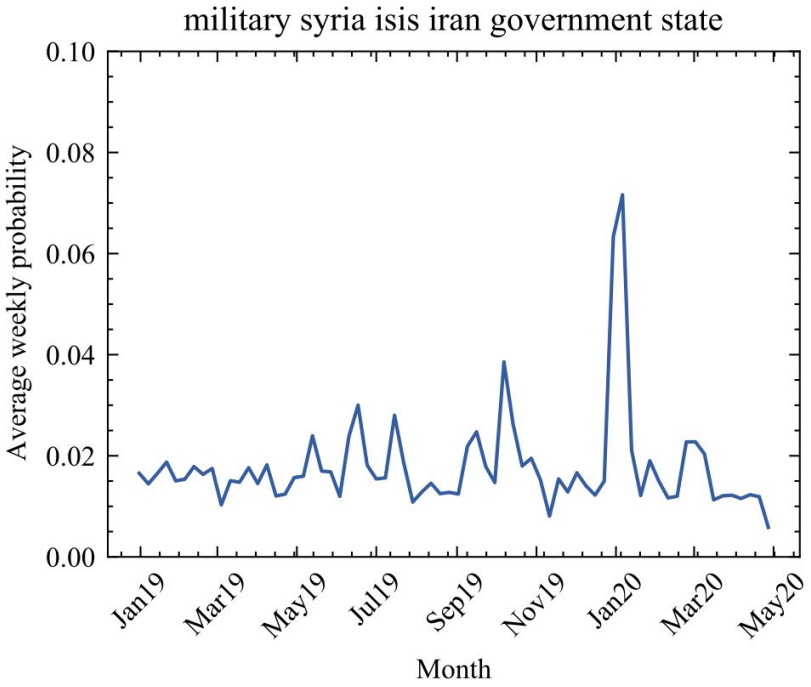
Another topic which presented spike in August 2017 has keywords "hate speech media university news students", most probably connected to to the Unite the Right rally in Charlottesville, Virginia in August 2017 and the Annual Stormfront Smoky Mountain Summit in September 2017.

Spikes in all the topics in August 2017 have different nature and we could not connect them together.



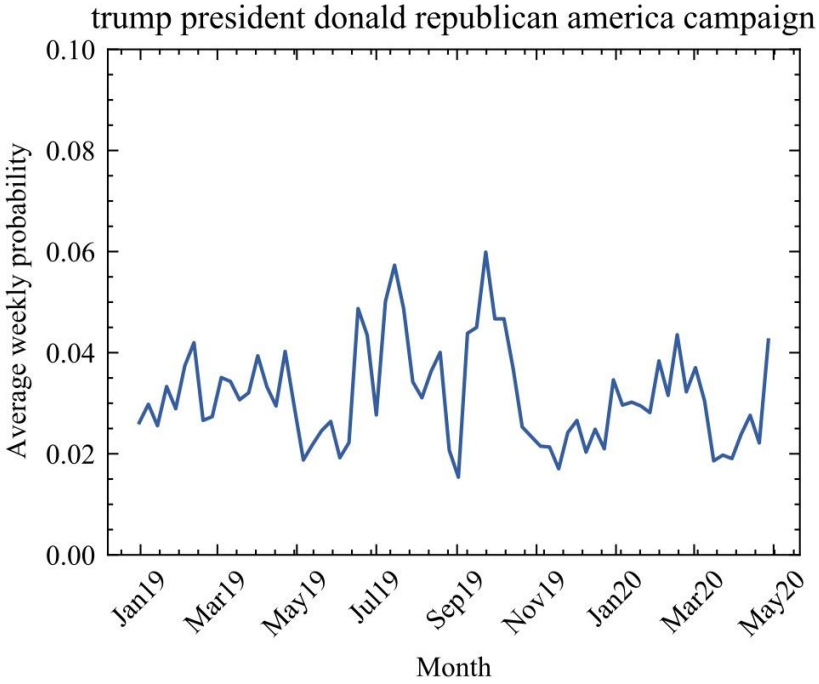
Results

Alongside we decided to validate how the method works on shorter scales, we calculated weekly average probabilities across posts. This chart shows such probabilities from 2019 for topic "military syria isis iran government state". The spike at the beginning of 2020 coincides with a US drone strike near Baghdad and an assassination of the major general Qasem Soleimani.



Results

On the same timescale, the topic, which has "trump" and "donald" as keywords, shows sudden increase and then two peaks. The increase, in June 2019, is likely related to Donald Trump's launch of his 2020 re-election campaign in Orlando, Florida. The peaks in July and September 2020 coincide with his July 4th speech and with the 74th Session of the United Nations General Assembly.



Conclusion

- The presented results show the StormFront agenda and its changes in connection with some events
- We can claim that anyone who analyses large-scale communities like StormFront - researchers, law enforcement, and security organisations can use the presented method to monitor communities' agendas
- One of the further improvement is scaling the method to other languages