



Topic Modeling of StormFront Forum Posts

Grigorii Nazarko

Master's Thesis
Master of Engineering - Big Data Analytics
June 2, 2021

MASTER'S THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Master of Engineering - Big Data Analytics
Identification number:	8276
Author:	Grigorii Nazarko
Title:	Topic Modeling of StormFront Forum Posts
Supervisor (Arcada):	Magnus Westerlund
Commissioned by:	Arcada University of Applied Sciences
<p>Abstract:</p> <p>The research of radical communities is crucial for preventing violent actions and affecting the community to avoid further radicalisation. In this thesis, we propose a way to analyse semantic topics which was assessed on the oldest right-wing forum StormFront.</p> <p>We obtained two million forum posts from 2015 to 2020 and applied several NLP techniques for topic modelling. The model that provided the best results was Latent Dirichlet Allocation (LDA). We used human experts, who estimated the connection between real-world events and the model's output to validate the value and sensibility of the framework. The validation showed that the framework is correct and valuable for analysing topics and affiliated discussions in StormFront.</p> <p>The thesis consists of two parts: formal thesis and associated conference paper (Nazarko, Frank & Westerlund 2021).</p>	
Keywords:	topic model; information extraction; NLP; LDA
Number of pages:	29
Language:	English
Date of acceptance:	25.05.2021

CONTENTS

1	Introduction	7
1.1	Background	7
1.2	Motivation	7
1.3	Data	8
1.4	Limitations	9
1.5	Ethical considerations	9
1.6	Author Contribution	9
2	Framework	10
2.1	Target group and value proposition	10
2.2	Frameworks review	10
2.3	Framework explanation	11
2.3.1	<i>Exploratory data analysis and models research</i>	11
2.3.2	<i>Modelling</i>	12
2.3.3	<i>Results analysis</i>	13
2.4	Results usage	14
3	Methods review	15
3.1	Latent Dirichlet Allocation	15
3.2	Embedding topic modelling	16
3.3	Top2vec	17
3.4	Method selection	17
3.5	Hyperparameters search	18
4	Validation	20
5	Results	23
5.1	Validation results	23
5.1.1	<i>Five years scale</i>	23
5.1.2	<i>One year scale</i>	25
5.2	Results usage	26
6	Conclusion	27
	References	28

FIGURES

Figure 1.	The sample representation of the most and the least discussed topics . . .	13
Figure 2.	Graphical representation (plate notation) for LDA.	16
Figure 3.	Average monthly topic probabilities for topic 24	22
Figure 4.	Average monthly topic probabilities for topic 2	23
Figure 5.	Average monthly topic probabilities for topic 33	24
Figure 6.	Average monthly topic probabilities for topic 16	24
Figure 7.	Average weekly topic probabilities for topic 23	25

TABLES

Table 1.	Sample topics with keywords from ETM outcome	18
----------	--	----

ABBREVIATIONS

NLP Natural Language Processing

LDA Latent Dirichlet Allocation

ETM Embedded Topic Modelling

FOREWORD

I want to thank Magnus Westerlund for all the help with this research. Many thanks to Richard Frank and Simon Fraser University for providing us with the data and participation in the research; Leonardo Espinosa-Leal, Anton Akusok, Lisa Staudinger and Andrey Scherbakov-Parland for all the knowledge and fun I got during my studies at Arcada University of Applied Sciences.

1 INTRODUCTION

1.1 Background

Communities play a significant role in society; people are integrating into different groups based on the same interests. Although those interests and communities are useful in most cases because they can help to solve common problems and make society more diverse, sometimes they are not so helpful, harmless and safe. Some communities can be angry against certain groups of people or governments, which converts into a problem when participants of those communities become more radical and start acting violently.

A recent example of such violent actions is storming of United States Capitol on the 6th of January 2020. Information about the rally and riot was distributed via social networks among right-wing communities. As a result, the rioters came into the Capitol during the joint session of Congress. Another example of the community's influence is mass shootings in mosques in New Zealand in 2019. The shooter live-streamed the shooting via Facebook and announced the manifesto and his plans on right-wing social network "8chan". A lot of people were killed and injured as a result of the attack (Bajak 2019).

Human casualties are not the only consequences of such actions but tension in society and escalation of radicalisation of different communities. Downscaling the tension, as well as human casualties, cost society a lot of resources. The possibility of such a negative impact on society implies that radical communities should be constantly monitored by law enforcement and security organisations (Nazarko, Frank & Westerlund 2021).

1.2 Motivation

Online resources like forums and social networks increasingly impact the mobilisation of violence in extremists groups (Gill, Corner, Conway, Thornton, Bloom & Horgan 2017). These resources look like the best place for communities monitoring because people interact there with each other, and these interactions can be tracked and analysed to predict and prevent dangerous behaviour.

Security organisations are developing different indicators to predict imminent dangers

analysing the communities behaviour. These indicators are related to personal users' actions, such as increasing participation in extremist discussions or posting content associated with extremist topics (2019). Although the personal indicators are valuable, it makes a lot of sense also to know the general ideas discussed in the communities, which occupies the community's members' minds at the moment. We call this momentary general picture as agenda (Nazarko, Frank & Westerlund 2021).

A community's agenda before such events can help to understand what preceded the violent act to take actions if the pattern appears again. For instance, if right before a terrorist act in a mosque, the topic "cars" prevailed in the agenda and then suddenly changed to "muslims". Then, next time such a sudden change happens, security organisations could strengthen security of mosques and prevent incidents.

On the other hand, the community's agenda after such events shows how the community reacts to the events; this can help to understand what actions should be taken for de-escalation. For example, government could make a marketing campaign with a certain community as a target group that needs information about its current general discussions. Furthermore, the manual research of communities' agendas in social media, requires many resources; a limited group of researchers should read and analyse the content produced by the massive community. Automatisation with the usage of machine learning techniques should decrease the cost of such studies. This research aims to find a way to analyse communities' agendas and their changes with machine learning techniques.

1.3 Data

The analysis of machine learning techniques requires data for comparison of models and validation of results. Our research's ideal data is text data from a community social network because it has the same nature and origin as the other data that can be used as an input to the method we present as an outcome of this research.

We used the data from StormFront. It is one of the most famous and oldest right-wing extremists' forum. The data was collected with the Dark Crawler (Scrivens, Gaudette, Davies & Frank 2019) and contained posts from the forum from the beginning of 2000 to

the May of 2020. Although we proved the method with StormFront data, it can be applied to almost any social media where any group of people discuss anything.

1.4 Limitations

The study is limited by the given data and some objective limitations of the considered models. For example, to provide the best performance of certain models, we selected only data in English. Also, we limited the volume of data and selected the most recent ones as the most relevant.

1.5 Ethical considerations

A possible ethical issue with the method itself is that the method can probably be used for unscrupulous purposes. For example, people who want to affect a community's behaviour can use this method to define actions based on the agenda and monitor the community feedback on these actions. Regarding the research, although some topics extracted with the method and explained in the results chapter can be very sensitive, there is no ethical problem because these topics are part of history and were actually discussed in the community.

1.6 Author Contribution

This thesis describes motivation and process of the research resulting in a conference paper written by Grigorii Nazarko, Richard Frank, and Magnus Westerlund, with the title "Topic Modeling of StormFront Posts" and published at the conference COGNITIVE 2021 (Nazarko, Frank & Westerlund 2021).

The publication deals with the topic modelling of posts from Stormfront. The original idea of topic modelling belongs to Richard Frank, who was also primarily responsible for the criminology aspects and implications of the study and acquiring the data. The methodology related to data preprocessing, model design, validation and visualization of the results was mainly the responsibility of the author of the thesis and Magnus Westerlund. The code implementation and computations were developed by the author of the thesis.

2 FRAMEWORK

2.1 Target group and value proposition

Various stakeholders can use the method of researching communities agendas, such as law enforcement and security organisations, various government institutions related to society and academic researchers, marketing specialists. Any individuals or organisations, which have a goal to research and affect communities.

We argue that we can provide a machine learning method to research agendas of communities and their changes. Agenda and its changes help to explore patterns in community discussions and feedback to irritation (Nazarko, Frank & Westerlund 2021). This information is beneficial as a support for making decisions that heavily connected with the community. For example, a decision about increase of vigilance or which content should be added to the advertisement.

2.2 Frameworks review

In this project's frame, we studied how researchers analysed communities' agendas in previous studies. For instance, this study showed the ability to use natural language processing for projecting posts to topic space which can be considered as an agenda (Rohani, Shayaa & Babanejaddehaki 2016). Another study presents the approach with topic modelling and critical discourse analysis to investigate patterns in discussions related to Muslims in a large Swedish forum (Törnberg & Törnberg 2016). This research used natural language processing for analysing tweets around COVID-19 pandemic (Mutanga & Abayomi 2020).

Another approach was used to detect emerging of news with Google Trends (Choi & Varian 2012) and Twitter Hashtag Trends (Lu & Yang 2012). That validates that topics can be used in connection with reality, the same way community-related hashtags or keywords can be used to reveal the agenda. However, it cannot be extended to any community social network because it is based on hashtags or trend indicators that are not presenting, for example, on forums.

The last considered method is the manual analysis of a community agenda; for example, this paper (Perry 2018) analysed the landscape of right-wing extremists community. This approach provides an extensive research and sheds light on different aspects of the community's life. The main drawback is the high costs that do not allow to make an analysis frequently and on a large volume of information.

We decided to proceed with the NLP approach because it can be used with any textual data, not only certain social networks with hashtags, and capable of processing a lot of information. Additionally, it does not require a lot of manual resources, which means that the method costs will be relatively small and the stakeholders can use the method more often.

2.3 Framework explanation

2.3.1 Exploratory data analysis and models research

When we obtained the data and conducted an exploratory analysis; consequently, we realised that the information volume is huge. We decided to limit it by the last five years of forum history, as this data is the most relevant. Besides, the dataset contained several languages; thus, we decided to limit the scope to only English posts because some NLP techniques are sensitive to vocabulary size.

The further step was to analyse different possible option to use an automatization of revealing the general agenda information from StormFront's posts. The most plausible task turned out to be the natural language processing task - topic modelling. This task's input is the set of text documents, and output is the distribution of documents across a set of topics to which the input documents belong. Usually, each of these topics is provided as a list of keywords that describe the topic in the best way.

The next stage was the research of topic modelling methods. We selected three families of models. The first family was generative models of a bag-of-words representation (Harris 1954) of the dataset and a representation of topic distribution across the dataset. For instance, such methods are probabilistic latent semantic analysis (Hofmann 1999) and latent Dirichlet allocation (LDA) (Blei, Ng & Jordan 2001). Advantages of these methods

are comparatively low resources requirements, few hyperparameters and the ability to provide distribution of topics for each document as a result.

The next family was LDA techniques with word-embedding representation (Mikolov, Sutskever, Chen, Corrado & Dean 2013). Word-embedding representation allows catching a context of words to provide a better semantic approximation. Disadvantages of these methods are high resources requirements and strong dependency on embedding, which means that the dataset for embedding training should have the same semantic distribution as the given dataset.

As a third family, we selected techniques based on the distribution of documents and words and further clustering. The example is top2vec (Angelov 2020), which allocates all the documents and words in the same semantic space and then extracts clusters on this set. The methods are fast in training and do not have many hyperparameters; however, they associate the document with only one topic, which is not always semantically correct.

2.3.2 Modelling

Once we had researched different approaches to topic modelling, we selected one method from each methods family and trained on our dataset. We selected LDA, Embedded Topic Model (ETM) (Dieng, Ruiz & Blei 2019) and top2vec. After training, we used human experts' estimation and detected which one produces the most discerning results on the given dataset for further analysis. In our case, it appeared to be LDA.

When we had selected the model, the next step was tuning of hyperparameters value. The LDA has several hyperparameters, such as the density of values in each sparse matrix and the number of topics that represents the shape of matrices. We decided to find the best value only for the number of topics and used the perplexity metric for that. Once we found the hyperparameter's best value, we applied the model results to the filtered dataset. It was a matrix where each row represents the topics distribution of the document. Accordingly, the number of rows was the number of documents.

2.3.3 Results analysis

At this stage, we had two issues. First, we wanted to validate that the model produces something which reflects reality. The other one, we had to present the results to show topics and their changes clearly. Thus, we decided to create a chart for each topic with the topic's indication and the dynamic of the topic's changes.

We calculated posts' average probability of belonging to a topic across all the posts created in the considered month for each month and each topic. Figure 1 shows the sample of two topics - the most and the least frequent topics; topics are indicated with the keywords above the chart area. The chart shows how often this topic appeared in the given period. Charts are on the same scale, so we can clearly see that the first discussed more often in all the months than the other one. This representation can show the changes in the probability of a topic to appear in the posts and indicates the topic with keywords. One can compare topics to each other if one locate these charts in a grid.

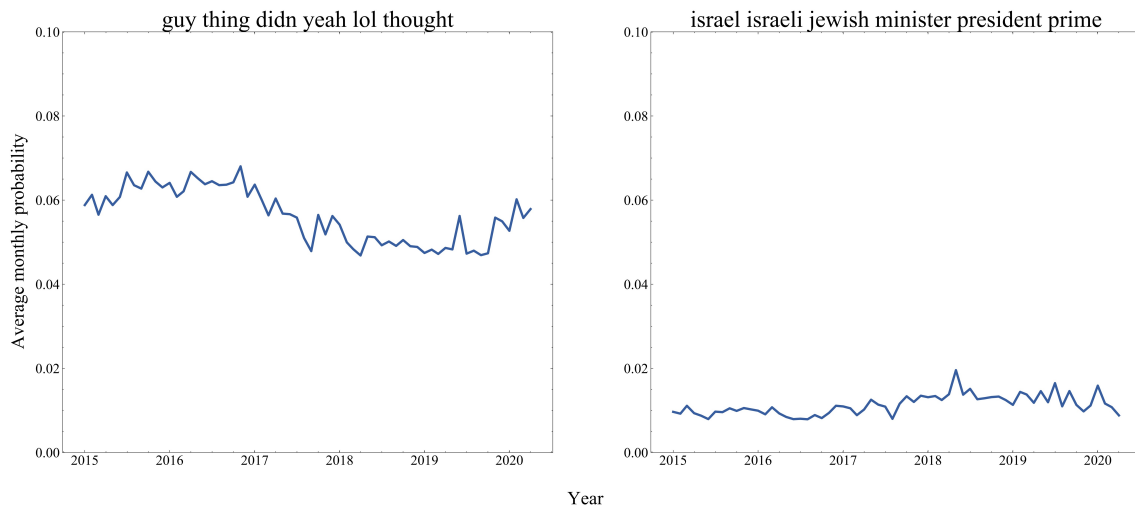


Figure 1. The sample representation of the most and the least discussed topics

The second issue was the validation of the results; with the charts above, we found co-occurrences of the spikes on the charts and the real-world events. Such co-occurrences can evidence that the model represents the real discussions because we know that the community agenda reacts to important events. We found co-occurrences on different scales - monthly data for five years and weekly data for one year. That means that the method can catch a community's feedback and consequently can be used.

2.4 Results usage

The results can be used to identify how the community reacted to the event. For instance, if all the charts like in Figure 1 are located in a grid, one can see the changes after the event has happened. The researchers can use this to analyse the historical data - the information discussed before something has happened needed to find patterns and probably predict the events in the future by the current community's agenda (Nazarko, Frank & Westerlund 2021).

Another usage is the monitoring of the agenda. As described above, if the researchers know patterns in the agenda, they can predict events by the community's agenda. For that, the researchers should monitor the current agenda and its changes in live mode. They can do that with the suggested representation of the results as well.

3 METHODS REVIEW

This section first presents a short theoretical explanation with the pros and cons of each method we tried, then the motivation for the selected method and the process of hyperparameters search.

3.1 Latent Dirichlet Allocation

Our group started topic modelling research with Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan 2001). LDA assumes that each document in a corpus contains one or more topics that humans can understand; usually, those topics can be described by a set of keywords. It means that when human who wrote the text meant something and these "somethings" can be somehow generalised and described with several keywords. For example, let us consider two sentences "I love Finnish nature, there are endless forest, rocks and lakes." and "The nest was found on the edge of the rock near the lake in one of Finnish national nature park." The sentences actually have different meanings, the first one is about one's preferences and the second one is about a found nest. However, there is something in common that we can describe with the words "Finnish, nature, lake, rock". This set of words or, more precisely, the idea which can be described with these words is an abstract topic.

LDA is a generative model. It is a class of models that generates the input distribution and then estimates if the generated data is close to the real data with a certain confidence level. If the confidence level is not achieved, the model creates another way of generating data, generates a new input and so on in a loop.

LDA takes bag-of-words as an input, then using two Dirichlet distributions (let's call them Θ and ϕ), generates topic per document distribution (Z) and eventually words per document (W). Words per document distribution is known as bag-of-words. The next step is the inference if the synthetic bag-of-words distribution is close to the input one; in the LDA implementation we used, it is done with the variational Bayes algorithm. Figure 2 illustrates this process in plate notation, where α and β are parameters for Dirichlet distributions, M is a number of documents, N is a number of words in a certain document, K is the number of latent topics.

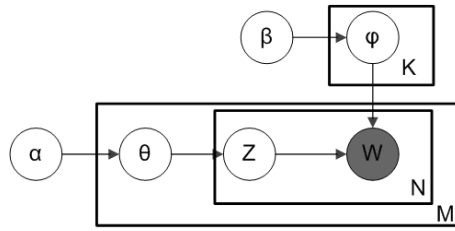


Figure 2. Graphical representation (plate notation) for LDA.

LDA's main drawback comes from the nature of the method; it takes a bag-of-words and does not work satisfying with huge and skewed dictionaries. Huge dictionaries significantly increase the dimensionality of the input and, consequently, the number of operations to implement the method. On the other hand, the method is comparatively easy, and the results can be effortlessly interpreted. Another upside is that the method does not require much training resources compared to deep learning methods.

3.2 Embedding topic modelling

Embedded Topic Model (ETM) (Dieng, Ruiz & Blei 2019) is an extended LDA that uses distributed word representation (Mikolov, Sutskever, Chen, Corrado & Dean 2013) instead of bag-of-words as an input; distributed representation provides dimensionality reduction. In the current research, we used GloVe representation (Pennington, Socher & Manning 2014).

The biggest advantage of this model is that, due to the dimensionality reduction, the method works better than LDA if distributed word representation was pretrained with the documents similar to documents in the corpus. The model considers that words can have similar meanings and appearing in similar contexts.

However, there are several drawbacks as well; first of all, as mentioned above, the documents for embedding training should have the same nature as documents in the researched corpus. In our case, the researched dataset has a peculiar lexicon that determines topics. That can cause the problem that topics are determined not by the suitable keywords but by the contextually close keywords in the corpus for embedding training.

Another drawback is more hyperparameters to search; embedding adds another hyperparameter - the size of each word vector, and the model is pretty sensitive to it. More

hyperparameters require a more thorough search of the best values. Which even more complicated with the third drawback - the model requires many resources to train. One attempt with the pre-trained embedding took three days of training on the cluster machine.

3.3 Top2vec

Another approach to topic modelling is top2vec (Angelov 2020); the approach converts documents and words into distributed representation in the same semantic space, clusters the documents, and represents topics as centroids of clusters. The closest words to the topics can represent the keywords.

The upside is that the method does not require a predefined number of topics. However, in our case, it produced several thousand topics with many duplications, which were not useful at all due to the limitation of human analysis. We could perhaps merge similar topics with some further research and get reasonable for manual analysis number of topics. Another upside is the ability of relatively fast training compared to deep learning methods and LDA.

The main drawback is that with this approach, each document can belong to only one topic. For example, if a document is a mixture of topics, then LDA-like methods would assign probabilities for each topic accordingly, but top2vec instead assign the document to a certain topic. In such a case, the information about the other topics in the post will be lost. This simplified approach can affect future analysis.

3.4 Method selection

As a first step, we at once rejected the consideration of the embedding topic modelling. The trial run showed that each training would take such much time that we cannot afford a proper hyperparameter search in a reasonable time, because each train session takes several days. Furthermore, the model has more hyperparameters than others. Another factor, which affected this decision was that topics did not have much sense in the trial run. Table 1 shows that several topics suspiciously had almost the same set of top keywords. This outcome may be due to another drawback of this model that the specific lexicon that

Table 1. Sample topics with keywords from ETM outcome

Topic #	Keywords
2	people, time, good, make, back, years, find, lot, long
10	people, time, good, make, back, made, things, put, lot
20	people, time, good, make, back, years, long, lot, made
43	people, time, good, make, years, made, back, long, put
44	people, time, good, make, back, made, things, lot, thing
47	people, time, good, make, back, years, made, find, lot
9	people, time, make, good, back, years, long, made, lot
23	people, good, time, make, back, lot, years, made, things
26	people, good, make, time, put, back, things, lot, give
31	people, good, time, make, back, years, lot, made, put
39	people, good, make, time, thing, back, things, years, long
40	people, good, make, time, thing, things, long, lot, back
48	people, good, make, time, back, years, lot, things, made
33	good, people, time, make, back, find, lot, things, years

defines topics was not considered in the corpus for training GloVe, and the model just found the semantically closest words from the context. Consequently, there was a risk that we will not succeed in the modelling even after the proper hyperparameter search, which we anyway could not afford.

Between LDA and Top2Vec, we selected LDA mainly because it provides a mixture of topics for each document as a result, which is a more complex model than each document belongs to one topic, but at the same time more representative, at the same time, LDA does not have significant downsides. Furthermore, Top2Vec required additional work for merging duplicated topics that would not guarantee that topics will be meaningful.

3.5 Hyperparameters search

Once the model had been selected, we needed to find the proper value for hyperparameters. For LDA, three main hyperparameters are α and β , which are parameters for Dirichlet distributions and the number of latent topics. α and β define how specific the initial distribution will distribute topics per document and words per topic; we wanted this distribution to represent each document with few topics and a few words. For that, the Dirichlet distribution should be sparse, consequently have a parameter value less than 1. We selected the default for sklearn LDA implementation, which is 1 divided into a number of topics.

To estimate the number of topics, we selected the perplexity metric. Perplexity for a topic modelling shows an ability to generalise the results. In other words, it shows how a model performs if it applied to the data which was not used for training. Another metric could be the topic coherence; it estimates a semantic distance between the top keywords for each topic. Although it can present how good the topic was generated, it can tell nothing about a model's ability to represent reality. Considering that, we decided to proceed with perplexity.

In our experiment, we split the data into test and train sets, 25% and 75% accordingly. Then we calculated perplexity on the test set for LDA outcomes trained on the train set with the following number of topics values 5, 10, 30, 40, 50, 70, 100 and 200. The lowest perplexity value was for 40 topics; this value we used in further analysis.

4 VALIDATION

When we acquired the model output, the next step was to validate that the model produces results that can reflect reality and serve the goal of our research. The validation should approve that the model results are sensible, interpretable, and reproducible; simultaneously, results can be used to define the community’s agenda.

The biggest impediment was that the model we used is unsupervised. Validation of unsupervised models is always more complex to conduct than that of supervised ones. The reason is that a quality metric for validation is tough to estimate because the data points have no labels comparable to the ones produced by the model.

We considered several approaches to such validation, including the calculation of metrics like perplexity or topics coherence as well as a manual estimation of the model outcome by humans. These approaches are most common for topic modelling tasks. We kept in mind that the goal was to validate the model’s performance and whether the model serves our research’s goal.

First, we started researching metrics, starting with perplexity (eq. 1, where M is the number of documents, w_d represents the words in document d , N_d the number of words in document d). This metric is used for hyperparameter search. The metric is based on likelihood and shows how the model generalises results and performs on new unseen data.

$$perplexity = exp \left\{ - \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \right\} \quad (1)$$

The second metric that we considered was topic coherence — the metric estimates for each topic how top keywords for a topic are semantically close to each other (eq. 2 for coherence of single topic, where w_i and w_j are the topic’s keywords; eq. 3 is U_{Mass} score, where D is number of documents which contain given keywords). This metric shows the

interpretability of outcome topics.

$$coherence = \sum_{i < j} U_{Mass}(w_i, w_j) \quad (2)$$

$$U_{Mass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)} \quad (3)$$

Although it is straightforward to calculate the metric, we faced two problems; firstly, we needed a threshold to use metrics to validate the model based on it, and we did not have such a threshold. These metrics are helpful if there is a need to compare different outcomes or models, but it is not easy to estimate if the model performs well in general.

The second problem of metrics was that based on the definition of these metrics, none of them validates how good our model serves the primary goal of the research as well as how good the model finds topics representing the community's agenda.

Thus, we decided to apply another validation method — judgment of the model outcome by human experts. It required more resources than calculating metrics, but experts could estimate both the model performance and the ability to represent the community's agenda. In our case, it means that we, as human experts, need to validate if topics distribution is proper on average for documents in a certain period. If we know that this is correct, we can inductively extend that into the whole timeline.

We decided to validate that by finding the co-occurrences of real-world event and related topics discussions reactions. We know that those loud events, which especially lit by the media, raise discussions in the community. Such co-occurrences can validate that the model assigns many documents, or posts in our case, to belong to the specific topic. This, in turn, validates that the discussion reflects reality and the distribution of topics is correct.

To find such co-occurrences, we needed a representation of results that shows dynamic changes in the volume of the discussions around each topic. For that, we calculated

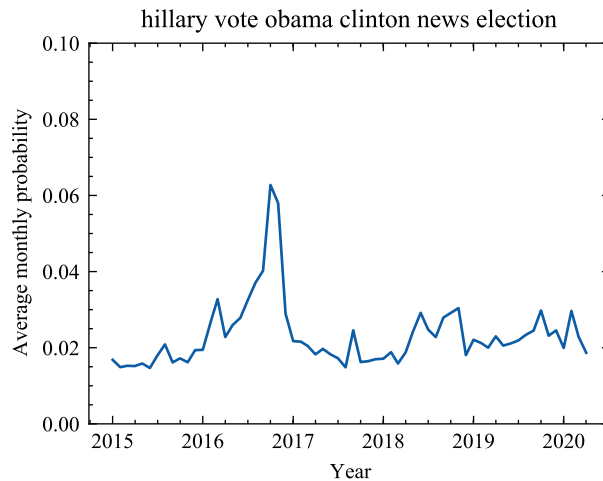


Figure 3. Average monthly topic probabilities for topic 24

mean probabilities across the posts for a period to belong to the topic and plotted it. The appeared charts turned to fit well for this task. Figure 3 shows an example of such a chart and co-occurrence of real-world event and the volume of discussion of a particular topic. It is seen that the peak in discussion happened at the same time as the US election at the end of 2016.

Changing the period for calculating the average probability, we can plot charts on different scales; e.g. in figure 3, we calculated the average probability of topic across posts per month, and each point on this chart was a month. However, a month is a significant period, and only significant events which raise many discussions can change the shape of the line. To validate that smaller events also affect discussions of topics, we also researched the same charts but plotted the average probability of topic across posts per week within one year. The results and figures of the five-year interval validation are presented in section 5.1.1 and for the one-year interval in section 5.1.2.

5 RESULTS

When we agreed on the validation process, the next step was conducting validation and getting results. Results are valuable as evidence that the community reacts to the events, and the model produces something valuable. This chapter contains two parts. First, we explain the validation results and then elaborate on how the model's results can serve the goal of the research by extracting an agenda of the community from the forum posts.

5.1 Validation results

As was mentioned in the validation section, we calculated the average probability of a topic appearing within a period and plotted charts for each topic. We considered two scales for the validation and assumed that these scales are sufficient to validate the model results (Nazarko, Frank & Westerlund 2021). These scales reflect monthly data for the interval from January 2015 to May 2020 and weekly data for 2019. In this subsection, we will present the results of both scales.

5.1.1 Five years scale

During the interval from January 2015 to May 2020, the most noticeable political event in North America was the 2016 United States presidential election. Our model found two topics connected to the presidential election according to keywords of these topics. Charts of these topics (charts 3 and 4) show that both topics have a comparatively high probability of appearing in posts at the end of 2016.

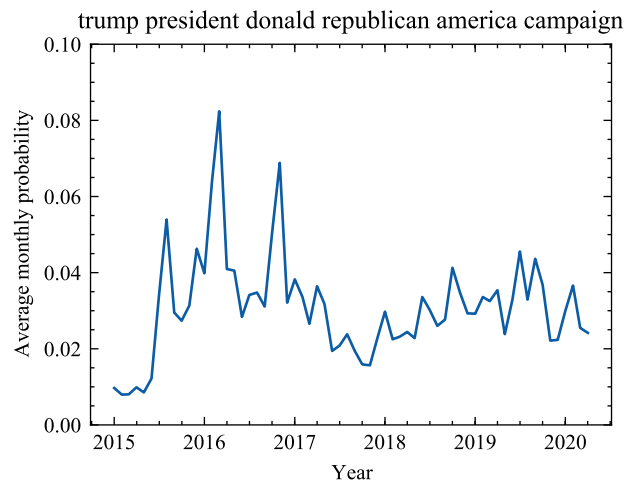


Figure 4. Average monthly topic probabilities for topic 2

Another thunderous event for this specific community in this period was the outage of Stormfront forum in August of 2017. The model also found the topic with keywords *forum Stormfront site*. One can see in figure 5 that the average probability plot of this topic has a spike in the middle of 2017, which evidence that the community discussed this topic a lot at this particular time.

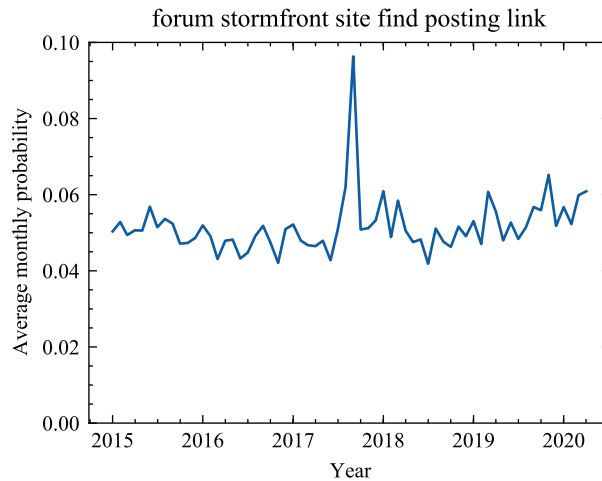


Figure 5. Average monthly topic probabilities for topic 33

The last example on this scale is the event that affected the world economy and re-established world order - the COVID-19 pandemic. The pandemic affected all the communities, including Stormfront users. The model highlighted a topic with keywords *health medical drug disease*, which are connected to health-related discussions. Figure 6 shows that this topic has a tremendous increase in the beginning of 2020.

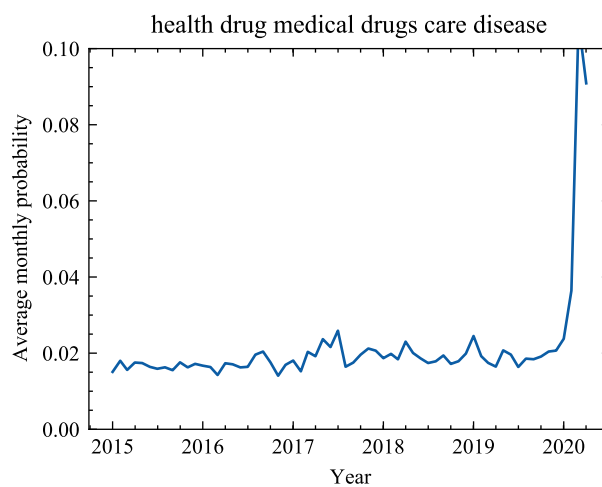


Figure 6. Average monthly topic probabilities for topic 16

5.1.2 One year scale

For one year scale, we took posts from January 2019 to May 2020. Similar to what we did for months on a bigger scale, we calculated average probabilities for each topic for each week within the interval. The fluctuations on this scale can evidence that less significant events raise fewer discussions than global ones, but the model still can catch this.

The first considered event happened in January, just at the beginning of 2020, and it was a drone strike at Baghdad International Airport conducted by the United States, which concluded with the assassination of an Iranian major general Qasem Soleimani. The topic with keywords military syria isis iran security had a peak in the first weeks of 2020, as shown in figure 7.

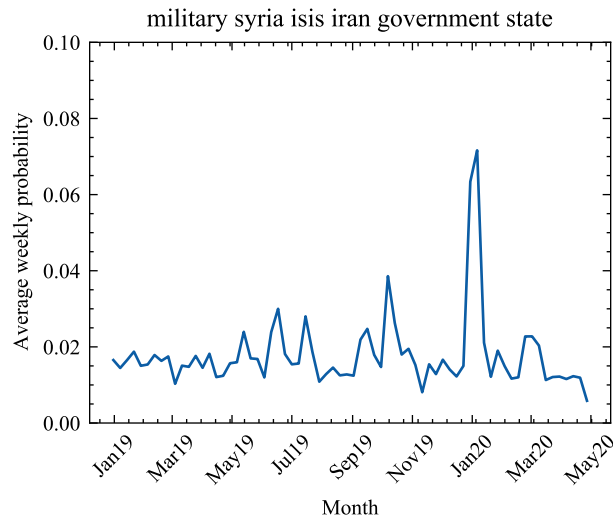


Figure 7. Average weekly topic probabilities for topic 23

More results of validation of a one-year interval scale can be found in the original article (Nazarko, Frank & Westerlund 2021).

Concluding the part about model validation results, we argue that these co-occurrences prove that our model can capture topics in the community alongside proving that the community reacts to specific events, and these events capture the focus of attention and affect the discussed topics.

5.2 Results usage

The validation results subsection showed that we could rely on the selected method, and the last part is to present that the results could serve the stakeholders' purposes and bring the proposed value. In particular, we claimed to find a way to define the community's agenda and its changes.

The agenda can be presented by placing charts like figure 3 for each topic on the same plot scale. For instance, figure 1 shows that the topic related to Israel is discussed less likely in all the points than a random topic with keywords "guy thing didn yeah lol thought", representing some reactions to others messages. It means that Israel related topics always contribute less to the agenda than the reactions one do (Nazarko, Frank & Westerlund 2021).

Crucial for the stakeholders, the agenda also changes. These changes can be presented through accommodation charts like figure 6. As we can see at the beginning of 2020, the topic increased significantly, which means the community agenda changed and topic related to health was discussed much more often than before (Nazarko, Frank & Westerlund 2021).

In conclusion, we argue that the provided results of the research allow us to claim that the goal was achieved. The model produces the correct outcome; the outcome is valuable for the global results. The selected framework brought us to the point where we can bring value to the stakeholders and provide them with the information they need.

6 CONCLUSION

Communities are an important part of society. Some of them can easily radicalise, and researchers need to find a way to get information about what is going on there. We limited the whole community research to the analysis of the community's agenda and its changes. Stakeholders can use this information as support for making decisions that are heavily connected with the community. For example, a decision about an increase of vigilance or which content should be added to the advertisement. The analysis of previous research showed that the StormFront community agenda was mostly analysed manually. This way does not provide the required level of scalability, allowing the processing of a significant amount of information. Our intention was to use NLP techniques to solve this problem (Nazarko, Frank & Westerlund 2021).

We defined a framework for this research. As a first step, the data should be gathered, cleaned, and processed to suit the NLP models. Also, language for each detection was conducted at this step. The data was filtered to only include posts in English, created not earlier than 2015. The next step in the research was the model selection; we considered three models - Latent Dirichlet Allocation, Embedding Topic Modelling, and top2vec. Weighing the pros and cons of considering models, we decided to proceed with Latent Dirichlet Allocation. To find the optimal values of the hyperparameters, we used perplexity as a metric.

Once the model's results were gathered, we presented the process of model result validation. We validated these results by finding co-occurrences of real-world events with the spikes on the charts of mean probabilities across the post to belong to a certain topic. Such co-occurrences are evidence that the model produces meaningful results and reflects reality. In the end, we showed how the validated model outcome brings value for the stakeholders. We found a way how to represent the community's agenda and its changes (Nazarko, Frank & Westerlund 2021). This way, we showed that the selected framework could provide the results and obtain the proposed value.

REFERENCES

- Angelov, Dimo. 2020, *Top2Vec: Distributed Representations of Topics*.
- Bajak, F. 2019, *Online providers knock 8chan offline after mass shooting*. Available: <https://abcnews.go.com/Technology/wireStory/security-cut-off-cesspool-hate-8chan-forum-64778026>.
- Blei, David; Ng, Andrew & Jordan, Michael. 2001, Latent Dirichlet Allocation, vol. 3, pp. 601–608.
- Choi, Hyunyoung & Varian, Hal. 2012, Predicting the present with Google Trends, *Economic record*, vol. 88, , pp. 2–9.
- Dieng, Adji B.; Ruiz, Francisco J. R. & Blei, David M. 2019, *Topic Modeling in Embedding Spaces*.
- Gill, Paul; Corner, Emily; Conway, Maura; Thornton, Amy; Bloom, Mia & Horgan, John. 2017, Terrorist use of the Internet by the numbers: Quantifying behaviors, patterns, and processes, *Criminology & Public Policy*, vol. 16, no. 1, pp. 99–117.
- Harris, Zellig S. 1954, Distributional Structure, *WORD*, vol. 10, no. 2-3, pp. 146–162. Available: <https://doi.org/10.1080/00437956.1954.11659520>.
- Hofmann, Thomas. 1999, Probabilistic Latent Semantic Analysis, In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., p. 289–296.
- 2019, *Homegrown Violent Extremist Mobilization Indicators 2019 Edition*, Federal Bureau of Investigation, Washington, DC: Office of the Director of National Intelligence. Available: <https://www.dni.gov/index.php/nctc-newsroom/nctc-resources/item/1945-homegrown-violent-extremist-mobilization-indicators-2019>.
- Lu, Rong & Yang, Qing. 2012, Trend analysis of news topics on twitter, *International Journal of Machine Learning and Computing*, vol. 2, no. 3, p. 327.
- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg & Dean, Jeffrey. 2013, *Distributed Representations of Words and Phrases and their Compositionality*.

- Mutanga, Murimo Bethel & Abayomi, Abdultaofeek. 2020, Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach, *African Journal of Science, Technology, Innovation and Development*, vol. 0, no. 0, pp. 1–10. Available: <https://doi.org/10.1080/20421338.2020.1817262>.
- Nazarko, Grigorii; Frank, Richard & Westerlund, Magnus. 2021, Topic Modeling of StormFront Forum Posts, In: *COGNITIVE 2021, The Thirteenth International Conference on Advanced Cognitive Technologies and Applications*, IARIA, pp. 37–43. Available: https://www.thinkmind.org/articles/cognitive_2021_2_30_40018.pdf.
- Pennington, Jeffrey; Socher, Richard & Manning, Christopher D. 2014, GloVe: Global Vectors for Word Representation, In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Available: <http://www.aclweb.org/anthology/D14-1162>.
- Perry, Scrivens Ryan, Barbara. 2018, A climate for hate? An exploration of the right-wing extremist landscape in Canada, *Critical Criminology*, vol. 26, no. 2, pp. 169–187.
- Rohani, Vala; Shayaa, Shahid & Babanejaddehaki, Ghazaleh. 2016, Topic modeling for social media content: A practical approach, pp. 397–402.
- Scrivens, Ryan; Gaudette, Tiana; Davies, Garth & Frank, Richard. 2019, Searching for extremist content online using the dark crawler and sentiment analysis, In: *Methods of Criminology and Criminal Justice Research*, Emerald Publishing Limited.
- Törnberg, Anton & Törnberg, Petter. 2016, Muslims in social media discourse: Combining topic modeling and critical discourse analysis, *Discourse, Context Media*, vol. 13, , pp. 132–142. Available: <https://www.sciencedirect.com/science/article/pii/S2211695816300290>.