What's in the News? Identification of Trending Topics in Alternative and Mainstream Lithuanian Media

Justina Mandravickaitė[1], Monika Briedienė[1,2[0000-0001-6165-1702]], Jonas Uus[1] and Tomas Krilavičius[1,2[0000-0001-8509-420X]]

[1] Baltic Institute of Advanced Technology, Pilies str. 16, Vilnius 01124, Lithuania
[2] Vytautas Magnus University, K. Donelaičio str. 58, Kaunas 44248, Lithuania
justina@bpti.lt

Abstract. It is no longer surprising that internet media is a significant appliance in reflecting and shaping public opinion. Tracking topics dynamics and focus in different media channels is an important tool for opinion-forming mechanisms and process analysis. Information collect, text analytics and Artificial Intelligence tools allows identification of trending topics in different media sources, while exploratory visual analytics tools provide means to identify prevalence of topics in different sources, and their dynamics. In this paper we discuss an ongoing research and demonstrate applicability of such approach to main Lithuanian news portal (delfi.lt) and alternative unconventional media channels – sarmatas.lt and netiesa.lt.

Keywords: Topic modelling, Framing, Media Monitoring, NLP, Lithuanian language, Artificial Intelligence, LDA, stm.

1 Introduction

Internet media is an important tool in reflecting and shaping public opinion. Modern tools and technologies allow automatic tracking and comparing dynamics of different topics in different media channels, and analysis of the results using visual tools. We apply a set of such tools for the two types of Lithuanian news portals: main WWW news channel - delfi.lt[1] and two alternative unconventional media channels - sarmatas.lt[2] and netiesa.lt[3]. We apply topic modelling methods for (trending) topics identification, and visual results for the further analysis.

Topic modelling is a text mining technique to discover common topics in a collection of documents. In practice researchers attempt to fit appropriate model parameters to the data corpus using one of several heuristics for maximum likelihood fit.

1 https://www.delfi.lt/, last accessed 2020/03/15
2 http://www.sarmatas.lt/, last accessed 2020/03/15
3 http://netiesa.lt/, last accessed 2020/03/15

Media Framing Dynamics of the ‘European Refugee Crisis’ is analyzed in [1]. This study investigates the national media discourses in Hungary, Germany, Sweden, the United Kingdom and Spain for this time period. LDA was applied 130,042 articles in 5 languages from 24 news outlets. It shows that country-specific media tracks the overall course of the refugee debate, uncovers dynamics and shifts in discourses.

Turkish news analysis is presented in [2]. The dataset consists of 4200 Turkish news titles belonging to 7 classes. NMF was the most successful method for three classes, while for five and seven classes LSA was the most successful method. Comparative study is presented in [3] as well.

There is an interesting study of topic modelling of news articles for two consecutive elections in South Africa [4]. Articles are classified using pairwise cosine similarity to identify similar topics in different periods of elections.

Critical evaluation of the utility of the thematic grouping of texts into ‘topics’ emerging from a large collection of online patient comments about the National Health Service (NHS) in England is presented in [5]. Results show that topic modelling allowed to group texts into topics that were truly thematically coherent with a mixed degree of success, while the more traditional approaches to discourse analysis consistently provided a more nuanced perspective on the data which was ultimately closer to the reality of the texts it contains.

In [6] paper, authors describe their work in developing a model for topic modelling and detection of hot topics being discussed in the local Malay news publisher. This model explored different features for article clustering and topic modelling, and then applied the TextRank algorithm to identify hot topics in the news.

The tremendous growth of social media content on the Internet has inspired the development of the text analytics to understand and solve real-life problems. Leveraging statistical topic modelling helps researchers in better comprehension of textual content as well as provides useful information for further analysis.

Authors [7] have tested Dengue epidemics tracking using Twitter content classification and topic modelling. Classifier achieves a prediction accuracy of about 80 % based on a small training set of about 1,000 instances, but the need for manual annotation makes it hard to track seasonal changes in the nature of the epidemics, such as the emergence of new types of virus in certain geographical locations. In contrast, LDA-based topic modelling scales well, generating cohesive and well-separated clusters from larger samples.

Another experiment with Twitter data set on topic modelling was for identification of vaccine reactions. The study [8] compared Gensim LDA, MALLET, and jLDADMM DMM models to determine the most effective model for detecting vaccine safety signals, assisted by an evaluation process that used an adjusted F-Scoring technique over a labelled subset of the documents.

Paper [9] uses 18,552 tweets dated from 2015 up to 2018 to analyze the dynamics of the LGBT conversation among Indonesian peoples. In this research, they explore the main topic of the LGBT conversation using LDA. The result shows that there are seven main categories that people normally talked about regarding LGBT.

Study [10] summarizes the message content of four data sets of Twitter messages relating to challenging social events in Kenya. They use LDA topic modelling to analyze the content. This study uses two evaluation measures: Normalized Mutual Information (NMI) and topic coherence analysis, to select the best LDA models. The obtained LDA results show that the tool can be effectively used to extract discussion topics and summarize them for further manual analysis.

Investigations can be done with short texts as well. [11] conduct a topic modelling of 6854 Instagram posts made by Ramzan Kadyrov (the head of the autonomous Chechen Republic in the Russian Federation). Researchers analyze the verbal framing of 24 dominant topics. The study concludes that the main rhetorical device that Kadyrov employs is a merging of personal and political themes throughout his posts.

2 Data and Methods

2.1 Corpora

Corpus consists of 5000 delfi.lt articles (a random sample from News category of delfi.lt corpus [12]), 1145 sarmatas.lt articles and 2411 netiesa.lt articles, both published in a period of 2014 – 2016 years. Delfi.lt is the mainstream news portal, the most readable and visited channel in Lithuania, while sarmatas.lt and netiesa.lt are alternative source of media in selected geographical indication. Sarmatas.lt is one of the most important sources in terms of dissemination of information (project Research Meadow[4], 2014) and netiesa.lt is unconventional but quite popular news portal among Lithuanian portal readers.

2.2 Methods

Topic analysis is a Natural Language Processing (NLP) technique that allows automatically extract meaning from texts by identifying recurrent themes or topics. The goal of the structural topic model is to discover topics and estimate their relationship to document metadata. LDA is a particularly popular method for fitting a topic model [13]. It treats each document as a mixture of topics and handles each topic as a mixture of words [13]. This allows documents to "overlap" with content, rather than grouping them in a way that reflects the normal use of natural language.

The structural topic model allows researchers to flexibly estimate a topic model that includes document-level metadata [14]. Estimation is accomplished through a fast variation approximation. In this research the stm package [14] was used, it provides many useful features, including rich ways to explore topics, estimate uncertainty, and visualize quantities of interest. Structural topic modeling operating principle:

1. The generative model begins at the top, with document-topic and topic-word distributions generating documents that have metadata associated with them;
    - a topic is defined as a mixture over words where each word has a probability of belonging to a topic.

[4] http://mokslopieva.lt/, last accessed 2020/03/15

- a document is a mixture over topics, meaning that a single document can be composed of multiple topics. As such, the sum of the topic proportions across all topics for a document is one, and the sum of the word probabilities for a given topic is one.

2. Topical prevalence refers to how much of a document is associated with a topic (described on the left hand side) and topical content refers to the words used within a topic (described on the right hand side). Hence metadata that explain topical prevalence are referred to as topical prevalence covariates, and variables that explain topical content are referred to as topical content covariates

In this work, the R [15] package stm [14] for structural topic modeling was used.

2.3 Overall Process

We used the following process for the analysis:

1. corpora were collected from the corresponding portals (not part of this research);
2. corpora were created from the random sample from delfi.lt and selected sarmatas.lt, netiesa.lt articles;
3. all texts were lemmatized and lowercased using SpaCy[5] Core Lithuania models;
4. stopwords[6], numbers, symbols and punctuation marks were removed;
5. documents were represented as bag-of-words (a text is represented as the bag (multiset) of words, disregarding grammar and even word order but keeping frequencies.);
6. low frequency words (5% of the least frequent words in the whole corpora) and 5% of words that occurred in all the texts were removed;
7. Latent Dirichlet Allocation (LDA) [16] and stm R function [14] were applied for structural topic modelling;
8. results were visualized.

3 Results

Topic modeling is part of a class of text analysis methods that analyze “bags” or groups of words together—instead of counting them individually–in order to capture how the meaning of words is dependent upon the broader context in which they are used in natural language. So foremost investigation was for finding the expected proportions in the data (see Fig. 1).

---

[5] https://spacy.io/, last accessed 2020/03/15

[6] https://github.com/tokenmill/ltlangpack, last accessed 2020/03/15

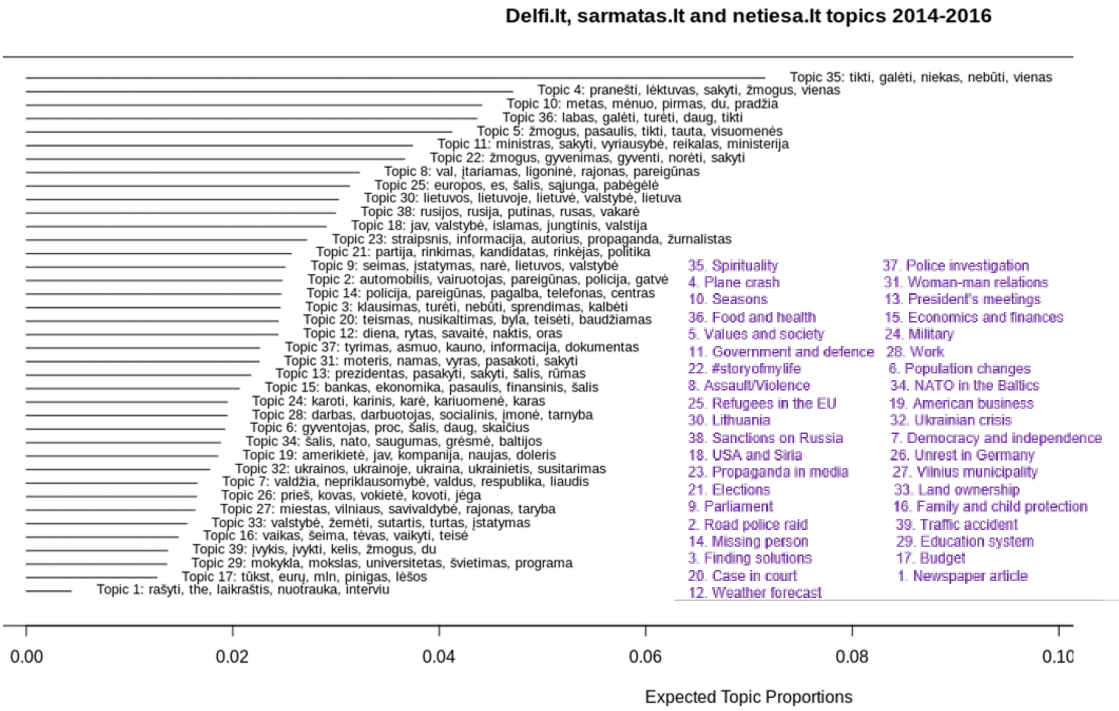
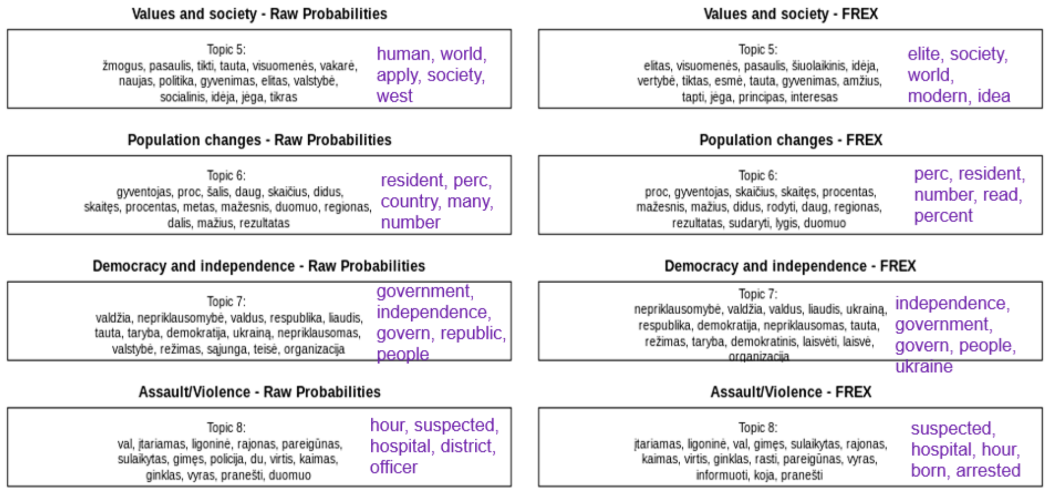


Fig. 1. Topics by expected proportions in the data.

After this approach we have to select and take into account the words with the highest (raw) probabilities and the highest FREX (Frequency and Exclusivity, i.e., words that are most frequent and exclusive to the topic) (see Fig. 2).



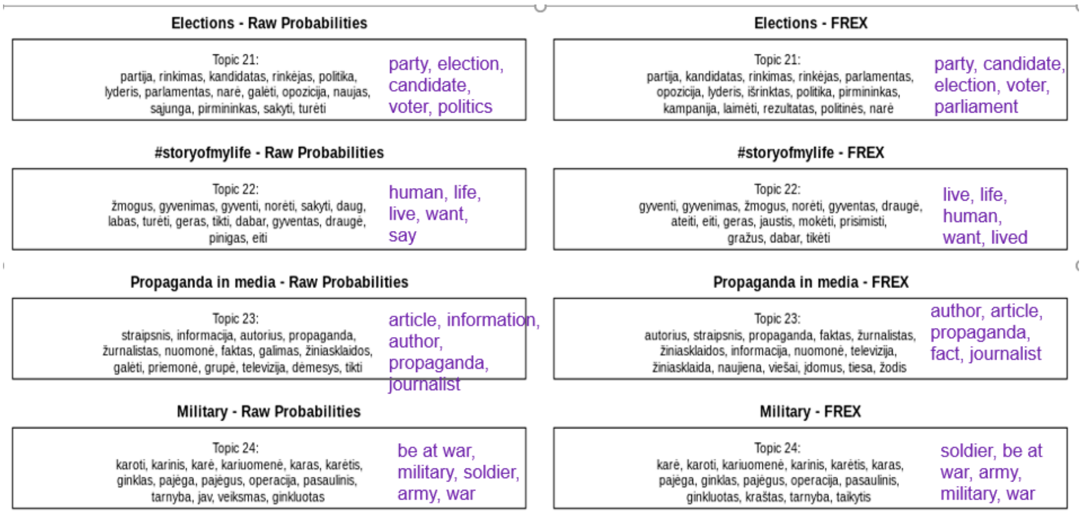


Fig. 2. The words with the highest (raw) probabilities and the highest FREX.

We apply LDA for delfi.lt and sarmatas.lt dataset. Analysis shows rather different combination of topics in the portals, see Fig. 3. In delfi.lt (left) orange topics (*democracy, traffic accidents, referendum on preventing foreigners from owning land in Lithuania, ceasefire negotiations in Ukraine, activities of the state security department of Lithuania*, etc.) prevail, while in sarmatas.lt (right) blue-purple topics (*Islam and terrorism, industry, Maidan, taxes, migrants and refugees*, etc.) are significant part of content. Summaries of identified topics (highly probable words) were assigned by experts after qualitative analysis.

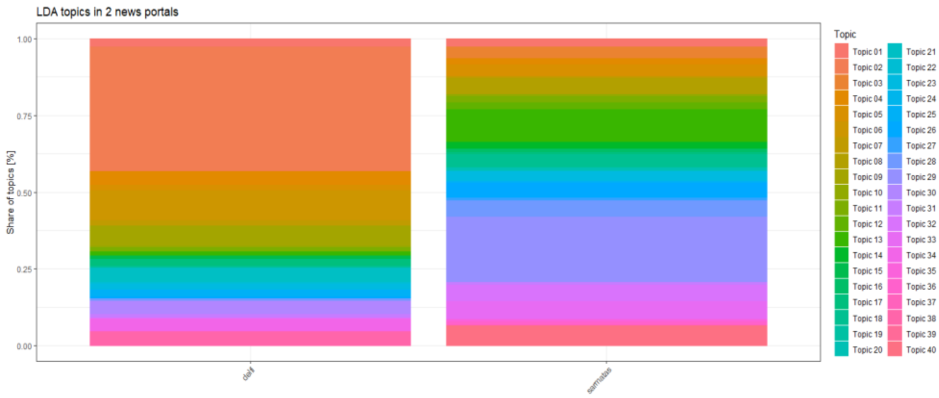


Fig. 3. LDA topic prevalence and distribution in delfi.lt (left) and sarmatas.lt (right).

Interpretability of topics built by topic modeling is an important issue for researchers applying this technique. Our investigation showed that higher semantic coherence indicates topics that have more consistent words (more interpretable) while exclusivity measures how exclusive the words are to the topic relative to other topics (e.g. low values mean topics that are vague and share a lot of words with other topics while high values indicate words that are very unique/exclusive to the topic) (see Fig. 4).

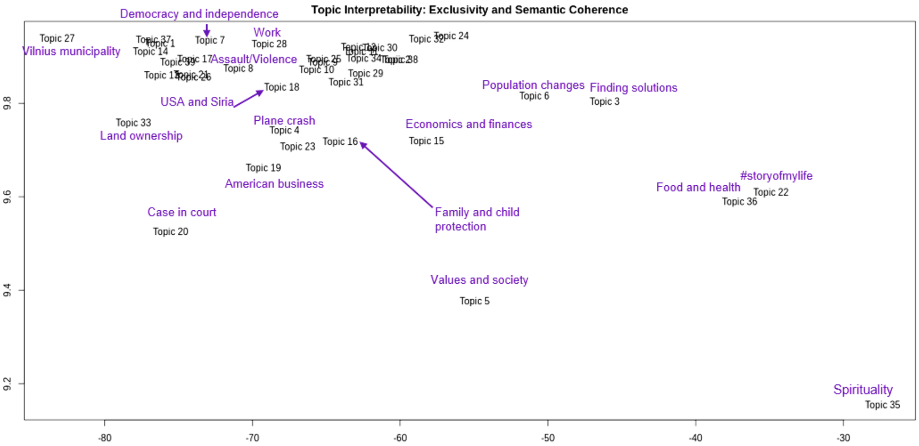


Fig. 4. Topic interpretability: the exclusivity and semantic coherence (X axis represents semantic coherence, Y axis – exclusivity).

After examination of the whole set, we focused on the distribution of topics across different media channels. The stm is a general framework for topic modeling with document-level covariate information. The covariates can improve inference and qualitative interpretability and are allowed to affect topical prevalence, topical content or both. The software package implements the estimation algorithms for the model and also includes tools for every stage of a standard workflow from reading in and processing raw text through making publication quality figures. Topical prevalence refers to how much of a document is associated with a topic and topical content refers to the words used within a topic. Expected difference in topic probability be media type (with 95 % confidence intervals) is shown below (see Fig. 5 and Fig. 6).

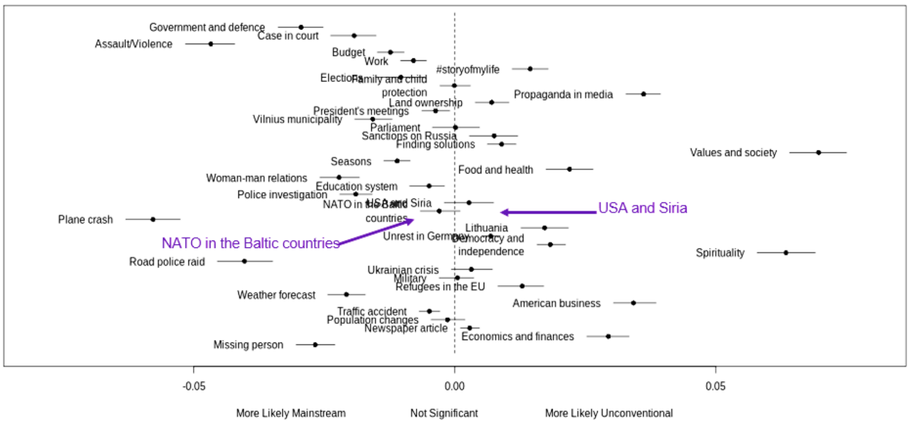


Fig. 5. Effect of media Type on Topic Prevalence in 2014-2016.

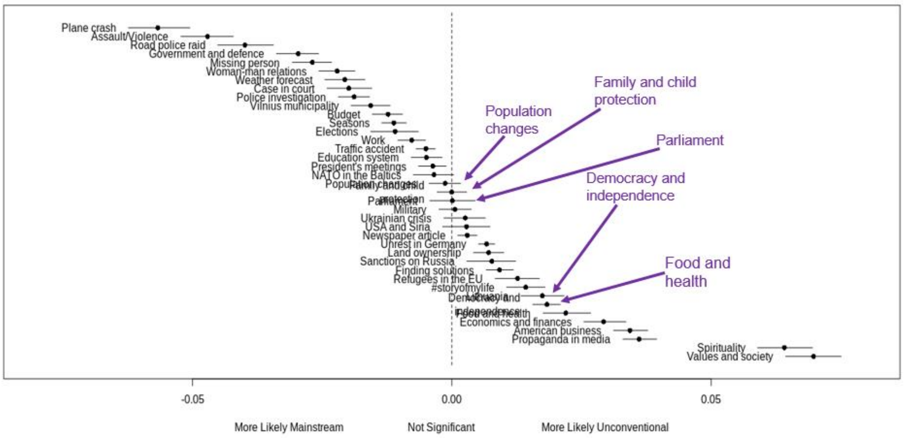


Fig. 6. Effect of media Type on Topic Prevalence in 2014-2016.

Following examining the distribution of all topics, we focused our research on key sensitive topics. We find that the model captures important events and differences between different media channel' depictions of these events (see Annex 1).

Topic correlation network creation results are depicted in Fig. 7. The way these algorithms work is by assuming that each document is composed of a mixture of topics, and then trying to find out how strong a presence each topic has in a given document. This is done by grouping together the documents based on the words they contain, and noticing correlations between them. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

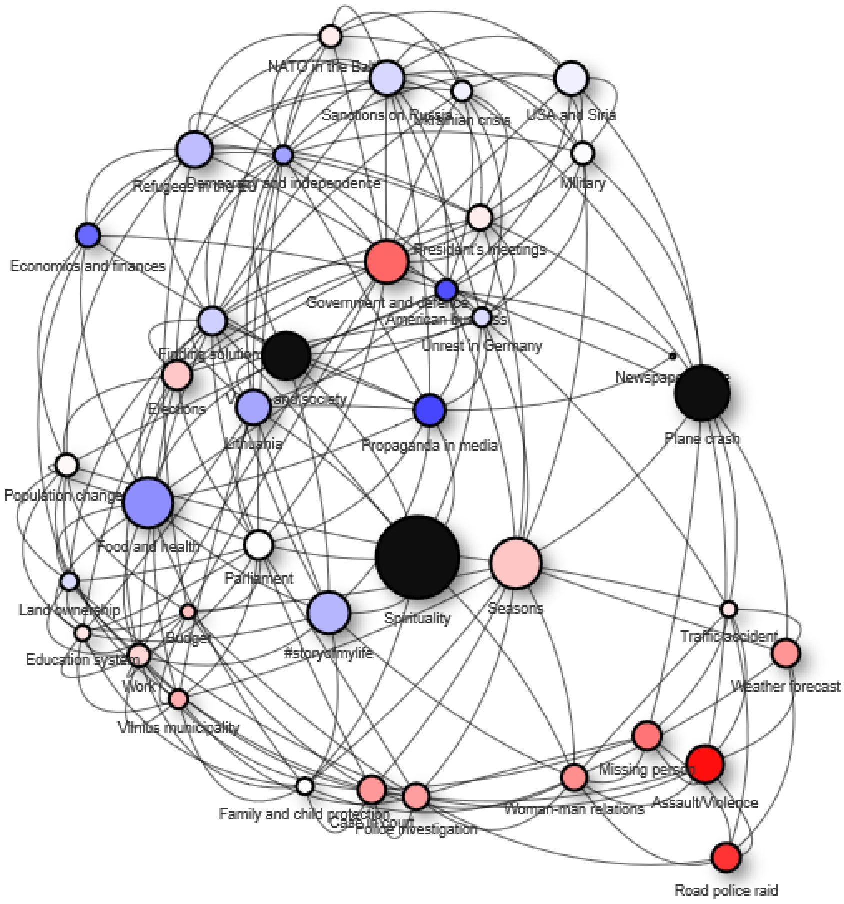


Fig. 7. Delfi.lt, Sarmatas.lt and Netiesa.lt Topic (Correlation) Network. Explanation: Blue – more typical to unconventional media; Red – more typical to mainstream media; Black – topics that differ delfi.lt (mainstream) and alternative/unconventional (sarmatas.lt and netiesa.lt) news sources most.

Topic models have become a standard tool within quantitative text analysis for many different reasons. Topic models can be much more useful than simple word frequency or dictionary based approaches depending upon the use case. Topic models tend to produce the best results when applied to texts that are not too short (e.g. tweets), and those that have a consistent structure.

4 Conclusion and Future Plans

We discussed an ongoing research of: (1) text analytics and Artificial Intelligence tools to identify trending topics in different media sources; (2) exploratory visual analytics tools to identify prevalence of topics in different sources & their dynamics. We demonstrated the applicability of such approach to mainstream Lithuanian news portal (delfi.lt) and two alternative/unconventional media channels – sarmatas.lt and netiesa.lt. Early stage analysis shows considerable difference of prevalent topics in different media channels, which allows identifying targets of the channel.

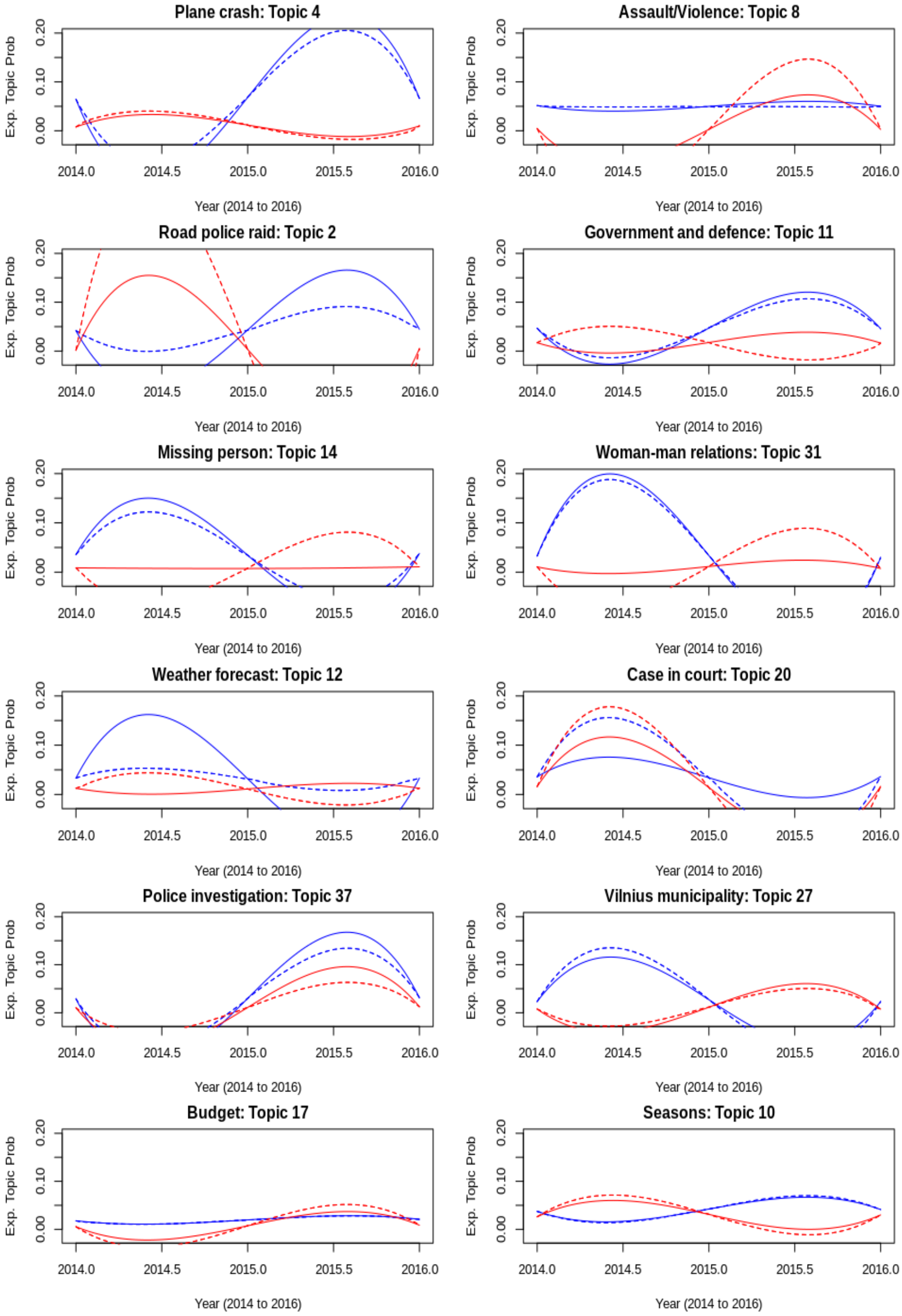
We plan to extend research to wider set of media sources, change of topics in time (more detailed) and relations between topics and media channels (more detailed).

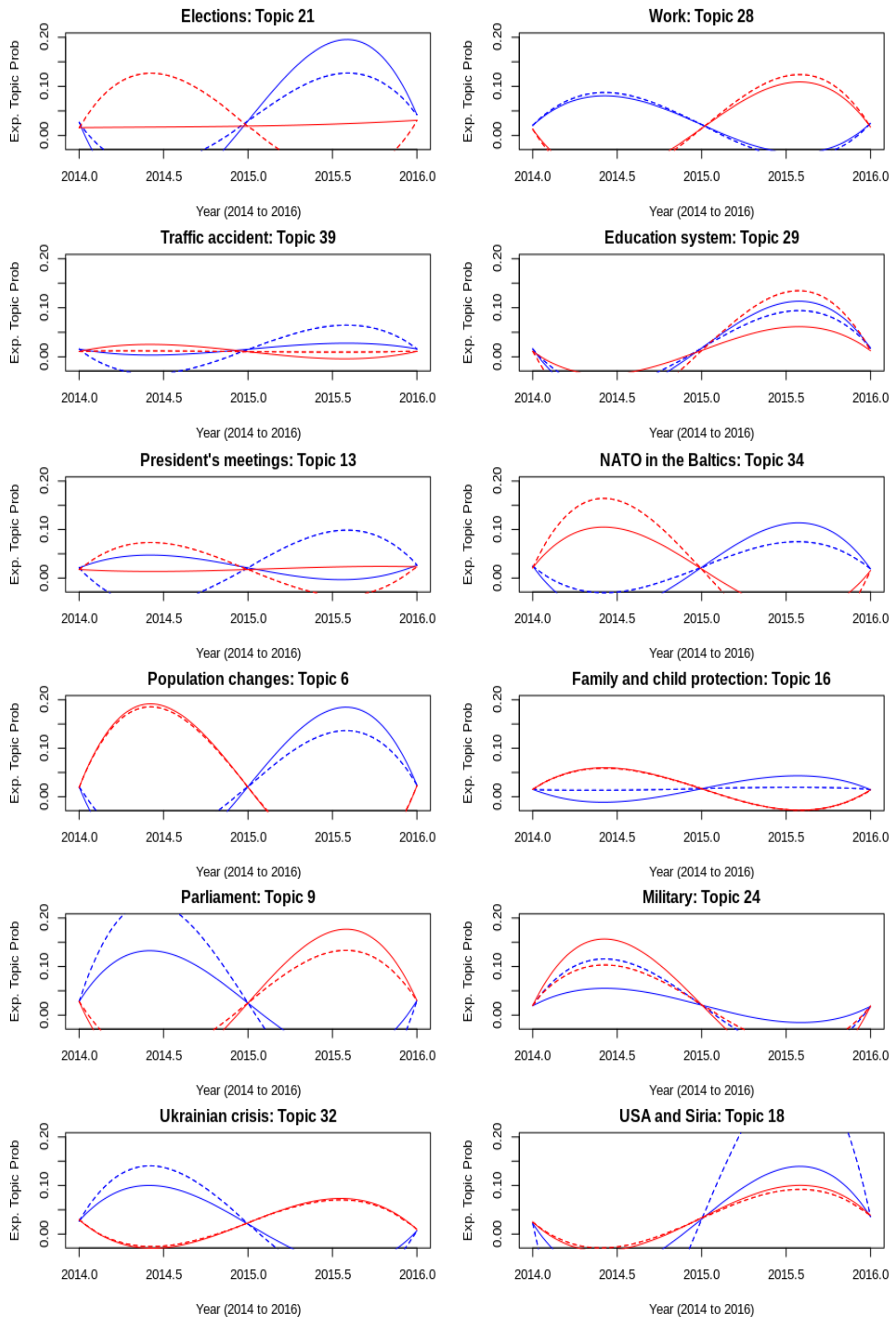
References

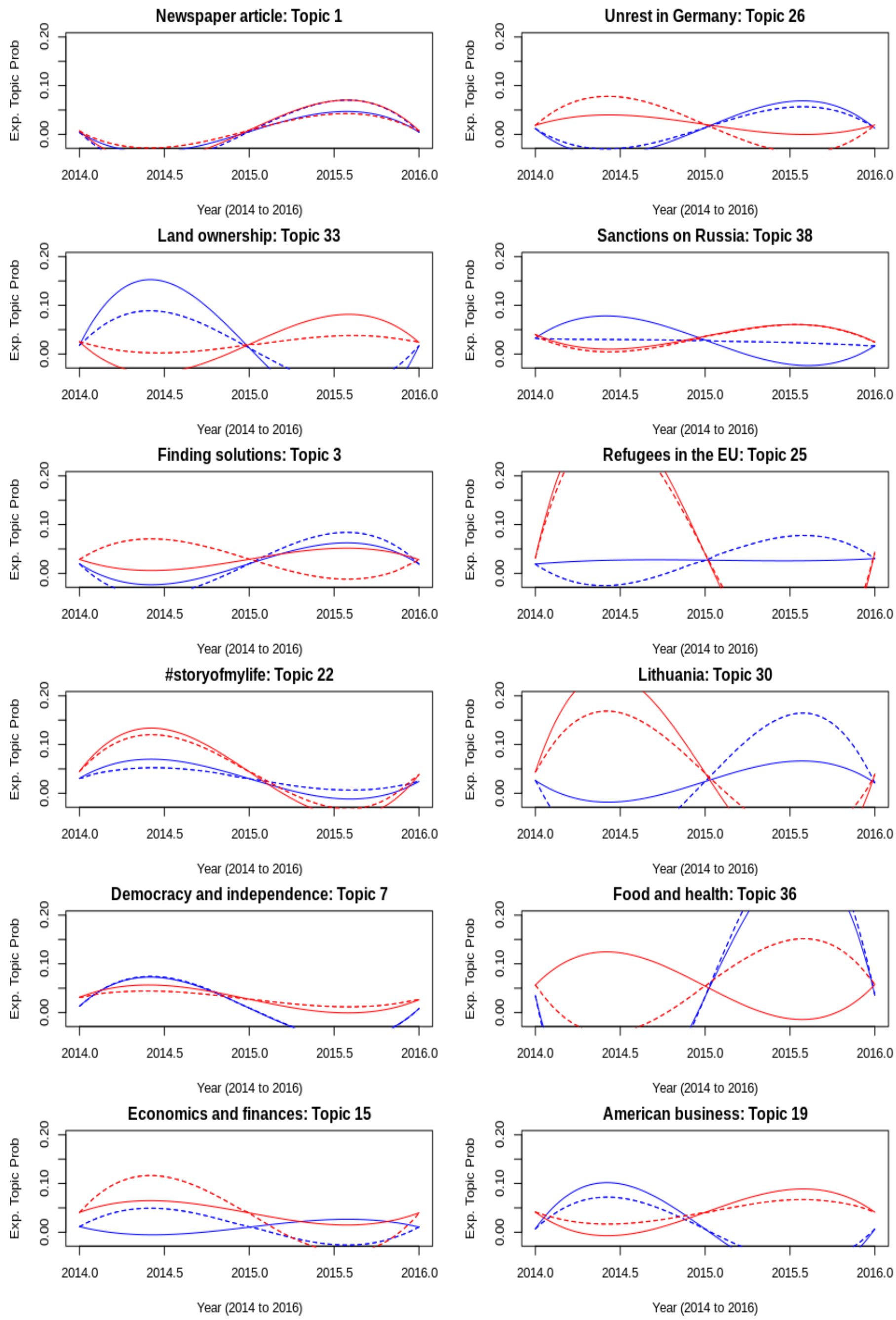
1. Heidenreich, T., Lind, F., Eberl, J. M., Boomgaarden, H. G.: Media Framing Dynamics of the 'European Refugee Crisis'. A Comparative Topic Modelling Approach, Journal of Refugee Studies, 32(1), i172–i182 (2019), https://doi.org/10.1093/jrs/fez025, last accessed 2020/03/15.
2. Güven, Z. A., Diri, B., Çakaloğlu, T.: Comparison of Topic Modeling Methods for Type Detection of Turkish News. In: 4th International Conference on Computer Science and Engineering (UBMK), pp. 150-154, Samsun, Turkey (2019).
3. Kherwa, P., Bansa,l P.: Topic Modeling: A Comprehensive Review, SIS, EAI (2019), doi: 10.4108/eai.13-7-2018.159623, last accessed 2020/03/15.
4. Moodley, A., Marivate, V.: Topic Modelling of News Articles for Two Consecutive Elections in South Africa. In: 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 131-136, Johannesburg, South Africa (2019).
5. Brookes, G., McEnery, T.: The utility of topic modelling for discourse studies: A critical evaluation'. Discourse Studies 21(1), 3–21 (2019), doi: 10.1177/1461445618814032, last accessed 2020/03/15.
6. Weiying, K., Pham, D.N., Hai, N.C., Ong, H. H.:Topic Modelling for Malay News Aggregator. In: Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), pp. 1-6, Subang Jaya, Malaysia (2018).
7. Missier P. et al.: Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. In: Casteleyn S., Dolog P., Pautasso C. (eds) Current Trends in Web Engineering, ICWE 2016, Lecture Notes in Computer Science, vol 9881, Springer, Cham (2016).
8. Habibabadi, S. K., Haghighi, P. D.: Topic Modelling for Identification of Vaccine Reactions in Twitter. In: Proceedings of the Australasian Computer Science Week Multiconference (ACSW 2019), Association for Computing Machinery, New York, NY, USA, Article 31, 1–10 (2019), https://doi.org/10.1145/3290688.3290735, last accessed 2020/03/15.
9. Arslina, A., Liebenlito, M.: Sequential Topic Modelling: A Case Study on Indonesian LGBT Conversation on Twitter. In: Prime: Indonesian Journal of Pure and Applied Mathematics, 1. 10.15408/inprime.v1i1.12726 (2019).
10. Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., Orwa, C., Ochieng, S., Sambuli, N.: Topic Modelling and Event Identification from Twitter Textual Data (2016).
11. Rodina E., Dligach D.: Dictator's Instagram: personal and political narratives in a Chechen leader's social network. Caucasus Survey, 7(2), 95-109 (2019), doi: 10.1080/23761199.2019.1567145, last accessed 2020/03/15.

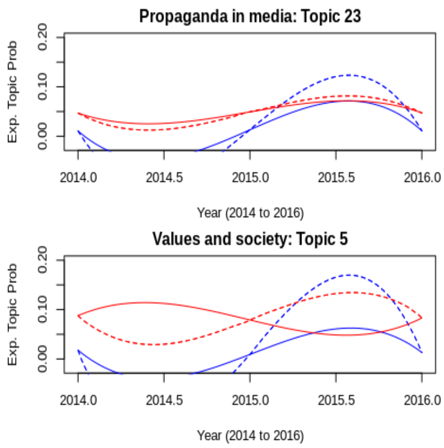
12. Bielinskienė, A., Boizou, L., Bumbulienė, I., Kovalevskaitė, J., Krilavičius, T., Mandravickaitė, J., Rimkutė, E., Vilkaitė-Lozdienė, L.: DELFI.lt corpus, Vilnius, Lithuania (2019), https://www.clarin.vdu.lt/xmlui/handle/20.500.11821/30, last accessed 2020/03/15.
13. Silge J., Robinson D. Text Mining with R– A Tidy Approach. O'Reilly Media, Sebastopol, California, USA (2017).
14. Roberts M. E., Stewart B. M., Tingley D.: Stm: An R Package for Structural Topic Models. Journal of Statistical Software 91(2), 1-40 (2019), doi: 10.18637/jss.v091.i0, last accessed 2020/03/15.
15. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2014), http://www.R-project.org/, last accessed 2020/03/15.
16. Blei, D. M., Lafferty, J. D.: Topic models. In Text mining, pp. 101-124, Chapman and Hall/CRC, Boca Raton (2009).

Annex 1


Plane crash: Topic 4
Assault/Violence: Topic 8
Road police raid: Topic 2
Government and defence: Topic 11
Missing person: Topic 14
Woman-man relations: Topic 31
Weather forecast: Topic 12
Case in court: Topic 20
Police investigation: Topic 37
Vilnius municipality: Topic 27
Budget: Topic 17
Seasons: Topic 10
Exp. Topic Prob
0.00
0.10
0.20
2014.0
2014.5
2015.0
2015.5
2016.0
Year (2014 to 2016)


Elections: Topic 21
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Work: Topic 28
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Traffic accident: Topic 39
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Education system: Topic 29
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
President's meetings: Topic 13
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
NATO in the Baltics: Topic 34
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Population changes: Topic 6
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Family and child protection: Topic 16
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Parliament: Topic 9
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Military: Topic 24
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
Ukrainian crisis: Topic 32
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)
USA and Siria: Topic 18
Exp. Topic Prob
0.00 0.10 0.20
2014.0 2014.5 2015.0 2015.5 2016.0
Year (2014 to 2016)


Newspaper article: Topic 1
Unrest in Germany: Topic 26
Land ownership: Topic 33
Sanctions on Russia: Topic 38
Finding solutions: Topic 3
Refugees in the EU: Topic 25
#storyofmylife: Topic 22
Lithuania: Topic 30
Democracy and independence: Topic 7
Food and health: Topic 36
Economics and finances: Topic 15
American business: Topic 19
Exp. Topic Prob
0.00
0.10
0.20
2014.0
2014.5
2015.0
2015.5
2016.0
Year (2014 to 2016)



Explanation:

- Blue – mainstream media portal;
- Red – unconventional media portal;
- Line -- expected probabilities;
- Dash line – sample median.