Topic Modeling for Frame Analysis of News Media

Sergey Pashakhin
HSE University
Saint-Petersburg, Russia
pashakhin@gmail.com

***Abstract*—Media frames have been traditionally extracted via manual content and discourse analysis. Such approach has a limited ability to deal with large text collections and is prone to subjectivity both in terms of text selection and interpretation. We illustrate possibilities and limitations of topic modeling for frame detection applying this method to a collection of 50,000 news items related to the Ukrainian crisis and retrieved from a Russian and a Ukrainian TV channels websites. We conclude that although topic modeling results allow to make assumptions about how topic is framed it is still not as precise as human reading of texts.**

I. Introduction

Media framing is an important issue in media studies and political science. Framing usually refers to selecting some aspects of a perceived reality and making them more salient in a message. It promotes a particular problem definition, causal interpretation, moral evaluation and/or treatment recommendation [1].

To find out how a subject is framed researchers usually use some form of manual content or discourse analysis. Generally, it is done in one of four ways. One way is simply to read texts and produce interpretations based on understanding. Another is to define a set of themes or codes, create a coding sheet and code texts by reading them. A third strategy involves software to search for keywords chosen based on the research question [2]. Finally, supervised machine learning can be used to determine which pre-defined frames are more salient than others [3]. Approaches based on researchers understanding of texts are limited by difficulty to reproduce the results and by high cost of using sufficient number of human assessors to achieve acceptable levels of inter-coder reliability. The larger a corpus is, the harder it is to use this approach. Keyword and supervised approaches are limited in their scope by the necessity to know a priori of what is worth looking in texts. However, when the frames and even issues to which they are attached are not known beforehand, inductive, that is unsupervised methods are preferable.

Finding an inductive approach for frame analysis is a non-trivial task since operationalization of a frame is difficult. Research on media framing is criticized for theoretical and empirical vagueness and is still considered to be far from being integrated into a consistent theoretical model [4]. One of the simplest inductive approach to discover frames in a corpora is word-frequency and co-occurrence analysis [5]. But for a large and heterogenous corpora, such as annual news coverage of a channel, where issues are unknown, some form of topic detection is required beforehand [6]. In this case, topic modeling is one of the most promising methods for frame analysis.

Topic modeling is inductive and allows researchers to explore data without imposing their prior knowledge on the analysis. Topic modeling detects topics in a corpus of texts in a way that closely reflects the way they are constructed by the authors [7]. Klebanov et al. propose to understand framing as a process of drawing words used to discuss an issue from a particular part of the relevant semantic field, at the expense of other parts [7]. Their work suggest that such operationalization enables to observe framing of an issue in a topic content produced by topic modeling algorithm. To our knowledge, the only application of topic modeling for frame analysis was conducted on an English language corpus [8]. That study didn't address the problem of finding stable topics in a corpora. Each new run of topic modeling algorithm on the same corpora will result in a new topic solution. This makes hard to reliably reproduce such research. In this work, we illustrate how topic modeling could be used for frame analysis on a Russian language corpus and assess the results.

II. Data

We illustrate our approach addressing the task of comparing coverage of the Ukrainian crisis by the Russian and the Ukrainian media. The research question is how the news framing of this crisis differs in the conflicting countries? We use data from the Russian TV Channel One and Ukrainian Channel Five. Both of them have been described as strongly affiliated with the political authorities in the respective countries, and both possess big audiences. We use written news available from their websites.

To capture news coverage of the major events of Ukraine crisis the time frame for data collection is defined between September 1, 2013 and October 1, 2014. This time period includes: street protests at the Maidan square in Kyiv (Euromaidan), Ukrainian presidential elections, succession of Crimea, armed conflicts in the South-East parts of Ukraine, the related international sanctions against Russia, and the crush of the Malaysian plane over the rebel territory. Transcribes of news broadcasts from this time period have been parsed from the websites. The resulting collection consists of 44,989 texts, of which 24,964 are from Channel One and 20,025 are from Channel Five. Since the channels broadcast in different languages, Channel Five texts were automatically translated into Russian. This has made joint topic modeling possible.

III. Method

The study utilizes Latent Dirichlet Allocation (LDA) with Gibbs sampling for topic modeling [9]. A comprehensive survey of directed probabilistic topic models by Daud et al. suggests that LDA is the state-of-the-art method best equipped

ISBN 978-952-68397-8-3 (PDF)

to for such task [10]. LDA requires data to be represented as a bag-of-words model. The model assumes that each document is a mixture of topics. Topics are represented by words often occurring together in the documents. The number of topics in a collection specified in advance. For this study we had chosen 100 topics based on interpretability and analytic utility as Blei and Lafferty suggest [11].

The nature of LDA is stochastic meaning that each run of the algorithm on the same collection with the same parameters will result in different topics. Research by Koltsov et al. suggest ways to solve this problem [12]. In order to find stable topics, the following strategy was used. First, five different topic modeling solutions were obtained. Then topic similarity of each pair of topics was estimated as defined in [13]: topics with the similarity value of more than 90% were considered the same topic. If a topic repeated itself in the three of five runs, it was considered stable and, therefore, really existing.

IV. Results

Topic modeling on combined collection of the channels produced 49 stable topics. Each topic was labeled manually based on the top twenty most relevant terms as well as texts with the highest probability in the topic.

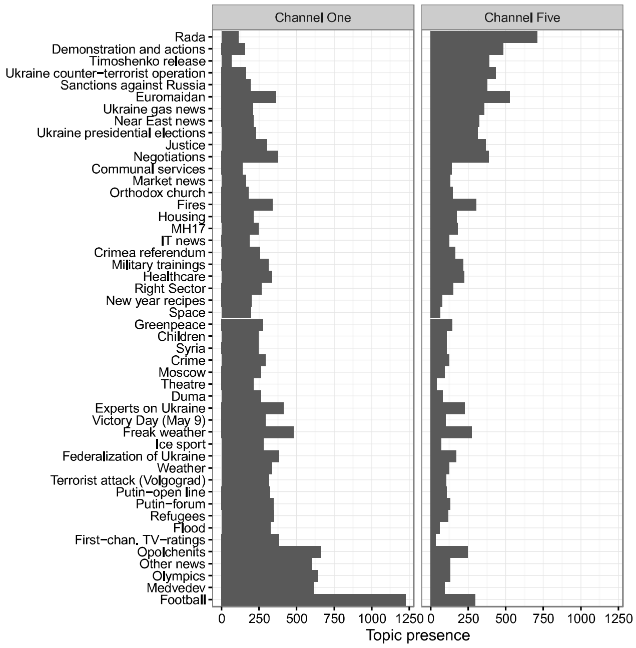

Fig. 1. Topic distribution by channel

Russian Channel One reported considerably less about the release of the Ukrainian oppositional politician Julia Timoshenko, sanctions against Russia, gas supply and Ukrainian presidential elections. Ukrainian Channel Five gave more attention to Euromaidan. Results show that news about armed conflicts in South-East Ukraine were framed so differently that the algorithm yielded two separate topics. One narrated the vision of rebels (federalization supporters); the other termed the event anti-terrorist operation (ATO) a term officially used by the Ukrainian government. This could be seen in Table I.

TABLE I. South-East Ukrainian armed conflict representation by different channels

Channel 1 (Russia)	Channel 5 (Ukraine)
ukrainian	Military
Guerrilla	Fighter
Town	Platoon
slavyansk	Ukrainian
Donetsk	Solder
Military	Serviceman
bombing	army
District	ukraine
Lugansk	power
Silovik	commander
Citizen	ATO
Army	Service
Fight	arm
Fire	defense
Peaceful	brigade
Donetsk (adj)	east
power	zone
wounded	officer
missile	operation
report	unit

Topic content shows that Channel One reports armed conflicts in East Ukraine as a war. It is suggested by the usage of such words as military, army, bombing, fire, missile and wounded. This topic contains some key actors of these stories: citizen, silovik and guerrilla (militiamen). Channel Five topic does not suggest any war meanings; it is framed as an operation against terrorists.

Reading texts with high probability in these topics confirms frames discovered by LDA. Channel One narrates armed conflicts in East Ukraine as a war waged by Ukrainian government on its own people and met with resistance from ordinary people who are forced to get armed. Channel Five coverage resembles crime reports rather than war news. It is framed as a series of actions by which the government forces eliminate criminals (terrorists).

V. Conclusion and limitations

Topic modeling showed promising preliminary results. LDA is able to grasp the contrasting difference in the coverage of the Ukrainian crisis by two different sources. While manual methods have been often criticized for subjectivity and suspected of political biases held by researchers, topic modeling objectifies intuitive human findings and liberates frame analysis from such criticism. However, there are some limitations. Fist, topic labeling has been done by one human assessor and cannot be considered reliable; at least one more assessor has to be involved. Second, to answer how exactly coverage differs across channels, discovering stable topics is not enough. Although topic modeling results allow to make assumptions about frames, it does not substitute further reading by humans. Third, as we have seen, the algorithm finds issues, but only for one issue it has shown the ability to differentiate between two distinct frames. We assume that it is not due to absence of difference in the coverage of such issues as refugees, but due to inability of topic modeling to capture these differences. Supposedly, iterative semi-supervised approaches could be a solution. Is is also interesting to examine how machine translation affects the results by reproducing the work with Russian corpus translated in Ukrainian.

Acknowledgment

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2016. We thank HSE Internet Studies Lab and its director Olessia Koltsova for resources and crucial help which made this work possible.

References

[1] R. M. Entman, “Framing: Toward clarification of a fractured paradigm”, *Journal of communication*, vol.43(4), 1993, pp. 51–58.

[2] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology (2nd ed.).* Thousand Oaks, CA: Sage, 2004.

[3] B. Burscher, D. Odijk, R. Vliegenthart, M. Rijke & C. H. de Vreese, “Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis”, *Communication Methods and Measures*, vol. 8:3, 2014, pp. 190-206.

[4] D.A. Scheufele, “Framing as a theory of media effects”, *Journal of Communication*, vol. 49(1), 1999, pp. 103–122.

[5] C. David, C. Atun, & A. La Via, “Framing the population debate: A comparison of source and news frames in the Philippines”, *Asian Journal of Communication*, vol. 20(3), 2010, pp. 337–353.

[6] C. David , J.M. Atun , E. Fille & C. Monterola, “Finding Frames: Comparing Two Methods of Frame Analysis”, *Communication Methods and Measures*, vol. 5:4, 2011, pp. 329–351.

[7] B. B. Klebanov, D. Diermeier, & E. Beigman, “Automatic annotation of semantic fields for political science research”, *Journal of Information Technology & Politics*, vol. 5(1), 2008, pp. 95–120.

[8] P. DiMaggio, M. Nag, & D. Blei, “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding”, *Poetics*, vol. 41(6), 2013, pp. 570–606.

[9] T. Griffiths, M. Steyvers, “Finding Scientific Topics”, *Proceedings of the National Academy of Sciences*, Vol. 101, 2004, pp. 5228–5235.

[10] A. Daud, J. Li, L. Zhou, & F. Muhammad, “Knowledge discovery through directed probabilistic topic models: a survey”, *Frontiers of computer science in China*, vol. 4(2), 2013, pp. 280–301.

[11] D.M. Blei, J.D. Lafferty, “Topic models”. In A.N. Srivastava, M. Sahami, (Eds.), *Text Mining: Classification, Clustering, and Applications.* London, Taylor and Francis, pp. 7194.

[12] S. Koltsov, S. I. Nikolenko, O. Koltsova, S. Bodrunova, “Stable topic modeling for web science: Granulated LDA”, *WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference*, 2016, pp. 342–343.

[13] S. Koltcov, O. Koltsova, & S. Nikolenko, “Latent dirichlet allocation: stability and applications to studies of user-generated content”, *Proceedings of the 2014 ACM conference on Web science*, June 2014, pp. 161–165.