IFLA Section
News Media

IFLA International News Media Conference 2024

Testing an Inductive Mixed-method Computational Approach to News Frame Analysis: An analysis of Hungarian online reporting of the 2014 Russia-Ukraine conflict

Mihály Nagy
Department of Digital Humanities, Eötvös Loránd University, Budapest, Hungary
E-mail address: nagy.mihaly@btk.elte.hu

Abstract:

Computational frame analysis is a highly contested, yet widely researched area of study, promising novel approaches for developing a broader understanding of news-reporting practices. Such an approach developed by Walter and Ophir (2019) is applied to inductively discover news frames used by Hungarian online news portals during the coverage of the events in Ukraine in 2014. The approach, named Analysis of Topic Model Networks, utilises LDA topic modelling and network community detection methods for inductively identifying frame packages. The applicability of the approach to Hungarian language text is explored using the BERTopic algorithm in place of LDA. Two subcorpora are analysed from the webarchive developed by Indig et al. (2019).

Keywords: ANTMN, BERTopic, War in Ukraine, Computational Frame Analysis

INTRODUCTION

The analysis of news frames used to cover the 2014 events leading to the war in Ukraine is an important step for understanding the sequence of narratives adopted on the topic. Several outstanding projects have targeted the subject within national media ecosystems (Lichtenstein et al., 2019; Nygren et al., 2018; Siiner and L'nyavskiy-Ekelund, 2017), however, conducting such projects is a lengthy and resource-intensive process, and limiting researcher bias remains an important challenge for contemporary research (Van Gorp, 2010). To increase speed and remove bias, recent research projects have demonstrated the applicability of several computational tools to framing analysis, specifically topic modelling, a technique which utilises unsupervised machine

learning (DiMaggio et al., 2013; Klebanov et al., 2008). The following analysis is the application of a specific topic modelling approach with the aim of gaining an initial impression of the way the Hungarian online media framed the war in Ukraine in 2014. The approach uses the Analysis of Topic Model Networks (ANTMN) developed by Walter and Ophir (2019), with one major technical alteration, namely the use of the BERTopic algorithm in place of the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modelling technique used by the authors. Walter and Ophir developed their technique as an answer to a specific methodological debate arising from questions on the validity and reliability of frames (Cacciatore et al., 2016). Following the suggestion to use computational tools for inductively finding frame components based on linguistic patterns (Matthes and Kohring, 2008; Van Gorp, 2010), their three-step method strives to overcome researcher bias, by relying on linguistic patterns for automatically clustering components into framing packages (Walter and Ophir, 2019, p. 260).

From Nygren et al. (2018) we learn how four different European national news ecosystems emphasised different aspects of the Ukrainian crisis. In Poland international politics, sanctions, and Russian military engagement were the main topics; in Ukraine, the conflict, fighting in the Donbas, the military, and displaced population were in focus; in Russia fighting, civilians, international politics, and sanctions gained much attention; and in Sweden, the most important topic was the downing of flight MH17 beside mostly international topics. Regarding the Hungarian news coverage of the events in 2014, Péter Apor's (2014) apt summary describes a somewhat unaware and disconnected public media. The events leading up to the Maidan Revolution were presented as the result of a divide between "Westernizers" and "Russophiles", however, discussions turned "globalised" following the Russian military intervention in Crimea. The most salient topics were diplomacy, economics (especially fossil fuel exports from Russia), NATO, and the East-West competition. Interestingly, Apor concludes that the Ukrainian nation and its people were generally viewed as a somewhat abstract concept. Often describing the country as a post-Soviet construction, at times even remarking on its validity, Hungarian mainstream media drew very few links between Hungary and Ukraine, even though the two nations share a common border. Beyond the impact on the Hungarian energy sector, the most prevalent common issue was the Hungarian minorities living in Transcarpathia, a topic which gained traction in far right news following the annexation of Crimea by Russia (Apor, 2014).

Two overlapping, different-sized sub-corpora selected from the webarchive created by Indig et al (2020) are analysed. The themes of the two collections are the overarching coverage of the conflict in Ukraine, and the downing of Malaysia Airlines flight MH17. To explore the applicability of the BERTopic algorithm, two separate embedding processes were applied, out of which only some are presented below. For the qualitative analysis of framing packages produced by the ANTMN process, the following study relies on the five generic categories introduced by Semetko and Valkenburg (2000). Based on the above-mentioned, the hypothesis of the following study is that the Hungarian online media ecosystem framed the beginnings of the war in Ukraine in terms of international diplomacy with attention to economic consequences. The primary purpose of this article is to test the applicability of BERTopic to the ANTMN method and provide preliminary findings regarding the trends in news coverage. Generating concrete conclusions regarding the Hungarian news ecosystem is beyond the scope of this analysis.

Frame analysis and it automation

Among the several competing approaches to understanding broad patterns in the news media, frame analysis has grown to be a prolific area of study, filled with a range of divergent practices. Fifty years after the publication of Erving Goffman's (1974) book, to which, at least within the social sciences, the concept of frame analysis is widely attributed, varying interpretations have resulted in a fractured paradigm (Entman, 1993), as there are numerous theoretical and methodological differences across its different applications (Matthes, 2009). Nonetheless, Robert Entman's definition – adopted by Walter and Ophir – is a good starting point. Its premise is that framing is based on selection and salience – some aspects of reality are consciously selected and highlighted through particular devices to influence the audience (Entman, 1993). The pattern of a frame "repeatedly invokes the same objects and traits, using identical or synonymous words and symbols in a series of similar communications that are concentrated in time" (Entman et al., 2009, p. 177). Framing is often contrasted to agenda-setting, and although the two overlap, it can be said that where agenda-setting is more concerned with dictating what to think about, framing influences how the receiver should think about a certain topic (Cacciatore et al., 2016; Walter and Ophir, 2019).

Disregarding its ambiguity, there are several categorical distinctions that more or less define each research project. Methodologically two general approaches can be differentiated; a deductive approach presupposes existing frames within the source material, and an inductive approach attempts to identify possible frames during the course of analysis without any prior conceptualisation (De Vreese and Lecheler, 2012). Traditionally, both have required substantial amounts of manual coding of frames and qualitative analysis. Conceptually there are two competing approaches in the field. One is "emphasis framing" which creates frames by structuring specific arguments and information, thereby influencing the focus and ultimately the overall perception of audiences (Entman, 1993). Although comparable to agenda-setting, it is important to note that whereas agenda-setting is based on the filtering of information, emphasis framing relies on structuring information. The other competing conceptualisation is "equivalency framing". It presents different but logically equivalent words or phrases (e.g. "gains" versus "losses", "full" versus "empty") to describe an event or topic, and ultimately influences audience perceptions (Cacciatore et al., 2016).

One area that has gained much attention in recent years is the application of computational approaches (Ali and Hassan, 2022), intended to limit researcher bias and speed up analysis in comparison to qualitative methods. As Eisele et al. (2023) point out, the three main automation methods applied to frame analysis – topic modelling, keyword-assisted topic modelling, and supervised machine learning – all adopt an inductive approach, with different levels of deductive control at different parts of the analysis. The present study – adopting Walter and Ophir's framework – applies topic modelling, which is an inductive text analysis method utilising unsupervised machine learning for identifying distinct topics within a corpus (elaborated on further in the method section). Its conventional methods are by definition primarily applicable to detecting emphasis frames, as the structure of texts is not preserved in the algorithm's output.

DiMaggio et al. (2013) proposed that topic modelling can be employed to automatically detect frames in texts, thereby minimising researcher bias and lowering costs and time due to its automated process. They argued that many of the identified topics could constitute frames if they are understood as "semantic contexts that prime particular associations or interpretations of a phenomenon in a reader" (DiMaggio et al., 2013, p. 578). Walter and Ophir have raised three

reasons why this may be problematic. First, the content and scope of topics depend on the number of topics identified by the algorithm, a metric which is set by the researcher (explicitly in LDA and implicitly in BERTopic). Second, they argue that identified topics may represent either generic or context-specific frames. As subcategories of emphasis frames, generic frames are "established routines" practised by journalists to cover a story, and context-specific (or topic-specific) frames are those only relevant to individual contexts (e.g. coverage of social protests or the EU's economy) (Walter and Ophir, 2019, p. 250). Although their second argument is a relevant observation, Walter and Ophir themselves refrain from distinguishing the two types of emphasis frames as the boundary between them is often not discernable. Third, the authors argue that as topics identified by conventional topic modelling algorithms are only linguistically coherent (not semantically), and therefore, cannot be considered frames.

To overcome the limitations summarised in these three arguments Walter and Ophir added two further processing steps to help interpret the output of topic modelling algorithms. They consider topics to be framing devices – structurally located lexical choices that are used to build a frame –, and building on Van Atteveldt and Peng's (2018) reasoning, they argue that by looking at the structural relationship between framing devices we can identify frames. In this conceptualisation, frames are considered linguistic patterns – "a community in a network of topics" (Walter and Ophir, 2019, p. 248). Therefore, the first additional step is creating a semantic network out of topics to map the relationships between framing devices. In such a network the topics are the nodes and the edge between a pair of topics is calculated based on their mutual presence in documents. The second additional step involves running a community detection algorithm on the generated network to identify topic clusters which are finally interpreted as frames. Their method promises a faster and more bias-free computational framing analysis method, only requiring qualitative analysis at the topic labelling step and the final interpretation of network communities.

The corpus

The source material used for the following measurements of Hungarian online news from 2014 is the webarchive of Indig et al. (2020), created between 2019 and 2022. The archive is distinct in its preservation method because all HTML documents are parsed and cleaned using a semi-automatic rule-based process that allows the curation of both the text and the metadata content. Crawled materials are saved as WARC files, and the filtering process is archived as a TEI-XML document. This process results in metadata-rich, relatively clean data regarding the textual content of the portal articles and preserves body text hierarchies such as paragraphs. More importantly, the crawling methodology traverses all available taxonomy pages of the target portal, therefore all articles present on the given portal (at the time of crawling) are captured in the archive. The portals archived for the 2014 period include major news portals from all ideological and political leanings, and even the most popular far right portal Kuruc.info. One limitation of the archive is that not all news portals are included for 2014, specifically index.hu, the most popular portal.

The archive was filtered to the following two different scale sub-corpora: 1) the overall coverage of events in Ukraine between February (the month leading up to the Crimean crisis) and September (the month of the first Minks agreement) (n=7636), and 2) coverage of the downing of the MH17 Malaysian Airlines flight (n=767). Filtering was done through two approaches. A keyword list was

created for each sub-corpora from the keywords (often referred to as hashtags) included in the archive that were annotated on the original HTML pages, and another keyword list was created from the results of a named entity recognition[1] (NER) processing of all articles between February and September (see Appendix 1. for both lists and filtered portals). Documents were chosen where any of the annotated keywords were present in the first list or where over 50% of paragraphs contained at least one keyword from the NER list. The NER solution was required as not all portals had annotated keywords. Using the Spacy Sentencizer function[2] the body text of articles was split into individual sentences and their language was identified with the Python langdetect algorithm[3]. Finally, only Hungarian sentences were selected from each article, ensuring that no foreign language or unintelligible text was included[4].

Method: ANTMN with BERTopic

The LDA topic modelling algorithm used by Walter and Ophir is a generative statistical model which describes each document as a "bag-of-words" and models them as "random mixtures over latent topics, where each topic is characterised by a distribution over words" (Blei et al., 2003, p. 996). It is demonstrated to work well as part of the ANTMN workflow, however, as argued by Egger and Yu (2022) it has become the least advanced method available among contemporary topic modelling algorithms. For Hungarian language especially, optimal use of LDA requires considerable text filtering, including case conversion, removal of punctuations, and lemmatization, as demonstrated by the work of Gelányi et al. (2022). Moreover, as Grootendorst (2022) points out, by reducing a document to a collection of its words, LDA fails to account for the context of words within a sentence. This is the primary aspect that BERTopic provides a solution for.

Embeddings

The first step in BERTopic is converting each document to an embedding representation using a Bidirectional Encoder Representations from Transformers (BERT) language model. Embeddings are generated with the Sentence-BERT framework that can use pre-trained models to generate "semantically meaningful sentence embeddings" (Reimers and Gurevych, 2019). The pre-trained model used[5] for this study was developed by Osváth et al. (2023) specifically for the BERTopic algorithm on Hungarian text. As the Sentence Transformers library is designed to embed short texts (sentences or paragraphs) two approaches were tested. First, embeddings were calculated for sentences generated during the filtering stage with the intention of aggregating their results later based on shared documents, and second, only the lead paragraphs of each article were embedded.

Topic modelling

The BERTopic algorithm works as a pipeline of customisable operations. Following the calculation of embeddings, dimensionality reduction of vectors was done with the UMAP technique (McInnes et al., 2018), where the number of neighbouring sample points was set to 15. For clustering documents into similar groups, the HDBSCAN algorithm (McInnes et al., 2017) was used. The most important metric in BERTopic is the minimum size of a cluster which

[1] The NER analysis was done using HuSpaCy, see: https://spacy.io/universe/project/HuSpaCy

[2] See https://spacy.io/api/sentencizer

[3] See https://pypi.org/project/langdetect/

[4] All used corpus data is available on Zenodo.org: https://zenodo.org/communities/elte-dh/records?q=&l=list&p=1&s=10&sort=newest

[5] See: https://huggingface.co/NYTK/sentence-transformers-experimental-hubert-hungarian

influences the number of clusters. Silhouette scores (Rousseeuw, 1987) and Calinski and Harabasz scores (Caliński and Harabasz, 1974) were calculated on a range of minimum cluster size metrics between 15 and 50 for each sub-corpora, and the final metric was chosen by considering both scores and qualitative analysis of topic distributions. For generating topic representations, the default "custom class-based variation of TF-IDF" (Grootendorst, 2022) was used with a custom CountVectorizer function of the Scikit-learn library (Pedregosa et al., 2011) set up with a list of Hungarian stop-words.

Network and Community detection

ANTMN originally creates a network by calculating the pairwise cosine similarity between topics, taken from the theta matrix which represents a distribution of topics over documents (Walter and Ophir, 2019, p. 255). BERTopic also returns a matrix of topics over documents, which can be understood more as the set of probabilities that a document belongs to a certain cluster. Although the two matrices represent different things, they can be used in similar ways. The assumption of this study is that the ANTMN pairwise network construction method can be used with the probability matrix of the BERTopic algorithm to estimate the tendency of topics to appear in the same documents. Therefore, the fully connected networks were created by calculating the cosine similarity between all topic pairs where the two vectors were the topic membership probabilities for each document. For the separate sentence embedding method the topic probability matrices split documents into many parts, therefore, the maximum probability from sentences belonging to the same document was taken as the probability for a given document. Averaging the probabilities was also considered, however, it yielded inconsistent results. As suggested by Walter and Ophir, five community detection algorithms were trialled on each network. Namely, the "Louvain" (Blondel et al., 2008), "Walktrap" (Pons and Latapy, 2005), "Spinglass" (Reichardt and Bornholdt, 2006), "Fast-Greedy" (Clauset et al., 2004), and "Leading Eigenvector" (Newman, 2006) algorithms[6].

Results

For the first sub-corpora, containing all articles on Ukraine, the analysis of lead paragraphs yielded interpretable communities, however, the sentence-based embeddings approach resulted in hundreds of topics, most of which were highly specific and required manual coding. For the latter approaches the most appropriate minimum HDBSCAN cluster size was 30, where higher numbers resulted in disintegration of topic distributions, and lower numbers resulted in a higher number of topics. As no resources were available for the manual coding of the resulting topics – also required for meaningful topic number reduction – further analysis was not pursued. Theoretically, a large number of highly specific topics aligns well with the context-specific approach of emphasis framing, however, from the perspective of speeding up initial analysis of a corpora, it is overly resource-intensive. The analysis of lead paragraphs, however, had (admittedly) more convenient results. With the optimal minimum HDBSCAN cluster size of 15, 59 topics were generated, out of which 5 were discarded as unintelligible. Both Louvain, Eigenvector, and Spinglass community detection algorithms resulted in three very similar clusters, however, the Walktrap and Fast-Greedy communities were drastically different. Here the Louvain community detection results are

[6] See code used for full methodology: https://github.com/everybitmihaly/ANTMN_with_BERTopic

analysed. As can be seen in *Figure 1.* the light blue community appears to focus on international bodies such as states (*Germany*, *Russian State Duma*) or organisations (*IMF*, *NATO*) and also contains topics such as *EU Association Agreement*, *UKR-RUS Talks*, and *EU Sanctions* which address concepts on an international scale. The red community contains more specific, human- and conflic-focused topics (*Civilian casualties*, *Ukrainian Military casualties*, *Protest casualties*), and fighting related topics (*Shelling*, *Breaking ceasefire*, and *Naval clashes)*. Finally, in the yellow community the *Gas* and *UKR Economics* topics are related to economic impact, and the *RUS-UKR Tensions*, *Russian invasion*, and *Military capabilities* topics cover the military conflict.

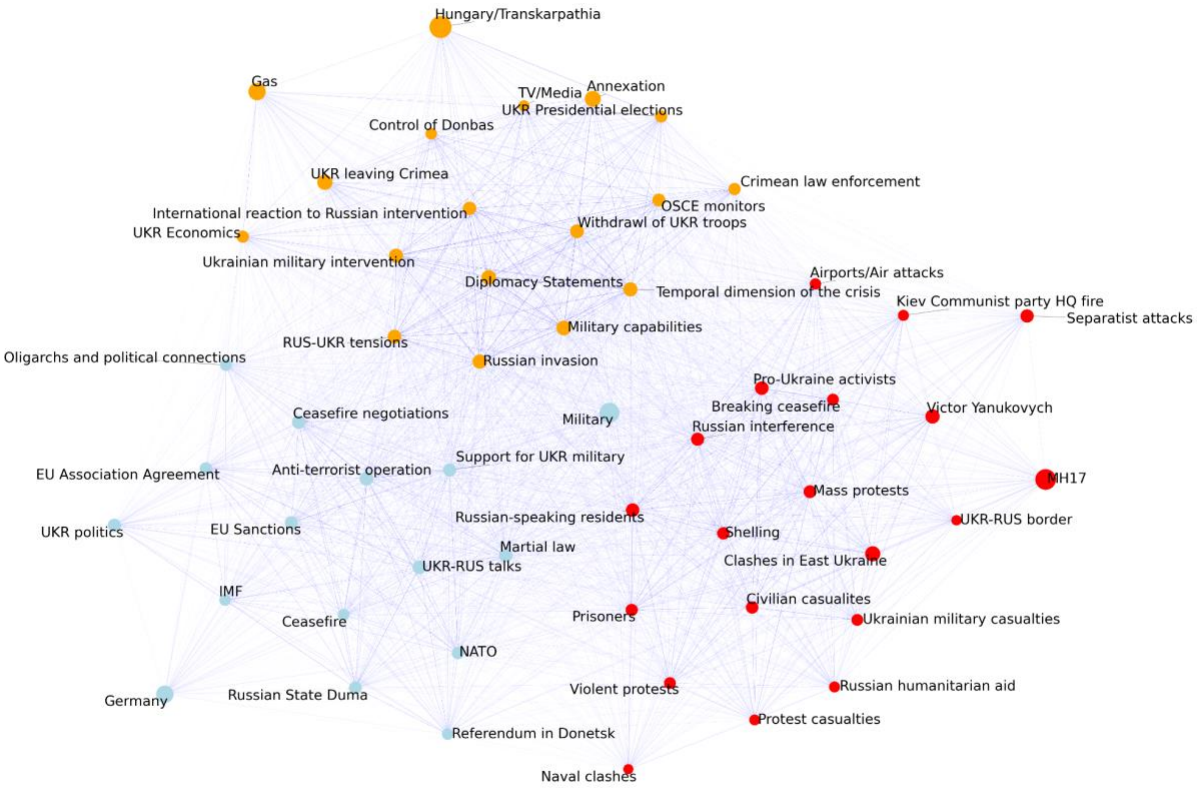

Figure 1. *Weighted,* fully connected, undirected network of topics generated from lead paragraphs of articles covering events in Ukraine in 2014 (Feburary to September). Nodes represent topics generated using the BERTopic algorithm, edges represent the co-occurrence of topics in documents. Colour is determined by Louvain community membership. Node size represents the sum probabilities of a topic.

For the MH17 sub-corpora, both lead paragraph embeddings and sentence embeddings resulted in an interpretable network. For the latter (*Figure 2.*) a minimum HDBSCAN cluster size of 25 returned 78 topics, out of which 7 were considered unintelligible. Taking the average of probabilities resulted in an uninterpretable network, however, the maximum method yielded

clearer results. It should be noted that with further work the topics could most likely be joined and refined to create a smaller and more concise network. Similarly to the previous example, community detection algorithms excluding Fast-Greedy and Walktrap resulted in four mostly similar clusters, therefore, the Louvain communities are presented in *Figure 2*. The red community contains several topics related to the victims' bodies and also the *Human rights law* topic. The grey community contains topics related to economics (*Sanctions*, *Foreign Exchange Market*, *Gazprom*) and diplomatic elements, such as *Blame from UKR/RUS* which contains state official's remarks on the cause of the crisis from both countries and *Act of War* which contains content qualifying Russia's actions. The lightblue community appears to cover the crash itself and contains several topics related to the passengers, but also many other topics, making its interpretation difficult. Finally, the yellow community contains topics relating to the victims and the human element, but also the conflict and the events around the crash itself.

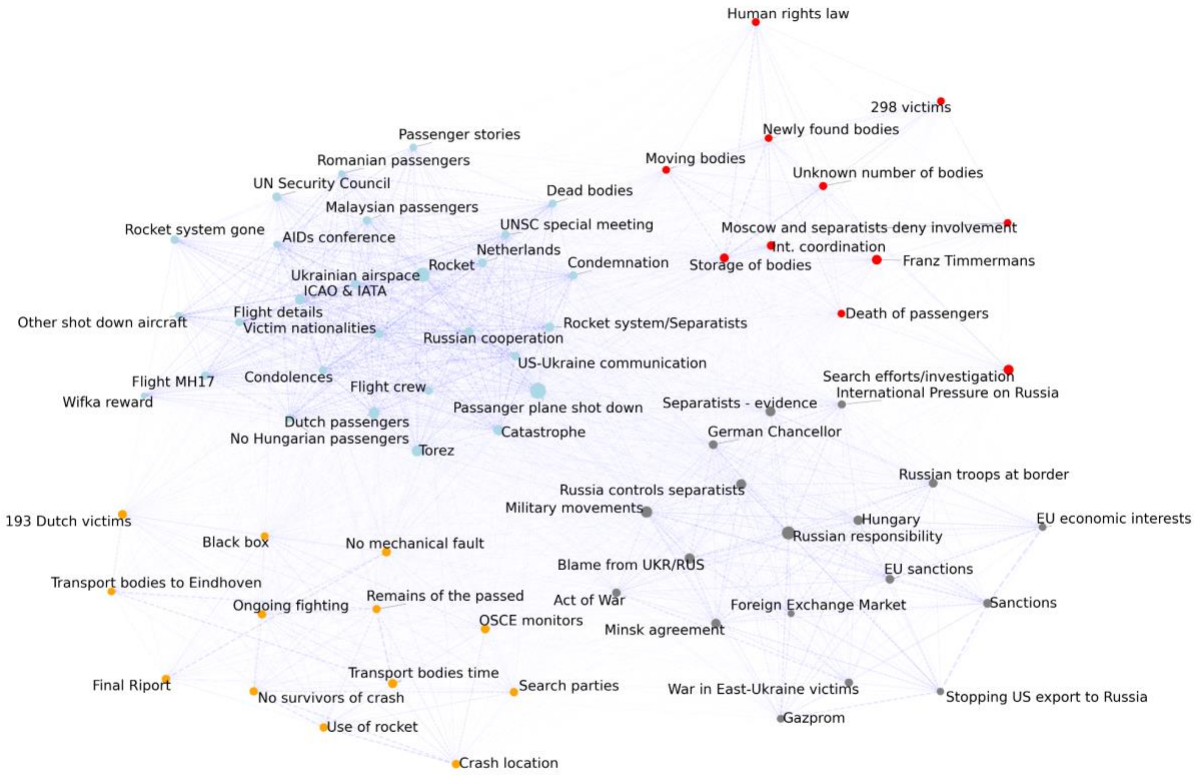

Figure 2. *Weighted,* fully connected, undirected network of topics generated from all sentences of articles covering the downing of Malaysian Airline flight MH17. Nodes represent topics generated using the BERTopic algorithm, edges represent the co-occurrence of topics in documents. Colour is determined by Louvain community membership. Node size represents the sum probabilities of a topic.

The results of both examples were analysed according to the political ideologies (left, right, centre, far-right) and pro- versus anti-government leanings of news portals (Table 1. and Table 2.). Probabilities were added up for each document within a category and normalised by the category size (for categorization see Appendix 2.).

	light blue	**orange**	**red**
left	0.269	0.207	0.222
right	0.254	0.202	0.255
centre	0.258	0.22	0.244
far-right	0.235	0.201	0.242
pro-gov	0.256	0.206	0.259
anti-gov	0.259	0.208	0.236

Table 1. Normalised distribution of Louvain topic clusters across ideological alignment (lightgrey) and political alignment (dark grey) for the lead paragraph based topic network generated from the general Ukraine events subcorpus.

	grey	**light blue**	**orange**	**red**
left	1.468	0.527	0.69	1.188
right	1.655	0.558	0.618	1.132
centre	1.641	0.522	0.648	1.355
far-right	1.393	0.58	0.778	1.097
pro-gov	1.646	0.521	0.675	1.192
anti-gov	1.58	0.553	0.639	1.202

Table 1. Normalised distribution of Louvain topic clusters across ideological alignment (lightgrey) and political alignment (dark grey) for the sentence-based topic network generated from the MH17 subcorpus.

Discussion and conclusions

In the first network of Ukrainian events, the light blue community could be indicative of the responsibility frame, as global actors are often positioned as those solving aspects of the conflict through aid (*IMF*) or sanctions (*EU Sanctions*). The topics of the red community point to a human interest frame in terms of a humanitarian perspective, and the *Shelling, Breaking ceasefire,* and *Naval clashes* indicate a conflict frame. Finally, the *Gas* and *UKR Economics* topics in the yellow community could be indicative of an economics frame, however, *RUS-UKR Tensions, Russian invasion,* and *Military capabilities* topics show that most of the conflict frames are present in this community. In the MH17 example, the most apparent frame comes through in the red community,

which contains topics serving as evidence of the human interest frame. The economics and international relations topics of the grey community are indicative of both an economic consequences frame and a responsibility frame, especially sentences within the *Russia control's separatists* topic that clearly question Russia's involvement. The two examples partially corroborate Péter Apor's (2014) observations about the Hungarian media's framing of the Ukrainian crisis. The international diplomacy and economic consequences frames align with Apor's description of the media's focus on globalised discussions.

Regarding the distribution of topic probabilities across news portals, there emerged no clear trends. Topics appear to be evenly presented by all portals, which could be explained by the nature of online news reporting and the homogeneous information flow within Hungarian media in 2014. The majority of articles on the events in Ukraine are short and usually act as an update regarding new developments. Repetition is high as most of the incoming information was taken from the only Hungarian news agency at the time, the *Magyar Távirati Iroda* ("MTI," 2009).

As embeddings capture the semantic similarity of different texts, using BERTopic by itself for frame analysis could theoretically address Walter and Ophir's argument against the idea of interpreting topics as frames suggested by DiMaggio et al. (2013). From a methodological perspective, the results above demonstrate that the probability matrix of BERTopic appears to be interchangeable with the LDA theta matrix in the ANTMN method. Regarding the automation of framing analysis, preliminary findings may be easier to attain, however thorough framing analysis would still require independent topic encoders and expert knowledge of the source material.

References

Ali, M., Hassan, N., 2022. A Survey of Computational Framing Analysis Approaches, in: Goldberg, Y., Kozareva, Z., Zhang, Y. (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 9335–9348. https://doi.org/10.18653/v1/2022.emnlp-main.633

Apor, P., 2014. Hungary-The Ukrainian Crisis in the Hungarian Media. Presented at the Cultures of History Forum, Imre Kertész Kolleg.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008.

Cacciatore, M.A., Scheufele, D.A., Iyengar, S., 2016. The End of Framing as we Know it … and the Future of Media Effects. Mass Commun. Soc. 19, 7–23. https://doi.org/10.1080/15205436.2015.1068811

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat.-Theory Methods 3, 1–27.

Clauset, A., Newman, M.E., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E 70, 066111.

De Vreese, C.H., Lecheler, S., 2012. News framing research: An overview and new developments. SAGE Handb. Polit. Commun. 292–306.

DiMaggio, P., Nag, M., Blei, D., 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding.

Poetics 41, 570–606.
Egger, R., Yu, J., 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Front. Sociol. 7, 886498.
Eisele, O., Heidenreich, T., Litvyak, O., Boomgaarden, H.G., 2023. Capturing a news frame–comparing machine-learning approaches to frame analysis with different degrees of supervision. Commun. Methods Meas. 17, 205–226.
Entman, R.M., 1993. Framing: Toward clarification of a fractured paradigm. J. Commun. 43, 51–58.
Entman, R.M., Matthes, J., Pellicano, L., 2009. Nature, sources, and effects of news framing, in: The Handbook of Journalism Studies. Routledge, pp. 195–210.
Gelányi, P., Sebők, M., Ring, O., 2022. A topikmodellezés lehetőségei és korlátai egy törvénykorpusz példáján= The opportunities and constraints of topic modelling–the case of a corpus of laws. STATISZTIKAI Szle. 100, 783–814.
Goffman, E., 1974. Frame analysis: An essay on the organization of experience. Harvard University Press.
Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv Prepr. ArXiv220305794.
Indig, B., Knap, A., Sárközi-Lindner, Z., Timári, M., Palkó, G., 2020. The ELTE. DH Pilot Corpus–Creating a Handcrafted Gigaword Web Corpus with Metadata, in: Proceedings of the 12th Web as Corpus Workshop. pp. 33–41.
Klebanov, B.B., Diermeier, D., Beigman, E., 2008. Automatic annotation of semantic fields for political science research. J. Inf. Technol. Polit. 5, 95–120.
Lichtenstein, D., Esau, K., Pavlova, L., Osipov, D., Argylov, N., 2019. Framing the Ukraine crisis: A comparison between talk show debates in Russian and German television. Int. Commun. Gaz. 81, 66–88.
Matthes, J., 2009. What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. Journal. Mass Commun. Q. 86, 349–367.
Matthes, J., Kohring, M., 2008. The content analysis of media frames: Toward improving reliability and validity. J. Commun. 58, 258–279.
McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. J Open Source Softw 2, 205.
McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv Prepr. ArXiv180203426.
MTI [WWW Document], 2009. URL https://web.archive.org/web/20090205004311/http://mti.hu/fixcikk/124 (accessed 5.14.24).
Newman, M.E., 2006. Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104.
Nygren, G., Glowacki, M., Hök, J., Kiria, I., Orlova, D., Taradai, D., 2018. Journalism in the crossfire: Media coverage of the war in Ukraine in 2014. Journal. Stud. 19, 1059–1078.
Osváth, M., Yang, Z.G., Kósa, K., 2023. Analyzing Narratives of Patient Experiences: A BERT Topic Modeling Approach. Acta Polytech. Hung. 20, 153–171. https://doi.org/10.12700/APH.20.7.2023.7.9
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks. Presented at the Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20, Springer, pp. 284–293.
Reichardt, J., Bornholdt, S., 2006. Statistical mechanics of community detection. Phys. Rev. E 74, 016110.
Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. ArXiv Prepr. ArXiv190810084.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Semetko, H.A., Valkenburg, P.M., 2000. Framing European politics: A content analysis of press and television news. J. Commun. 50, 93–109.
Siiner, M., L'nyavskiy-Ekelund, S., 2017. Priming language political issues as issues of state security: A corpus-assisted discourse analysis of language ideological debates in Estonian media before and after the Ukrainian crisis. Lang. Policy State 25–44.
Van Atteveldt, W., Peng, T.-Q., 2018. When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. Commun. Methods Meas. 12, 81–92.
Van Gorp, B., 2010. Strategies to take subjectivity out of framing analysis, in: Doing News Framing Analysis. Routledge, pp. 100–125.
Walter, D., Ophir, Y., 2019. News frame analysis: An inductive mixed-method computational approach. Commun. Methods Meas. 13, 248–266.

Appendix 1.

Subcorpus 1: Coverage of events in Ukraine 2014 February to September

Annotated keyword list:

Ukrán Biztonsági Tanács, ukrán édesség, ukrán kamion, ukrán himnusz, ukrán zavargás, Krími Autonóm Köztársaság, ukrán elhárítás, ukrán szociálpolitikai miniszter, Oroszország Ukrajna, ukrán hadsereg, g8 ukrajna, ukrajnai tűzszünet, Ukrajna, krími parlament, ukrán erők, ukrán nyelvtörvény, Kijevi tüntetés, ukrán titkosszolgálat, ukrán elnökválasztás, orosz-ukrán területi viták, Magyar-Ukrán Társaság, krími válság, ukrán vállalatok, Ukraja, Sahtar Donyeck, ukrán gazdasági miniszter, ukrán segély, összecsapások ukrajnában, Krím, ukrán-orosz válság, megtámadtak egy ukrán hajót, ukrán ellenzék, ukrán főügyész, Donyecki Népköztársaság, Ukrajna Független Média Szakszervezete, donyecki repülőtér, tüntetések ukrajnában, orosz-ukrán gázvita, krím-vélsziget, ukrán-válság, kijevi megállapodás, ukránkérdés, ukrán szakadárok, donyecki lázadók, ukrajnai szakadárok, Kelet-Ukrajna, ukrán tisztek, ukrán pilóta, ukrajna.ausztrália, ukrán zászló, ukrán ellenzéki, orosz ügynökök ukrajnában, ukrajna feldarabolása, kelet-ukrajnai szakadárok, krími népszavazás, ukrajnai elnökválasztás, Ukrán válság, orosz-ukrán, kijev, ukrajnai tüntetések, Ukrajna föderalizációja, ukrán turista, ukrán cigarettacsempész, ukrán vásag, ukrán pilótanő, ukrán légtér, kijevi hírportál, Donyeck, donyecki szakadárok, ukrán tüntetések, ukrán katona halála, shaktar donyeck, Dinamo Kijev, ukrajnai konfliktus, ukrán terület, ukrán védelmi miniszter, ukrajna, kijevi nagykövet, ukrán tanár, Ukrán Biztonsági Szolgálat, krími helyzet, krími blokád, ukrán légtérzár, ukrán nagykövet, ukrán tüntetés, ukrán nacionalisták, ukrán, Délkelet-Ukrajna, Krími félsziget, orosz-ukrán konfliktus, Ukrán Nemzeti Bank, Krím félsziget, ukrán választás, krími katonák, Ukrán Nemzetbiztonsági és Védelmi Tanács, ukrán befektetők, krím, Ukrán bíróság, ukrán válság, dél-ukrajna, ukrán miniszterelnök, ukrán vezeték, ukrán újságíró, ukrán gyermekpornó, donbasz, Szeverodonyeck, ukrán vezetés, Donyeck megye, orosz–ukrán konfliktus, krími referendum, ukrajnai zavargások, ukrán szeparatisták, ukrán elnök, Ukrán Népköztársaság, Donyeck-medence, Ukrajnai válság, krími konfliktus, ukrán–orosz válság, ukránpárti tüntetők, krími-félsziget, dinamo kijev, magyar-ukrán határ, Donyecki Köztársaság, ukrán-magyar, ukrán tűzszünet, donyeck, ukrajna. kelet-ukrajna, ukrán-orosz határ, ukrajnai magyarok, Ukrán Országos Önkormányzat, ukrajnai válság, ukrán hadművelet, ukrán tünetések, orosz-ukrán háború, donyecki köztársaság, ukrán tütnetések, MTSZ Ukrajina, Krími népszavazás, ukrán nyelv, ukrán parlament, kijevi tüntetések, EU-ukrán szabadkereskedelmi megállapodás, Ukrán Nemzeti Gárda, ukrán hadihajó, ukrán politika, ukránok, ukrán helyzet, "donyecki állam", krími arany, donyecki reptér, ukrán erőszak, ukrajinszka pravda, ukrán külügyminiszter, krími légtér, ukrán válogatott, ukrán katonák, Kijev, kijevi parlament, Krími Köztársaság, krím-félsziget, ukrán-orosz viszony, Ukrajnai Magyar Demokrata Szövetség, krími aranykincsek, kelet-ukrán, ukrán határőr, ukrajnai románok, volt ukrán

miniszterelnök, orosz-ukrán gázmegállapodás, Krím-félsziget, ukrán-magyar banda, ukrán köztévé, ukrán menekültek, donyecki népszavazás, Donbassz Aréna, ukrán flotta, ukrajnai népszavazások, donyecke, orosz-ukrán határ, ukrán kormányfő, ukrán határ, krími tatárok, Ukrán ENSZ-nagykövet, ukrán lázadók, kelett-ukrajna, Krími válság, ukrán-orosz konfliktus, kelet-ukrajna

NER keywords: Putyin, Poroshenko, Janukovics, Kijev, Krím, Donyeck, Luhanszk, Ukrajna,

Portal distribution:

'24.hu - Rangadó: 11', '444: 1010', 'HVG: 1882', 'Heti Válasz hetilap: 234', 'Kuruc.info: 705', 'Magyar Narancs: 58', 'Magyar Nemzet – Polgári Napilap: 298', 'Maszol: 322', 'Népszava: 1429', 'Origo: 856', 'VS: 155', 'alfahir: 676'

Subcorpus 2: Coverage of the downing of Malaysian Airlines flight MH17

Annotated keyword list:

Lelőtt maláj gép, maláj gép, maláj utasszállító, eltűnt maláj gép, maláj járat, maláj légitársaság, eltűnt maláj repülőgép, lelőtt maláj gép, maláj repülőgép, hova tűnt a maláj repülőgép, a malájziai gép tragédiája, maláj repülőgépkatasztrófa, maláj hatóságok, MH17, MH-17

NER keywords: Maláj Légitársaság, Malajzia Airlines, MH17, MH-17

Portal distribution:

'24.hu - Rangadó': 10, '444': 65, 'HVG': 165, 'Heti Válasz hetilap': 44, 'Kuruc.info': 44, 'Magyar Narancs': 9, 'Magyar Nemzet – Polgári Napilap': 53, 'Maszol': 65, 'Népszava': 88, 'Origo': 113, 'The Budapest Beacon': 6, 'Transindex': 37, 'VS': 67, 'alfahir': 48

Appendix 2.

Pro-government / opposition groups of portals:

Vadhajtasok, Origo, Magyar Nemzet – Polgári Napilap, VS

Abcúg, alfahir, Maszol, 24.hu - Isten tudja, 444, Heti Válasz hetilap, HVG, 24.hu - Rangadó, Kuruc.info, Transindex, The Budapest Beacon, 24.hu - Roboraptor, Mérce, Magyar Narancs, Népszava

Ideological groups

Left: Abcúg, 444, Mérce, Magyar Narancs, Népszava
Right: alfahir, Maszol, Vadhajtasok, Heti Válasz hetilap, Origo, Magyar Nemzet – Polgári Napilap, VS
Centre: 24.hu - Isten tudja, HVG, 24.hu - Rangadó, Transindex, The Budapest Beacon, 24.hu - Roboraptor
Far right: Kuruc.info

Categories were judged based on each portal's wikipedia description available at: https://hu.wikipedia.org/wiki/Kateg%C3%B3ria:Magyarorsz%C3%A1gi_internetes_sajt%C3%B3term%C3%A9kek