Doctoral School of Earth Sciences

# Spatiotemporal and multilingual Semantic Machine Learning Analysis of Social Media Data for the recent protests in Europe – based on Twitter data –

Main findings of the PhD dissertation

Tamás Kovács

Supervisors:
Dr. Ádám Németh
Prof. Dr. Klára Siposné Kecskeméthy

University of Pécs
Faculty of Sciences
Pécs, 2022

## I. INTRODUCTION

The perception of inherent tensions between justice and injustice (or the disproportion of good and bad) often press a group of people (or even the whole society) to seek change concerning politics and power, for example in the form of protests. In the last few decades, the advent and rapid expansion of internet-based communication technologies transformed the way of seeking change through connective action, where besides the two main elements—the people and their intentions—the role of the information along with its spread and accessibility gained more and more significance. Social media platforms such as Twitter are considered a new mediator of collective action, in which various forms of civil movements unite around public posts, often using a common hashtag, thereby strengthening the movements.

The data-driven analitical approach, relying on social media posts and activities, has many strengths—especially considering its high temporal resolution and rapid user-response to certain news and information. Twitter data serves as a unique and useful source of information for the analysis of civil movements, as the analysis can reveal important patterns in terms of spatiotemporal and sentimental aspects, which may also help to understand protest escalation over space and time. The investigation of social media in the case of events such as the murder or a protests in Belarus seems an important tool to track and understand the immediate reaction of people, unlike any other method or source of information.

The methodological workflow developed in this doctoral research combines time series clustering with semantic topic modeling and sentiment analysis, performed on georeferenced social media data, which provides multi-modal insights into the public's reactions to a specific political event. The proposed approach includes multi-lingual corpus translation, as well as location and sentiment extraction, using machine-learning topic modelling methods to reveal the hidden interests and motivators of collective action. Through this, the approach has a distinct advantage over the prior investigations that primarily focused either on hashtag-activism (ignoring the spatial dimensions) or, on the contrary, using only location-specific hashtags. Whereas by applying machine learning algorithms and techniques that are almost entirely automatable, the analysis can cover a much wider range of input data than existing studies, where the researchers solely evaluate posts manually. Overall, with this mixed-method approach, this work overcome the limitations of contemporary research on social movements that mainly focuse one one language and a restricted area.

The social media data analyzed in this dissertation were obtained using the Twitter Streaming Application Programming Interface (API), the US-based social networking and microblogging service. The first dataset's starting date is adjusted to the first official report of the murder of Ján Kuciak while the final day is adapted to the earliest statement of the resignation of Prime Minister, Robert Fico (26 February and 15 March 2018). The second dataset's starting date is adjusted to the day of the Belarus presidential election while the final day is adapted to the formally

inauguration of the president of Belarus, Alexander Lukashenko (9 August and 23 September 2020). Both datasets consists of the content of the tweets and additional attributes such as user name, user location, and the timestamp when the tweet was posted.

## II. RESEARCH OBJECTIVES AND GOALS

This dissertation suggests a comprehensive methodology to overcome the aforementioned limitations in the existing methods and handle the complexity of protest analyses. This work examines the similarities and correlations of spatial, temporal, and sentimental markers of Twitter data by developing a new data-driven combined approach investigating two distinct East European protest movements. First, the European influence of a Slovakian journalist's Ján Kuciak and his fiancée Martina Kusnirova assassination in 2018. Second, the influence of the Belarusian presidental election in 2020, which was not analyzed by the contemporary research.

The multi-spectral interpretation of the dynamic and challenging nature of the events requires an efficient analytical method to assist a more comprehensive understanding of the protest dynamics. Thus, this work goes beyond the state of the art in two distinct ways. First, it demonstrates how georeferenced social media data can be used for analyzing political events, even at a smaller spatial and societal scale and in unique non-English languages. Second, from a methodological viewpoint, the present work proposes a new algorithmic workflow that combines time-series clustering with semantic topic modelling and sentiment analyses on georeferenced social media data. Moreover, this research will put effort into deeply understanding the relevant research of the past decade.

By presenting an original perspective and considering the presented limitations of existing analyses, in this dissertation we intend to answer the following research questions:

1. Temporal and spatial aspects:
    1. How tweeting activity related to the murder of Kuciak varied over time throughout Europe? (RQ1a)
    2. How tweeting activity related to the presidential election of Belarus varied over time throughout Europe? (RQ1b)
    3. Can we identify the influence of specific events and incidents, such as media reports or findings of the investigation based on this tweeting activity? (RQ1c)

2. Content aspects:
    1. How does the sentiment of the tweets vary over time, and how does it relate to specific events and news? (RQ2a)
    2. Around what topics do the tweets revolve, beyond the murder itself?

(RQ2b)

3. Profiles:
    1. How can we characterize the countries based on the temporal aspects of the tweeting activity and the sentiment of the tweets? (RQ3a)
    2. Does the categorization of the countries also reflect differences in the identified topics or the changes of the sentiment values over time? (RQ3b)

## III. METHODOLOGY

The data pre-processing approach (performed on the raw tweets) consists of a thorough text cleaning workflow and the transformation of available meaningful location information to coordinates for further utilization in the spatial analysis step. In the first part of the pre-processing, we implement primary filtering on our dataset to ignore short tweets that hardly bear any semantic significance. Moreover, stop-words, rare, and too frequent words are removed to normalize the dataset and reduce the redundant noise from tweets. Then, we use the available location information of the user to localize the tweets without direct coordinates attached to a tweet (also called geoparsing) to further increase the size of the dataset for machine learning-based translation and spatiotemporal analysis.

### Preliminary Text Cleaning

The aim of this step is to increase the efficiency of the subsequent translation process. We remove short tweets (containing a single word), or those posts that contained only hashtags or URLs, because they hold unclear and hardly interpretable semantic value. Then, we remove replies (@user_name) from the text as that is considered as unnecessary noise in the analysis; in contrast, the hashtags are preserved but their sign (#) is removed. The consideration behind this step is that users tend to use hashtags as an integral part of the syntax. As a final cleaning step, we remove new line characters as well as additional whitespaces from the tweets.

### Locating Tweets Using Coordinates and User Profile Information

In general, the tweets that inherently include coordinates constitute only a small subset (1-2%) of all tweets. To overcome this limitation our analytical approach tried to locate those tweets that had no coordinates using location information available in the users' profile (account location and user description).

To increase the amount of tweets having some kind of spatial reference, we can extract location information from the profile of the Twitter users, which information can serve as a proxy to where this user might be active most of the time. To obtain useful information on user_location data, first we ranked individual user_location data by frequency then used the built-in map function of Python and a translator dictionary developed by us to transform all location items and to group similar entries. The subsequent step after user_location transformation is to apply

the Geolocator function of OpenStreetMap through Geopy, which provided latitude and longitude coordinates, that we will use for subsequent mapping applications.

Account Location and User Description
The user location and user description are fields that may contain geographical data sometimes in abstract form. These fields can be edited by the user and have a maximum length of 30 and 160 characters respectively, thus, the credibility of information relies on the user. Generally, these values are not frequently altered and do not necessarily describe the tweet's exact location, but they may represent the user's residence or work place at least on a city level. Most users use this field as intended, however, some fill it out in their native language while others use creative ways to share their locations, either by coordinates, fictitious locations, or communicating with emojis, thus creating noise for standardized geolocation approaches. These limitations can be significantly alleviated by extensive preprocessing steps.

**Transform various coordinate format to DMS**
Besides proper text format, some users choose creative modes to indicate their location, which won't be extracted by the above-mentioned approaches. Some provide precise geo-coordinates in various formats (DMS, MinDec, or DegDec) in the user_location field. Our approach use regex pattern search to extract these values and then converted them to DMS which is useable for our geolocating approach.

**Transform Emoji flags to text**
Another type of extractable user location is based on emoji flags. To minimize user_location's limited input character field, some specify their home country and additional locations as emoji flags, tiny country pictograms. Our approch transform pictograms to text To extract this meaningful data type, we use the Demoji Python library, which is useable for our geolocating approach.

**Securing consistent spelling over user-editable fields**
Although our dataset contains multilingual values, we apply the spell checking over the user location and description field to correct mistyping and increase the subsequent geolocating efficiency.

**Geoparshing and geocoding**
A subsequent step is to extract and standardize possible locations that may represent in the individual Tweets. The approach of geoparsing extracts places from text and matches them to a known place using a global place index. Our approach is based on Mordecai which is a full-text geoparsing system that extracts place names from text, resolves them into their correct entries in a gazetteer and returns structured geographic information for the resolved place name. Mordecai was created to provide several features missing in existing geoparsers, including better handling of non-US place names, easy and portable setup and use though a Docker REST

architecture, and easy customization with Python and swappable named entity recognition systems (NER).

**Translation using Google API and BERT transformer**
The majority of the tweets (68%) were in a language other than English, and as a result, they had to be translated or transformed before being used in the subsequent analysis steps. For the smaller Slovakian dataset we used the TextBlob text-processing library with Google Translate API, while over the larger Belarus dataset we used BERT pre-trained language "xlm-r-100langs-bert-base-nli-stsb-mean-tokens" embedding model.

**Emoji/Emoticon Transformation**
In general, text pre-processing approaches tend to remove any emojis (small images) and emoticons (facial expression representation using keyboard characters and punctuations) from the text. The main problem of such approaches is that users use these small images and characters as the lingua franca of social media to express feelings or ideas. In our methodology, we convert them to word format in order to preserve the emoji information for further analysis steps.

**Semantic Analysis**
The semantic text analysis process used in our approach is divided into two stages: first, we extend the list of stop words in the algorithm based on the characteristics of our data set and then we remove these words from the text, and second, we provide a dictionary-based sentiment analysis, which classifies the subjective sentiment information contained in each tweet.

**Removing Stop Words**
The literature considers auxiliary verbs, conjunctions and other parts of written text that do not bear significant semantic meaning as "stop words". A list of these words is predefined by the Natural Language Toolkit (NLTK) in the algorithm we used. Nonetheless, we added further words to the stop word dictionary that are unique to the unedited text, or the analyzed corpus, including special first names. We also remove these words from the dataset along with words with three or less characters, as they also have limited semantic significance.

***Sentiment Analysis***
Sentiment scores are used to identify how positive or negative the text of a given tweet is. This identification is performed by calculating the difference between the quantity of positive and negative terms using a vocabulary with positive and negative words in an automated way. We selected for this purpose the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon, a rule-based sentiment analysis tool that uses a lexicon-driven method and heuristics to assess the input data. This method is standardized to the sentiments presented in social media, and it has a higher classification accuracy than other methods in the light of the recent

literature.

**Spatiotemporal Data Processing and Clustering**

To understand the escalation of protests based on social media activity, we also performed spatiotemporal analysis using the tweets. Most of the tweets in our query results were posted from European countries; therefore, to keep the analysis concise we only considered countries from Europe, including Russia and Turkey. We kept the data aggregated at country level, as language and the political characteristics of a country might influence the tweeting behavior stronger than other characteristics at city level or other finer spatial scales.

Our approach performed clustering to find countries with similar tweeting trends, which would probably also indicate when protests took place or the presence of other influencing parameters such the media or politics. For this purpose, we used Time Series Clustering in ArcGIS Pro 2.8, where time series data can be clustered based on three criteria: having similar values across time, tending to increase and decrease at the same time, or having similar repeating patterns. By identifying countries with similar pattern, we might be able to reveal the influencing parameters and how these parameters changed over time. Moreover, it provides a more concise and informative visualization and interpretation of the result than statistical values for 39 countries one by one. Moreover, this approach is able to avoid problems related to different population sizes and no normalization based on the population is needed for the clustering.

**Topic Modeling Using Latent Dirichlet Allocation (LDA) Method**

**Preparing Steps for Topic Modeling**

The first step of topic modeling is tokenization, segmenting texts into smaller units. Our approach use the Gensim library's simple pre-process function for this step, which iteratively converts tokens to Unicode strings, removing accent marks and lowercasing the string. Then we filter out the most common bi- and trigrams (two and three-word expressions) from a stream of sentences. However, in order to set a proper filtering threshold, we first manually explored these multi-word expressions with the help of the scikit-learn CountVectorizer, which converts the text to a matrix of token counts. Then, we set up the Gensim threshold to ignore those bi- and trigrams that bear well-known information such as the fact (in the case of the Slovakian dataset) of the murder or the victim's occupation for each cluster identified in the spatiotemporal analysis.

**Lemmatization and Vectorization**

The aim of this step is to reduce inflectional and derivationally related forms of a word to a common base form similarly to a stemming approach. However, in the case of lemmatization, the part of speech of a word (POS tag) such as symbols, numbers or verbs should be first determined and the normalization rules will be different for the different parts of speech thus it is lexically more sophisticated. This method also involves the grouping of the inflected forms of each word, identified by

the word's lemma, or dictionary form so they can be analyzed as a single item, thus enhancing the significance of the topic–word associations. For lemmatizing, we use the spaCy Lemmatizer that provides a rule-based lemmatization with the setting to allow only proper nouns, verbs, and nouns related to our LDA corpus because our research is concentrating on topics that primarily answer the question of who did what and where.

**Performing the LDA Topic Modeling**

For topic modeling, we used LDA with the Gensim library on the final set of geolocated tweets in Python. To date, there is no generally established a priori parameter modeling approach for LDA. In order to find the most suitable parameters, the alpha, beta and the number of topics extractable from the dataset, we apply hyperparameter optimization that seeks after the best setting in a validation corpus set (75%). We used the topic coherence measure (C_v) for performance comparison, which is considered to have the strongest correlations with human ratings. The value of C_v combines an indirect confirmation measure that uses normalized pointwise mutual information (NPMI), cosine similarity, a Boolean sliding window, and the one-set segmentation of the top words. We have applied this optimization approach to all clusters identified in the spatiotemporal analysis. Finally, the tweets were classified according to the topic that produced the highest probability; then, we generated the 10 most frequent keywords for each topic.

**Alternative topic modeling option with Bertopic**

Although Latent Dirichlet allocation (LDA) is on of the most widespread topic modeling techniques nowadays we tried another method using Bidirectional Encoder Representations from Transformers (BERT), which was forced by our second larger Belarus dataset. We used BERTopic model with the pre-trained language "xlm-r-100langs-bert-base-nli-stsb-mean-tokens" embedding model for topic modeling. The embedding model is used to extract the contextualized word representation for all tokens and then passes it to BERTopic, which uses the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce the embedding dimensionality. Furthermore, it preserves both the local and global structure of embeddings. As a sebsequent step, the density-based algorithm (HDBSCAN) clustering the tweets, allowing the identification of the outliers. Practically, class-based TF-IDF is a TF-IDF formula adopted for multiple classes by joining all documents per class.

## IV. RESULTS

This research proposes a comprehensive technique to overcome the limits of current methodologies and to deal with the complexity of protest studies. The suggested method integrates multilingual corpus translation, location and sentiment extraction, and machine-learning topic modeling techniques to expose the underlying interests

and motivations behind collective behavior. Consequently, the given method has a major advantage over previous research that focused heavily on hashtag activism (while disregarding geographical dimensions) or, alternatively, used solely location-specific hashtags. In contrast, by using machine learning methods and methodologies that may be nearly totally automated, we can get a considerably broader variety of input data than in present studies, in which researchers manually review all submissions.

This study investigated the patterns and similarities of Twitter data's geographical, temporal, and emotional indicators by building a fresh data-driven integrated methodology to investigate two unique East European protest movements. First is the 2018 European impact of the murders of Slovakian journalist Ján Kuciak and his fiancée Martina Kusnirova (26 February and 15 March 2018). Second is the impact of the 2020 Belarusian presidential election (9 August–23 September), which has not been studied by current research.

The multi-spectral interpretation of the complex and dynamic nature of the events prompted an effective way of analysis to facilitate in a more thorough comprehension of the protest dynamics. Consequently, our work outperforms the state of the art (SOTA) in two separate ways:

1. This thesis showed that georeferenced social media data may be utilized to analyze political events, even on a smaller geographical and societal scale and in languages other than English.
2. The current study proposes a novel algorithmic approach that integrates time-series clustering with semantic topic modeling and sentiment analytics on geo-referenced social media data in the analytics of social phenomenon.

By presenting an original perspective and considering the presented limitations of existing analyses, this research put effort into profoundly understanding the relevant research of the past decade. The specific results of the research were described in the Results chapter of the dissertation over fifty pages. However, we would like to emphasize again that both the literature review and the methodological setup also provide a unique contribution of the research. Below, I briefly summarize the most important results, item by item:

- **Increased geoparshing accuracy**
  According to the Twitter Developer Platform, approximately 1-2% of Tweets are geo-tagged, while ~30-40% of Tweets contain some profile location information. Our Belarusian dataset had 655,423 non-empty rows in the user_location field and 10,353 places in the user_bio field. The described geoparsher method was able to find 346,162 places in the user_location field and 10,353 places in the user_bio field. The idea behind analyzing this field was that users tend to use it to indicate their occupation and their employer. Overall, the presented mixed-model geolocating approach was able to identify 356,515 geolocated places, which is 54,3% of the Belarusian dataset. The application of the same method on the

#Allforjan dataset returned coordinates for 8069 tweets, which is 61.2% of the original dataset. Kapanova and Stoykova (2019) conducted a comparable study and were able to analyze approximately 29 geolocated Tweets per day, while our methodology provided an average of 100 for each day for Slovakia. This is a 15–20% improvement on average compared to what we knew about the Twitter Developer Platform and a 244% increase compared to their similar study.

- **Increased semantic understanding of the data**
  Prior sociological studies on social media has primarily focused on hashtag activity, overlooking the written language, which is a crucial component of human communication (e.g, LeFebvre & Armstrong, 2018, Sinpeng, 2021). This dissertation focuses on the text, including multi-lingual corpus translation and sentiment extraction, utilizing machine-learning topic modeling techniques to expose the underlying interests and motivations of collective action. Consequently, our technique has a distinct advantage over previous studies, which often ignored these factors.

- **Better understanding the spatial factor in the dynamics of contemporary protest movements**
  In order to comprehend spatial aspects in group dynamics, sociological study on social media has depended primarily on hashtag activity, which provide limited understanding. To comprehend the escalation of demonstrations based on social media activity, the methodology of this thesis did spatiotemporal analysis using geolocated tweets (see above). The data was aggregated at the nation level due to the possibility that a country's linguistic and political factors may impact the movement that was evaluated. The subsequent clustering algorithm identifies nations with comparable tweeting patterns, which may also suggest when demonstrations occurred or the existence of other influential characteristics, such as media and politics. By selecting countries with a similar pattern, our method was able to highlight the contributing characteristics and how they evolved through time. Moreover, it gave a more concise and illuminating visualization and explanation of the result than statistical values for the countries of Europe, one by one.

- **The general spatial characteristics with advanced visualization revealed the supportive countries of the #Allforjan movement**
  The spatial outcomes of the thesis's multilayered spatial analysis demonstrated a direct association with European countries. Although the absolute activity statistics indicated that Slovakia, Germany, Italy, and France were the most active nations, our method eliminated the bias caused by the impact of population sizes. Normalization with each country's population and improved visualization uncovered two crucial factors: active

countries and outliers. Belgium, the Czech Republic, Germany, France, Hungary, Poland, Slovakia, and Switzerland were the eight countries with consistent tweeting activity during the course of the 18 days evaluated in our analysis. Czechia, Hungary, and Poland are three neighboring countries with a shared past. Three outliers were shown by the visualization: Malta, Italy, and Slovakia. A few months earlier to the assassination of Kuciak, a journalist from Malta, Daphne Caruana Galizia, was also murdered. This piqued Malta's curiosity. Due to the relationship between Italian 'Ndrangheta mafia and Slovakian officials, the Italian's position was widely debated throughout the period. Switzerland was the publisher of the periodical for which Kuciak was employed. The normalized heatmap depiction indicated the three most significant theme groups that contributed to the establishment of the protest network.

- **The general temporal characteristics revealed the crucial days of the #Allforjan movement**
  One of most notable spike occurred on 28 February, when Kuciak's incomplete study about the connections between the Italian mafia and Slovakian politicians was revealed, while there are other lesser peaks beginning on the main protest day (9 March) and the resignations of Slovakia's interior minister and prime minister (12 and 14 March).

- **Clustering of these countries based on similarity revealed the dynamics the main characteristics of the #Allforjan movement**
  Our algorithm made five groups: Cluster 1 (blue) had high tweeting activity (1 March) in the first few days and a second, smaller peak when PM Fico resigned (14 March). Italy has the most tweets in this group, mainly because of its mafia's involvement in Slovakia's murders and corruption. Cluster 2 (red) had the least Kuciak-related tweets. Cluster 3 (green) has a modest activity level as Cluster 2, with two smaller peaks on 28 February and 14 March. Cluster 4 only includes three countries, and there's no peak when the murder and motivation were uncovered. 12 and 15 March are peaks for these countries due to the PM's resignation and other indirect effects of the murder or journalist's work. If we look at the subject modeling results of various countries, we can see that tweets about the murder and subsequent events may have a political narrative embedded in their political systems. The trend is obvious, but the absolute quantity of tweets is low (similar to Cluster 3), suggesting Twitter may not be the most popular social media tool in these countries. Those that use it may be less representative of the broader population than in other nations, where more tweets were harvested, distorting the results. Clusters 2–4 lack enough tweets for a deep study of sentiment patterns and subject modeling. Cluster 5 (purple) has over 5500 tweets and includes Slovakia. The group peaked when the journalist was found dead and his unfinished work was published. The

number of tweets dropped to its lowest point on 8 March, before rising again on 9, 12, and 14 March.

- **Sentiment distribution per countries revealed the feedback of the #Allforjan movement and make insight of identity framing**
  Considering the government's ties to the mafia, most people supported the claims and were satisfied with the political consequences, such as the PM's resignation. Using the initial obtained sentiment compound score (before applying sentiment classes), we observed that Slovakia and Germany are among the nations with at least one tweet per day in the analysis period, a statistically rising trend. (Germany had 95% confidence, Slovakia 99%)

- **The topic modeling of the most active clusters (1 and 5) revelealed the main drivers of the #Allforjan movement and the weight of other political factors**
  Cluster 1 includes the PM's departure, the Italian mafia, and the EU's role. Most tweets denounce the murder, followed by concerns about press freedom and security. Kuciak's article, released after his death on February 28 in English and Slovakian, later appeared in other languages (e.g., in French). The article showed links between Slovakia's aristocracy and organized crime. The remaining discussions debate Kuciak's paper (Ndrangheta in Slovakia). Cluster 5 topic modeling findings show 7 subjects. These subjects are clearer Cluster 2 themes discussed not only the PM but also the government's role, protests, the mafia, Kuciak's fiancé, and Viktor Orban, the Hungarian PM, although Hungary was neither in this cluster nor among the most active countries in terms of tweeting behavior. Topic 2's representative tweet discussed the political turmoil in Slovakia, including resignations and new elections. The next topic covers press freedom in Europe, and its urgency is emphasized by the fact that several tweets claimed a clear connection between the deaths of Kuciak and Daphne Caruana Galizia, a Maltese investigative journalist who was assassinated previously, on 16th October 2017. This theme's overrepresentation (Topics 4 and 0) demonstrates that Caruana Galizia's case bolstered the Kuciak movement online. The tweets that linked Galizia's death to Kuciak's fiancée's case might be viewed as a clear denunciation of violence against women that may fuel this campaign. Martina Kusnirova is mentioned by name and as a fiancée, suggesting that users of this cluster not only lamented her death but also differentiated between an innocent and a work-related fatality. Topic modeling also identified tweets on the Hungarian PM and Hungarian-born American billionaire George Soros, a common issue in conspiracy theories and fake news. This may have two causes. First, six months before Kuciak's killing, the Hungarian government initiated a statewide billboard campaign depicting George Soros and declaring, "Don't let George Soros have the last laugh," making him a

scapegoat for the 2016 refugee crisis. It may have helped Slovakian PM Fico's 5 March 2018 political declaration. In this statement, the PM asked about George Soros and Andrej Kiska, who had proclaimed new elections a day earlier. PM Fico may have discredited the president with this statement. Second, PM Orban may have seen George Soros' "fingerprint" in the Slovakian crisis on 10 March. This portrayal of Soros may reflect similar political inclinations among different countries, as the clustering algorithm put Hungary, Belarus, and Turkey in the same group.

- **The general spatial characteristics with advanced visualization revealed the supportive countries of the #Belarusprotest movement**
  Multilayered spatial analysis in the research showed a link between the protest and other European countries. Our strategy eliminated the bias created by population size, which revealed that the UK, Germany, Belarus, and the Russian Federation were the most active nations. Normalizing population and improving visualization revealed active countries and outliers. 33 countries have continuous tweeting during our 44-day analysis. Among these, Western neighbors Latvia, Lithuania and Estonia were the most important ones. These nations first sanctioned Belarus. The Lithuanian parliament imposed economic sanctions on Alexander Lukashenko and 30 Belarusian officials on August 18. These countries also incorporated the judgments of the European Union. The analytics revealed an outlier, namely Austria where the narrative was stick to a possible energetic crisis.

- **The general temporal characteristics revealed the crucial days of the # Belarusprotest movement**
  One of most notable spike occurred when Konstantin Shishmakov, 29, director of Vaukavysk's Bagration Military History Museum, disappeared on 15 and 16 August 2020. As a deputy of the election committee, he refused to accept and sign the forged documents. He called his wife at 5 p.m. and said he was coming home, but he never arrived. On 18 August, Shishmakov's body was found in a river in Grodno. From 15 August, smaller peaks appear. Most likely, these are the biggest protest days because of the "solidarity chain" marches, which protested the crackdown in Belarus following the election and the police violence that followed, leading to multiple deaths and arrests. Josep Borrell, the EU's High Representative, said on August 14 in Brussels that the EU would prosecute Belarusian officials who rigged the election and used violence. On 16 and 23 August, Belarus's largest protest days, a "March of Unity" was held in Minsk with 200,000 participants and all major regional centers. 6,000 people protested in Homel, 4,000 in Hrodna, and 3,000 in Brest, Vitebsk, and Mogilev, which may explain why they were discussed across Europe. Sviatlana Tsikhanouskaya, the leading opposition candidate against Alexander Lukashenko, escaped to Lithuania after the election results were published.

- **Clustering of thes countries based on similarity revealed the dynamics the main characteristics of the #Belarusprotest movement**
  The most imporant result of the clustering is the numbers of the clusters which is two, revealing the spatial limitation of the movement. The first cluster (blue) covers European nations with strong tweeting activity in the first week. After the first peak (18 August), it displays a diminishing pattern with smaller peaks on 20, 24, 28, and 31 August, as well as 7, 14, and 21 September. In addition, tension rises between 7 and 9 September. Mondays have single peaks, indicating Sunday protests. Cluster 1 includes wise political decisions and occurrences. The first turning point came when state-controlled firms joined the demonstrators on 18 August. The opposition created the Coordination Council that day to aid a power transfer. The Belarusian chief prosecutor started a criminal probe into the Council two days later (20 August). Multiple global events affected web trends on August 24. First, publications helped spread the news that 50,000 Lithuanians joined human chain protests yesterday. Tsikhanouskaya met the U.S. deputy secretary as Belarusian police detained two Coordination Council members. EU foreign ministers resolved to penalize 20 Belarusian officials on August 28. On August 31, Coordination Council member Lilia Vlasova was detained. Between 7 and 9 September, Cluster 1 users focused on protest leaders who were kidnapped and carried to the Ukraine border but unable to leave. The 1st and 2nd clusters' trends diverge significantly until 24 August. Cluster 2 (red) Eastern countries had the least active users. In these countries, tweeting activity is modest and falling, with three smaller peaks on 24, 31 August, 7, 14, 21 September. Cluster 2 peaked a day early than Cluster 1 because of demonstrations. On the fifth day of the demonstration, 13 August, participants formed huge lines. If we look at the topic modeling results of Cluster 2, we can see a high possibility that tweets will mention protests, their participation, insults, and the rally's impact.

- **Sentiment distribution per countries revealed the feedback of the #Belarusprotest movement, revealing a certain distance from the events on the part of Western countries**
  Cluster 1 is an fluctuation of positive and negative values. This cluster's users viewed various September events positively, especially on 11, 15, 19, and 21. These were mostly political developments. The EU denounced the violence and demanded Belarus free all jailed demonstrators. The most unfavorable qualities were the largest protests and their aftermaths in Belarus, such as the long solidarity queues (13 August) and when three senior opposition organizers went missing a day after 10,000 Belarusians marched through Minsk. The tweets show a notable focus on protest days, Sundays and Mondays. Protest days dominate tweets. The majority of Monday's positive numbers occur after midnight, indicating protest

marches. Users tend to be negative in the middle, especially a day before protests, which shows how they feel about the government. Cluster 1 analytics show that most respondents supported the charges, given Lukashenko's election fraud and Europe's response. Overall, European users were satisfied with political outcomes like the Coordination Council and EU sanctions but nothing more. Cluster 2 (East) has two negative extremes on 13 August and 7 September. Cluster 1 sentiment changed more dramatically than Cluster 2. Cluster 2 users were greatly affected by the protests, whereas Belarusian factory workers' countrywide strike was regarded good. Cluster 2 users concentrated on negative incidents in Belarus, such as a self-immolation attempt near a police station in Smaliavichy or when 390 women were held by Belarusian soldiers on 19 September. Cluster 2 countries and users focused on how Belarusians live and suffer, while Cluster 1 focused on the shifting political scenario (for example, by putting sanctions on Belarus or seeing the Coordination Council as a legitimate power).

- **The topic modeling of the most active clusters revealed the main drivers of the #Belarusprotest movement together with political and limiting factors**
  The issues of Cluster 1 (West) include protests in Minsk, Alexander Lukashenko's relationship with Vladimir Putin, and EU sanctions against Belarus. Overall, leaders' tweets condemning events in Belarus were the most important. Emanuel Macron said Belarus leader Lukashenko "must go," and Joe Biden chastised Trump for not condemning the despot. Lukashenko's amicable relationship with the Russian president and its effect on events and Belarus' neighbors follow. Cluster 1 focused on European politics and Belarus. Cluster 2 (East) subjects have higher percentage contributions than West topics (Cluster 1). These subjects include protests, human rights, strikes, and the Energy Union. Topic 4's representative tweet focused on the significance of energy in Belarus' political turmoil, namely Russia's efforts to bypass traditional transit countries. The Yamal–Europe gas pipeline connects the Yamal Peninsula and Western Siberia to Poland and Germany through Belarus. Austria is included since it gets its gas from Belarus and the Yamal–Europe gas pipeline. Cluster 2's biggest issue was its reliance on Belarusian natural gas. Women's role in the protest is another important topic. Tweets discussed a Flower Power demonstration in Minsk. Belarusian women stopped police abuse by standing with flowers along major thoroughfares, showing unlimited love. They greeted drivers with white flowers and signals. "Flower Revolution" was the peaceful protest against tyrant Lukashenko. Overall, there is no interaction and multiole connection between Cluster 1 and 2.

- **The unique difference between Slovakia and Belarus was the handling**

**of deaths which revealed a limitation factor for Belarus**

Both case studies included people who tragically died, yet their online reputations were absolutely different. During the protests in Belarus, five men were killed by police brutality. In contrast, these supplementary occurrences were not revealed by the topic modeling strategy. Further investigation reveals that these points were not crucial to the discussion surrounding the presentation itself. A obvious indicator of rising protest levels in Slovakia was funerals, contrary to Belarus.

- **The comparison of the events in Slovakia and Belarus in the light of connective and collective logics revealed those factors that led success only in Slovakia**

  Both movements used technology to create a decentralized network. Their structures differ greatly. Slovakia had a large, flexible network with unlimited access. Movement fostered connections and deeper interpersonal bonds. In Belarus, the government partially controlled network access and exchanges, so the extensive, flexible network was not fully integrated. Slovakia's social norm legitimized engagement and cooperation by encouraging others to contribute to the common good. Belarus' societal norms may be limiting. Personal narratives must be connected to collective action frames because individual actions can become collective action frames through continual encounters and cooperation. Social pressure in Belarus maintains gender norms. Belarus lacks institutional and statutory foundations for women's equality, despite no legal barriers to women in administration. The "we" (women) versus "they" (men) division in politics might establish collective action frames, but its success may be limited, especially in Belarus.

**RELEVANT PUBLICATIONS DURING THE DOCTORAL STUDIES**

Kovács, T., Kovács-Győri, A., & Resch, B. (2022). #Belarus: Spatial Proximity—Classifying Protesters Based on Spatiotemporal and Sentiment Analysis of Twitter Data. *Sensors* (in progress) | WOS, SCI/SSCI Q1(Impact 3.576)

Kovács, T., Kovács-Győri, A., & Resch, B. (2021). #Allforjan: How Twitter Users in Europe Reacted to the Murder of Ján Kuciak—Revealing Spatiotemporal Patterns through Sentiment Analysis and Topic Modeling. *ISPRS International Journal of Geo-Information*, *10*(9), 585. | WOS, SCI/SSCI Q1(Impact 3.05)

Kovács, T. (2020). A vizuális tartalom társadalom befolyásoló hatása. *Geopolitikai szemle*, *2*(1), 221–233.

Kovács, T. (2020). Lengyelország felzárkózása a rendszerváltást követően. In Somkúti B. (Ed.), *Közép-Kelet Európa központi makrorégiója és a kapcsolódó államok gazdasági felzárkózásának sikerei és kudarcai 1990 után* (pp. 1-30) (accepted for publication).

Kovács, T. (2020). Bulgária felzárkózása a rendszerváltást követően. In Somkúti B. (Ed.) *Közép-Kelet Európa központi makrorégiója és a kapcsolódó államok gazdasági felzárkózásának sikerei és kudarcai 1990 után* (pp. 31-60) (accepted for publication).

Kovács, T. (2019). Euromaidan: The Route to the Crimean Crisis that went through Online. *Vojenské Reflexie*, *15*(Mimoriadne číslo 1), 47–58.

Kovács, T. (2018). Recent #Martyrs. In *National and International Security 2018: Proceedings of the International Conference on National and International Security* (pp. 199–205). Armed Forces Academy of General Milan Rastislav Štefánik.