Journal of Social Media Research

Journal of Social Media Research
ISSN: 3062-0945
2025, 2(1), 50-70
https://jsomer.org

REVIEW ARTICLE OPEN ACCESS

Extracting emerging events from social media: X/Twitter and the multilingual analysis of emerging geopolitical topics in near real time

John Corcoran Burns[1]*, Tom Kelsey[1], and Carl Donovan[2]

1 School of Computer Science, University of St Andrews, St Andrews, Scotland, United Kingdom.
2 School of Computer Science, University of St Andrews, St Andrews, Scotland, United Kingdom.
3 School of Mathematics and Statistics, University of St Andrews, St Andrews, Scotland, United Kingdom
* **Correspondence:** Jack Cole Building, North Haugh, St. Andrews, KY16 9SX, United Kingdom, Tel: +1-718-702-7724, jb370@st-andrews.ac.uk

Article Info

DOI: 10.29329/jsomer.14

Article History:

Received:	27/11/2024
Revised:	08/02/2025
Accepted:	01/03/2025

Keywords

Social Media Monitoring
X / Twitter,
Topic Modeling
Geopolitics
Google Trends

Highlights:

- X / Twitter can find emerging geopolitical topics faster than Google Trends, in some cases.
- Our X / Twitter algorithm can identify topics hours earlier than when they trend through Google searches.
- Our X / Twitter algorithm's use of multiple languages can identify trending topics worldwide for specific regions.

Abstract

This study uses multiple languages to investigate the emergence of geopolitical topics on X / Twitter across two different time intervals: daily and hourly. For the daily interval, we examined the emergence of topics from February 4th, 2023, to March 23rd, 2023, at random three-hour intervals, compiling the topic modeling results for each day into a time series. For the hourly interval, we considered two days of data, June 1st, 2023, and June 6th, 2023, where we tracked the growth of topics for those days. We collected our data through the X / Twitter Filtered Stream using key bigrams (two-word phrases) for various geopolitical topics for multiple languages to identify emerging geopolitical events at the global and regional levels. Lastly, we compared the trends created by tracking emerging topics over time to Google Trends data, another data source for emerging topics. At the daily level, we found that our X / Twitter-based algorithm was able to identify multiple geopolitical events at least a day before they became relevant on Google Trends, and in the case of North Korean missile launches during this period, several languages identified more missile launches than the Google Trends data. As for the hourly data, we again found several topics that emerged hours before they started appearing on Google Trends. Our analyses also found that the different languages allowed for greater diversity in topics that would not have been possible if only one language had been used.

Citation: Burns, J. C., Kelsey, T., & Donovan, C. (2025). Extracting emerging events from social media: X/Twitter and the multilingual analysis of emerging geopolitical topics in near real time. *Journal of Social Media Research, 2*(1), 50-70. https://doi.org/10.29329/jsomer.14

1. Introduction

From the Ukraine War to increasing tensions between the United States and China and conflicts in the Middle East, geopolitical pressures are at their highest levels in decades. Developments in geopolitical events can occur rapidly, forcing governmental and non-governmental actors to adjust their responses on short notice. Therefore, gaining as much lead time as possible to identify emerging geopolitical events is vital, enabling the best course of action to materialize. However, not all geopolitical events have global implications; some may be regional issues that are highly significant to specific areas. Thus, it is essential to use multiple languages to capture diverse opinions on widespread global events and to track more localized occurrences that a single language might miss. While relying on traditional media sources to monitor the emergence of geopolitical events is feasible, we have chosen to utilize X/Twitter for several reasons. As of 2025, X/Twitter is the eighth most popular social media platform globally and continues to grow rapidly, especially among younger generations (Duarte, 2025; Martin, 2023). Additionally, X/Twitter offers data in multiple languages and provides this information in real-time, facilitating the gathering of multilingual data for our study instead of relying on various traditional news sources. Furthermore, traditional news media often lag behind social media, which makes it possible to overlook the emergence of a story. In contrast, the spread of stories on X/Twitter can be tracked more easily through the available APIs during our research. Moreover, by analyzing tweets in English, Spanish, French, Portuguese, Arabic, Japanese, and Korean, we can capture 85% to 90% of all tweets posted (Vicinitas.com, 2018). This linguistic diversity allows us to identify specific regional geopolitical events and provides a global perspective on major worldwide geopolitical issues that might be overlooked if we relied solely on English.

Specifically, this paper aimed to investigate geopolitical events topic modeling at both the daily and hourly levels. By examining the data at the hourly level, it would be possible to find the emerging topics quicker than at a daily level, and we can also find topics that were only relevant within the day that might get lost if a larger time series was used. We only investigated three-hour intervals, which is short. However, we investigated over 40 days, which allowed for exploring topics generated within the day and throughout the entire study period. Finally, multilingual data adds more novelty to the study by investigating the topics generated across regions, allowing for comparison and contrasts between what geopolitical risk emerged in different parts of the world and which are more relevant to these regions.

The rest of the paper is as follows: Section 2 focuses on the key concepts that support our study, while Section 3 describes related work that helped form our methodology. Section 4 describes the methodology we developed for dynamically tracking emerging geopolitical risks. Section 5 details our results, Section 6 discusses our findings, and Section 7 concludes.

2. Key Concepts

Three concepts—topic tracking over time, geopolitical topic generation, and Google Trends — underpin our paper on emerging geopolitical topics.

Section 2.1. Topic Tracking over Time

For our research, we determined that the best way to examine emerging topics would be through an extension of Latent Dirichlet Allocation (“LDA”) (Blei, et al., 2003), however, some background into topic tracking will be helpful for greater understanding. The roots of topic tracking over time from news and social media lay in the mid to late 1990s with a DARPA funded study aimed at using various methods to “ (1) segmenting a stream of data, especially recognized speech, into distinct stories; (2) identifying those news stories that are the first to discuss a new event occurring in the news; and (3) given a small number of sample news stories about an event, finding all following stories in the stream.” (Allan, et al, 1998, p. 1). Thus, with a time series of text data, such as news articles, is it possible to identify emerging stories given the methods used at the time? Partners in the study: CMU, UMASS, and Dragon Systems, tried different ways based off clustering algorithms to group the different news articles into topics while examining emerging and fading topics with success (Allan, et al, 1998, p. 11 - 13). However, new methods for identifying topics with increased computational power have been applied to this research field (Blei & Lafferty, 2006; Cordeiro, 2012; Hurtado, et al, 2016; Weiringa, 2023). One of the most popular methods that have been employed, and the one our study uses, is topic modeling, specifically LDA (Blei, et al., 2003).

Topic modeling is a similar process to clustering. However, it uses text data, where each document in the corpus (the set of documents) is sorted into groups of related documents which form the topics (Kulshrestha, 2022). Topic modeling "can be considered as a fuzzy classification because it provides a soft degree of belonging of the documents to a specific topic" (Amara, 2021). This means that one document can belong to multiple topics. LDA extends this concept. As Blei et al. (2003) defined, "LDA is a three-level hierarchical Bayesian model, in which each collection item is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.". In other words, LDA uses an expectation maximization ("EM") process with Dirichlet – Multinomial conjugate priors based off the algorithmic hyperparameters set for number of topics, and the alpha and beta of the Dirichlet representing the document – topic density and the word – topic density respectively (Kapadia, 2019). Using the hyperparameters as a start for the EM process, LDA updates the probabilities of each word in the corpus belonging to each topic and the probabilities each document in the corpus belonging to each topic. Once the EM process converges, the LDA will have optimal values for the words in each topic and which topic each document most belongs too, which allows for interpretation of the topics based of the words that make up each topic, thus "discovering the hidden themes in the collection" (Ria, 2019). In plain English, LDA analyzes the documents of interest, in our case tweets, and builds topics based on words in the document corpus that have the greatest relationship to each other. Then, it assigns each document with differing probabilities for each topic. The highest probability associated with topic for a document is the topic the document most likely discusses. Thus, if you have a continuous data set over time, it is possible to graph the changes in topics over time based on when the new documents enter the system, thus changing topic probabilities. Also, as Amara states: "the LDA can be used as predictive model to detect novel trends in the behaviour of the social media users' topics or to give a deeper analysis about a specific topic by detailing the document providing the target topic." (Amara, 2021, p. 6). While LDA is powerful method to find topics in documents, the major drawback to LDA for our study is that it is "static" (Blei & Lafferty, 2006). This means that it evaluates all the documents in the corpus at once and is not designed to evaluate the data chronologically, which posed a problem for our analysis.

Luckily, several methods extend LDA for this capability. Dynamic Topic Modeling is one such method which "developed sequential topic models for discrete data by using Gaussian time series on the natural parameters of the multinomial topics and logistic normal topic proportion models" (Blei & Lafferty, 2006, p. 6). Using a normal logistic distribution over the Dirichlet distribution allows for the development of topics over time, and thus, we can see when topics emerge (Blei & Lafferty, 2006). Alternatively, one could do this process manually by running the LDA algorithm over each time frame, labeling the created topics and comparing them to the previous time frames to track the emergence of topics over time. Besaw employed a different method, using weights for the topics created and showing how they emerge and change over time (Besaw, 2021). This is the method we incorporate into our analysis for this study, and we will go into more depth in the Methods section.

Section 2.2. Geopolitical Topic Generation

As many geopolitical events constantly occur, we needed to identify generic geopolitical terms that would allow us to capture as much data as possible, as briefly described below.

For this study, we use the X / Twitter Filter API, which we describe further in the Methods section. However, here we describe the key terms used to generate our data. Initially, we only used terms from the "Goldstein Index" (Goldstein, 1992), which was developed from a study by Goldstein, which aimed to rank geopolitical events from most conflictual to most cooperative, also described in the Methods. Based on the table Goldstein developed (Goldstein, 1992), we created bigrams (two-word phrases) to capture geopolitical tweets better. However, we found through our initial testing, that we did not gather enough tweets per hour for the LDA algorithm to develop coherent topics, thus we decided to include more sources to expand our geopolitical topic gathering potential. We used two sources: Klement, 2021 and Caldara and Iacoviello, 2022. Klement is a textbook on geo-economics, and each chapter discusses factors that affect global economics and that occur from different geopolitical events. Caldara and Iacoviello created an index for tracking and evaluating geopolitical events through looking at different English news articles for specific keywords. We discuss how we created our keywords for our study in the Methods section, but we wanted to introduce the key Bigrams sources here. The key bigrams themselves can be found in Table 1 in Appendix A.

Section 2.3. Google Trends

The last concept we discuss is Google Trends ("Google Trends", 2022). Google Trends is a website that tracks topic popularity through the increase in Google search queries. This provides a valuable parallel test to contrast our X / Twitter emerging topic modeling, generating a trend that would occur concurrently over time. As described by Choi and Varian: “Google Trends provides *daily* and *weekly* reports on the volume of queries” (Choi & Varian, 2009, p. 3). Since 2009, Google has enhanced Google Trends by adding functionality that reports changes in the popularity of search terms and provides data at the minute level, extending up to four hours before report generation. Choi and Varian also describe the change in popularity data: “Google Trends data does not report the raw level of queries for a given search term. Instead, it provides a query index. The query index begins with the query share, the total query volume for a search term in a specific geographic region divided by the total number of queries in that region. The query share figures are normalized to start at 0 on January 1, 2004. Later numbers indicate the percentage deviation from the query share on that date.” (Choi and Varian, 2009, p. 4). While January 1, 2004, is the earliest date for which Google Trends data is available and was used by Choi and Varian in their study, we generated Google Trends reports from February 4, 2023, to March 23, 2023, at the daily level, and for the four-hour time frame from June 1, 2023, to June 6, 2023, focusing on the three-hour window during which we collected our tweets. We chose Google Trends for comparison with our emerging geopolitical topic analysis because Rill et al. demonstrated that Google Trends effectively compares to their PoliTwi System, which aimed at capturing emerging German political topics using X / Twitter (Rill et al., 2014). While examining similar concepts, we also opted to utilize Google Trends.

Numerous studies employ topic modeling concepts and LDA; however, our study particularly focused on research into topic modeling over time and multilingual topic modeling.

Section 3.1. Dynamic Topic Modeling Literature Review

While Blei and Lafferty developed dynamic topic modeling based on the LDA mathematical framework, other studies created methods to track topics over time. In a variation of LDA, Griffiths and Steyvers replace the EM process of LDA with a Markov Chain Monte Carlo (“MCMC”) process known as Gibbs Sampling to obtain the topic and word probability distributions (Griffiths & Steyvers, 2004, pp. 2 – 3). Additionally, they tracked topics over time in what they termed “Hot or Cold Topics” by using Linear Trend Analysis to identify the emergence of topics over time (Griffiths & Steyvers, 2004). Wang and McCallum adopt a different approach than the previous two; they directly incorporate time into the LDA algorithm via the Beta distribution, enabling “a continuous distribution over time associated with each topic, with topics responsible for generating both observed timestamps and words. Parameter estimation is thus driven to discover topics simultaneously capturing word co-occurrences and the locality of those patterns in time.” (Wang & McCallum, 2006, p. 1). Further studies built on the foundation established by this research, including Ahmed and Xing’s infinite Dynamic Topic Model, which utilized a recurrent Chinese restaurant franchise process to allow the words to vary dynamically between topics, facilitating the emergence of new topics while allowing others to fade without restriction on the number of topics. Hurtado et al. expand the research by attempting to forecast which topics will be popular soon. Meanwhile, Hida et al. combine the dynamic aspects of Blei and Lafferty’s Dynamic Topic Modeling to capture the emergence and changes of topics over time with the static nature of LDA, providing a better understanding of how topics relate to one another. A more recent method developed by Grootendorst, 2022 is BERTopic. BERTopic implements dynamic topic modeling in four steps: the first leverages Sentence-BERT (Reimers & Gurevych, 2019) to embed the documents (turn the document words into vector representations). Next, the dimensions of the newly embedded corpus are reduced and resulting data points are clustered based on density. Afterwards, the documents in each cluster are evaluated with the TF-IDF metric (this metric describes the importance of the word to the document) by expanding this to the c-TF-IDF variation, the important words can be found for each cluster (the c in cluster), allowing for topic descriptions to be found. Finally, for dynamic topic modeling, the previous steps are applied first to the whole dataset to develop the topics, then run again at each time interval to find the changes in topic importance (Grootendorst, 2022). While these are all innovative techniques, our method varies in a few key ways. First, we use hyperparameter tuning with LDA to improve topic coherence, which other studies did not. Additionally, unlike Grootendorst, we run our topic tracking over time algorithm at each time stamp with the full data gathered to understand the changes in topic emergence. We go into further detail of our algorithm in the Methods section. Lastly and importantly, we found several studies (Culotta, 2010; Cordeiro, 2012; Becker, et al., 2021) use the X / Twitter API temporal qualities to analyze topic identification and the growth of these topics over time. These studies proved that our method of identifying emerging geopolitical events on X / Twitter and tracking their changes was possible.

Section 3.2. Multilingual Topic Modeling Literature Review

Another area that influenced our research was multilingual topic modeling. This concern is important in topic modeling, as Lind et al. state, “automated methods of content analysis (such as topic modeling) are usually applied to text documents in just one language—mostly English” (Lind et al., 2019, p. 2). Researchers have employed various methods to develop topic modeling techniques suitable for a multilingual framework. For instance, Boyd-Graber and Blei created a topic model for multilingual text called MuTo. MuTo is based on the LDA algorithm; instead of focusing solely on words, MuTo utilizes matching pairs of words from multiple languages to create corresponding topics across languages. Boyd-Graber and Blei demonstrated that MuTo performed well with similar texts in English and German. In an updated study, Yang et al. revealed that their multilingual topic model “does not force one-to-one alignment across languages,” which facilitates better topic generation for languages with smaller corpora (Yang et al., 2019, p. 1). Similarly, Yuan et al. (2018) constructed multilingual topics through anchor word-based methods, where the exact words across various languages served as the foundation for creating topics, which users could then refine. Lastly, Zosa and Pivovarova (2022) advanced this concept further by incorporating images with linked multilingual text data. Like Boyd-Graber and Blei, their M3L-Contrast model outperformed other models that exclusively used text data. However, our research employs language as a proxy for location to examine the geopolitical topics in various regions. Therefore, we intentionally avoided linkages between topics; fortunately, many studies provided valuable insights. Zheng et al. utilized separate LDA models for Japanese and Chinese to compare the topics discussed in different online blogs. Notably, two multilingual papers that were particularly helpful were by Amara et al. and Sakamoto et al. Both papers explored trends—COVID-19 for Amara and the U.S. and Japanese legislatures for Sakamoto—across multiple languages using distinct LDA models for each language. They also monitored these trends over time, demonstrating the emergence and growth of the topics.

Section 3.3. Novelty

The novelty of our research in this space is combining multilingual and dynamic topic modeling methods. Furthermore, we examine a novel area of applying this combination on geopolitical events found through the analysis of tweets. Additionally, we apply this synthesis of methods over a smaller time frame than has been investigated previously, showing changes in geopolitics not just at the daily level, but also detailing the emergence of topics at the hourly level as well

2. Method

Our methodology has four sections: the Data Gathering section, the Data Cleaning section, the Multilingual Emerging Geopolitical Topic Modeling section, and finally, the Evaluation section.

Section 4.1. Data Gathering

We chose X /Twitter as our data source; thus, we decided to implement the X / Twitter API, specifically the Filter Stream API[1]. This X / Twitter API allows us to obtain real-time tweets that contain key phrases in the different languages we are investigating. We felt confident using this method of data collection as other studies, such as Metzler, 2012, and Culotta, 2010, both use key words in their studies to gather tweets. As for the ethical considerations for using X / Twitter data, we only collected the bare minimum data through the API, just the username and text, no geographic information. We only analyzed public tweets, public posts online, not private messages. The users know that anyone can see these public tweets, and according to Reuter et al. 2019, most Twitter users do not find monitoring X / Twitter an “inappropriate surveillance or a violation of privacy (Reuter et al., 2019); thus, our research did not violate their privacy.

As described earlier, we initially used the “Goldstein Index” topics in Appendix A in Table 1. However, we decided to expand our reach of potential geopolitical events and included topics from other sources, such as Klement, Caldara, and Iacoviello. With our geopolitical topics in hand, we gathered key words and phrases for these topics. For Klement, we chose the key phrases based on our readings, breaking the different chapters he discusses on geoeconomics into different topics. As for Caldara and Iacoviello, who focused exclusively on geopolitical risks, had a table that broke down the different geopolitical topics they were searching for and the key phrases they used to search for them (Caldara & Iacoviello, 2022). After determining our topics, we also

[1] X / Twitter. "Filtered Stream."

gathered the most common key phrases from web scraping Wikipedia articles related to our topics (Terrorism[2], Oil Supply Shock[3], US – China Relations[4], Cyberwarfare[5], Nuclear Threats[6]). Additionally, we tested whether monogram (one-word phrases) or bigrams (two-word phrases) were better for our study for gathering tweets. We found that while the monograms gathered more tweets, they were less accurate in terms of content than the bigram phrases, thus we decided to use bigrams as their increased relevance would produce better geopolitical topics. Combining these methods gave us a list of relevant bigrams for each topic, however, the next step was to translate these bigrams across the six other languages we were investigating.

We chose human translation over machine translation for our translation process, as Kravariti (2016) describes: "It is a translator's job to ensure the highest accuracy… Humans can interpret context and capture the same meaning, rather than simply translating words… Humans can spot content where literal translation is impossible and find the most suitable alternative" (Kravariti, 2016). While not as fast as machine translation, a human translator's increased accuracy and flexibility were vital for our data gathering needs, as obtaining the correct keywords allows us to collect the most relevant tweets for our topics. Thus, we employed translators from Gengo.com[7] for French, Portuguese, Arabic, Japanese, and Korean, while a private translator handled Spanish[8]. Unfortunately, the X/Twitter Filter Stream API rules have limits, so we could only choose a limited number of key bigrams for each topic when gathering the data. We decided to select five bigrams for each topic, except for the "Goldstein Index" topics, for which we increased the limits to ten bigrams each to gather more tweets with a defined geopolitical focus. We ran the Filter Stream API for an hour (our testing showed that a time interval under an hour would not produce viable topics), while storing all the tweets that came in as JSON files. At the end of that hour, we take the JSON files in our storage folder and convert them into a Python data frame, where we begin our data cleaning process.

Section 4.2. Data Cleaning

The first step in the data cleaning process is labeling the languages of the tweets. While the Filter Stream API can include a language indicator as one of its filtering rules, the tweet output does not contain a language tag. Thus, we must create this tag to separate the languages for later analysis (Danilak, 2021). Once the languages of the tweets are identified, we split the data into individual language data frames and sort these data frames chronologically. We then tokenize (i.e., split the tweets' text into individual words) the tweets in the language data frames and remove the language-specific stop words, while also including certain stop words related to URLs and X/Twitter: (["http", "https", "co", "com", "app", "go", "amp", "RT", "rt"]). Afterward, we group all the remaining words into bigrams, i.e., two-term phrases of words that occur next to each other (Rehurek & Sojika, 2010). This process helps identify distinctive word combinations and build the corpus necessary to implement the LDA algorithm. Once the bigrams are constructed for each language, we apply lemmatization to the bigrams for English, Spanish, French, Portuguese, and Japanese (Honnibal & Montani, 2023). Lemmatization removes the endings of words to retain the word root (for example, "bus" is derived from "buses"), so that these word roots can be more easily grouped to create clearer topics later. We do not include Arabic or Korean in the lemmatization for various reasons. For Arabic, lemmatization generally involves removing the diacritical marks, which are "short vowel symbols inscribed atop regular letters" that are a part of written Arabic (Hegazi, et. al, 2021, p 1). However, "Arabic abjad or letter can be seen in several words which do not carry a similar meaning, so in order to remove that confusion that can even occur to native speakers, we use diacritical marks in a wide range of texts" (International House Cairo, 2023). Thus, we tested to see if removing the diacritical marks would greatly affect the comprehension of the topics created. We found that the topics with diacritical marks made more sense than those without; thus, we decided to skip the lemmatization process for Arabic. With Korean, however, we encountered a different problem. Through testing, we found that many out-of-the-box lemmatization programs for Koreans lost much of the tweet's context since we were working with such short texts. However, studies such as Lee and Song, 2020, skip the lemmatization step in their analysis step of Korean text data, so we feel safe doing the same. Once our data was cleaned, we continued to the first step of the LDA process, which is creating the corpus and dictionary for each language out of the either the lemmatization

[2] Wikipedia contributors, "Terrorism."

[3] Wikipedia contributors, "1973 Oil Crisis."

[4] Wikipedia contributors, "China–United States Relations."

[5] Wikipedia contributors, "Cyberwarfare."

[6] Wikipedia contributors, "Nuclear Warfare."

[7] https://gengo.com/

[8] See Acknowledgements for Details

bigrams data for English, Spanish, French, Portuguese, and Japanese, or the regular bigrams data for Arabic and Korean. We can move on to creating the emerging geopolitical topic models with these created.

Section 4.3. Multilingual Emerging Geopolitical Topic Modeling

Our emerging geopolitical topic modeling analysis methodology mainly comes from two sources: Kapadia (2019), who provided the framework for building our dynamic selection of the hyperparameters for our LDA algorithm, and Besaw (2020), who developed a way to track the growth of topics over time that interfaced well with the rest of our design. Our study's novelty is synthesizing these two methodologies into one framework.

The first part of our process involves the dynamic selection of hyperparameters for the LDA algorithm. As explained in the Key Concepts section, the LDA algorithm operates by first setting hyperparameters, which are parameters determined before training the LDA model. While these can be chosen randomly, the best results arise when these hyperparameters are fine-tuned for data analysis. Therefore, for each of the seven languages, we created hyperparameter tuning functions to dynamically set alpha (the document-topic density), beta (the word-topic density), and the number of topics. To achieve this, we rely on the concept of coherence. As defined by Roberts et al., coherence is “maximized when the most probable words in a given topic frequently co-occur together” (Roberts et al., 2013, p. 10). This indicates that a topic can be deemed to have high coherence if the bigrams frequently appear together; thus, the topic is more easily interpretable by a human reader of its words. Consequently, higher coherence generally correlates with better-quality topics (Roberts et al., 2013, p. 10). Therefore, finding the combination of hyperparameters that maximizes coherence is crucial to the LDA topic modeling process. To do this, we follow the procedure outlined by Kapadia, 2019, who developed a method to loop through each combination of a predetermined list for each hyperparameter (alpha, beta, and the number of topics) and record the coherence value from each run of the LDA model with those hyperparameters in a data frame. We enhanced this process by automatically identifying the maximum coherence from the output data frame and inputting the hyperparameters corresponding to this value into our primary LDA modeling function within our algorithm.

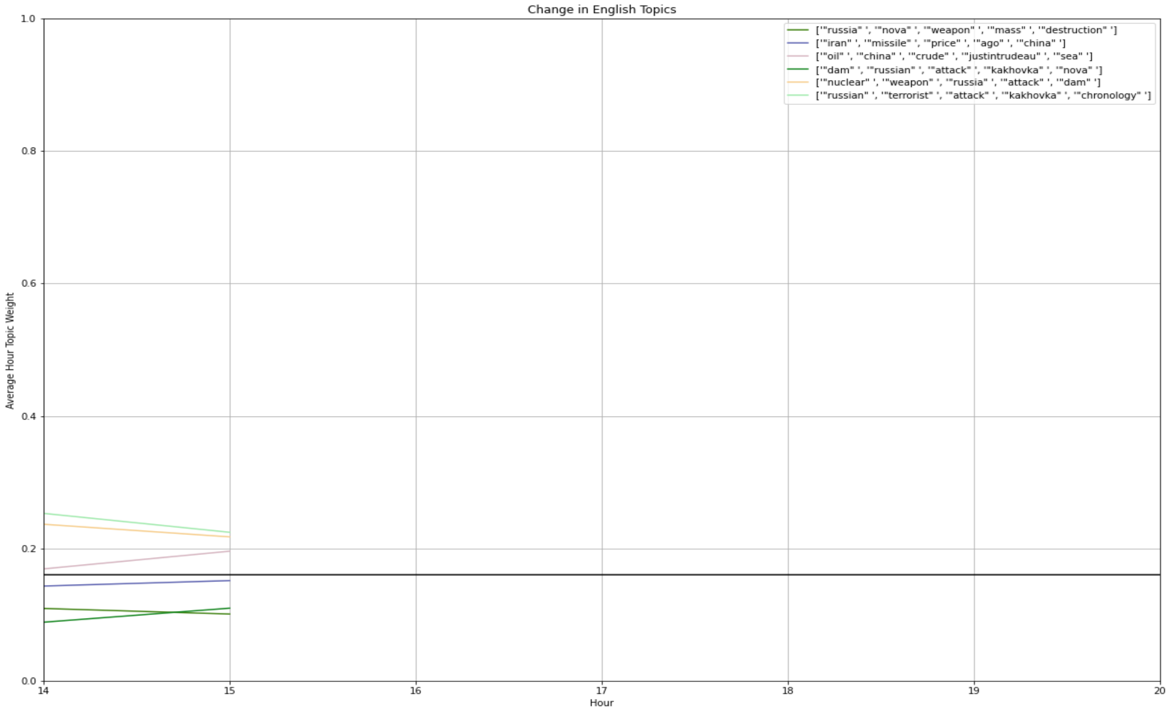


Figure 1. This is the first hour of data for the English topics for June 6th, 2023, starting at 2 pm GMT (14:00). On this day, there was major news out of Ukraine with the destruction of the Nova Kakhovka Dam[9]. However, the US also placed sanctions on Iran and China over Iran’s hypersonic missile program[10]. The solid black line is the median value for the average weights of the topic proportions across the entire time

[9] Lakezina, Viktoriia. "War Zone Villagers Flee after Massive Ukraine Dam Destroyed."

[10] Psaledakis, Daphne. "Us Slaps Sanctions on Iranian, Chinese Targets over Tehran's Missile, Military Programs."

period. Trends under this line are less relevant than the ones above it. Trends that cross it show an increase in relevance over time and could be considered emerging topics

With our hyperparameters tuned, we move to implementing the primary LDA model. We plug in the hyperparameters and obtain the LDA model with the topics with the highest coherence. However, these topics do not describe how they changed or if any emerged, so we need to manipulate the topic modeling results to work as a time series. First, we need the topic proportions of each tweet for each topic, this is the proportion of words in a tweet that belong to a particular topic, also known as the weight (i.e., how much each topic holds weight in a document). Once we have each topic proportions for each document (in our case, a tweet), we can create a time series by breaking the sum of the individual topic proportions by hour. We divide that value by the tweets count each hour, giving the average topic weight by hour. This average topic weight allows us to track the changes in the topic by hours and evaluate if certain topics are emerging or losing relevance (Besaw, 2021; Wieringa, 2023). Finally, we used the first five topic words to label each topic. However, for all non-English languages, we use Google Translate (Nidhaloff, 2020) to translate the topic words to English before labeling the topic trends over time for easier comparison between the language topics. The labels give us a better understanding of the trends we visualize to track the emergence of topics. Lastly, this process is repeated every hour, where more data is gathered and reprocessed with the previous data collected to see how topics have changed from the previous hour. This is done during each hour of the three-hour capture period. In Figures 1 – 3, below, we provide an example visualization of the process, these are the topics for English from June 6th, 2023, starting 2 pm GMT (14:00)

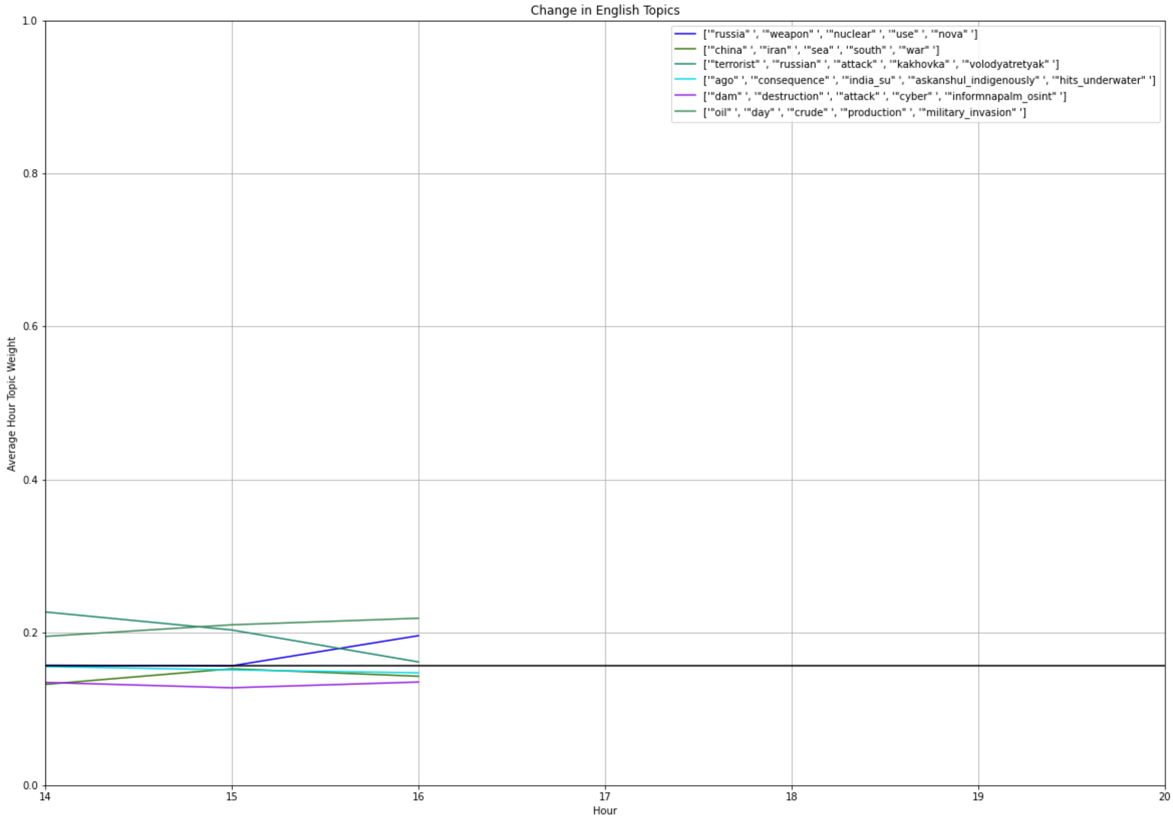


Figure 2. This is the second hour of data topics now added to the first hour, showing the changes in topics over the time, as the figure shows, the Nova Kakhovka Dam story is starting to fall in relevance, while the China and Iran sanction starts emerging more

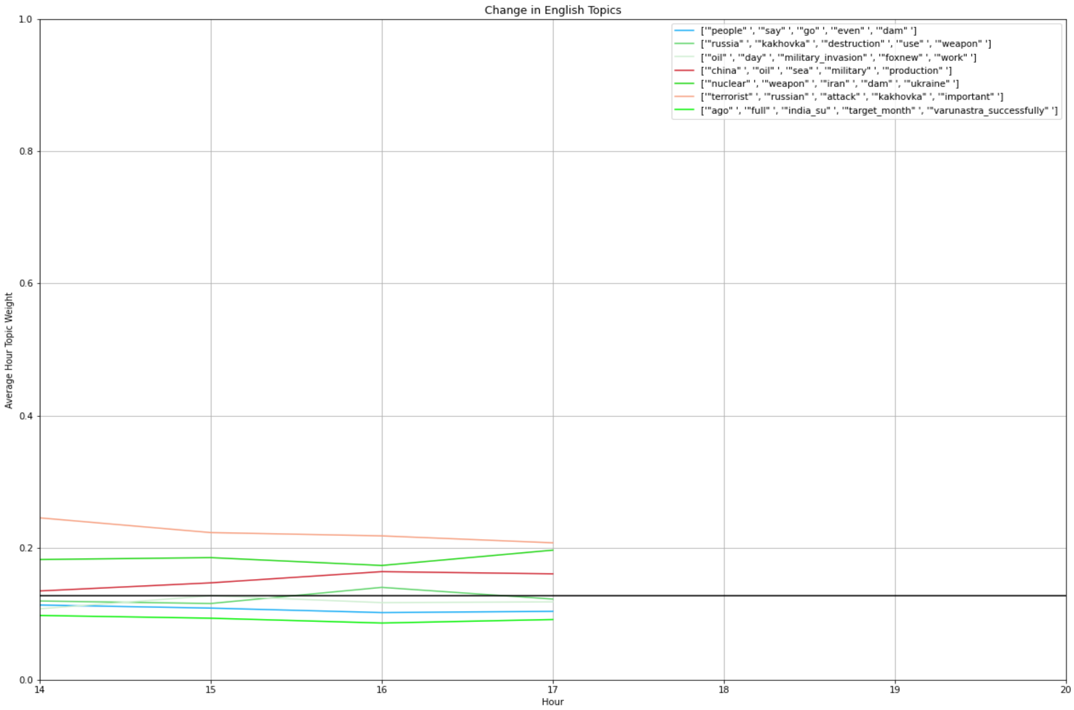


Figure 3. This is the third hour of data topics now added to the previous hours, the Dam is still relevant but decreasing, while Iran and China have split into separate, but still relevant topics, with the China topic leveling off, but the Iran topic still growing

For our study at the daily level, we examined the period from February 4th, 2023, to March 23rd, 2023. We take the final hour of data topics, like Figure 3, and get the count of number of topics that appeared for their labels. For example, in Figure 3, we would say that there were three topics related to the Dam, one for China, one for Iran, and one for India. This gives us a time series trend we can evaluate using the Google Trends described below. The Supplemental Material file: Appendix F contains a full illustration of the methodology.

Section 4.4: Evaluation and Validation

For both the Daily and Hourly trends of emerging geopolitical topics data, we followed the lead of Rill et al. (2014), who evaluated German political topics on X/Twitter against Google Trends. We also aimed to determine if our methods for emerging geopolitical topics could outperform Google Trends. For the hourly analysis, we collected minute-level data from Google Trends over a four-hour period during the analysis timeframe of the X/Twitter model. We compared the Google Trends data to the trend lines generated through changes in X/Twitter topics over time. This allows us to assess whether a topic identified in the tweets appeared in Google Trends first or through the X/Twitter topic models. For the daily analysis, we utilized daily Google Trends data covering the entire study period from February 4, 2023, to March 23, 2023, and compared it to the number of topics generated for each day (specifically, the topics that remained at the end of our model analysis timeframe). This establishes a trend in counts over time for the X/Twitter topics, which can be compared to Google Trends for the same topic, enabling an evaluation of the emergence and relevance of each topic to see which method captures it first. For the validation of the emerging topic models, after the completion of the analysis run, we take the topic words generated and search them to see if any events took place during the analysis time frame that would correspond to the emerging topics for each language. In the Results section and Appendices B, C, D, and Supplemental Material section in Appendix E, we have cited articles related to the generated topics.

3. Results

We obtained our Hourly Topic Results for our analyses across two days: June 1st, 2023, and June 6th, 2023. As for the daily level, we examined at the period from February 4th, 2023, to March 23rd, 2023. We obtained many emerging geopolitical topics across our seven study languages, however, for the sake of clarity in the main body of this study, we only include one example that best exemplified each type of analysis. However, we include our remaining hourly results in Appendix B, and our remaining daily results in Appendix C. Additional results investigating more topics can be found in Supplemental Material file: Appendix E.

Section 5.1. Hourly

Figure 4 below displays the English topics generated and tracked during the final hour of our three-hour time frame on June 1, 2023. The time is set to GMT, as this is the time zone that X/Twitter uses to record when a tweet is created. As the chart illustrates, seven topics were monitored (selected due to computational power constraints). While many topics lost relevance during this period, the green-colored topic of “Iran" appears to have emerged around Hour 19. This emergence was related to the announcement by the U.S. Department of State11 regarding sanctions on Iranian operatives involved in external plots, such as assassination attempts that occurred outside of Iran. The NCRI12 reported these new sanctions at 20:30 GMT, just after the “Iran” topic emerged.

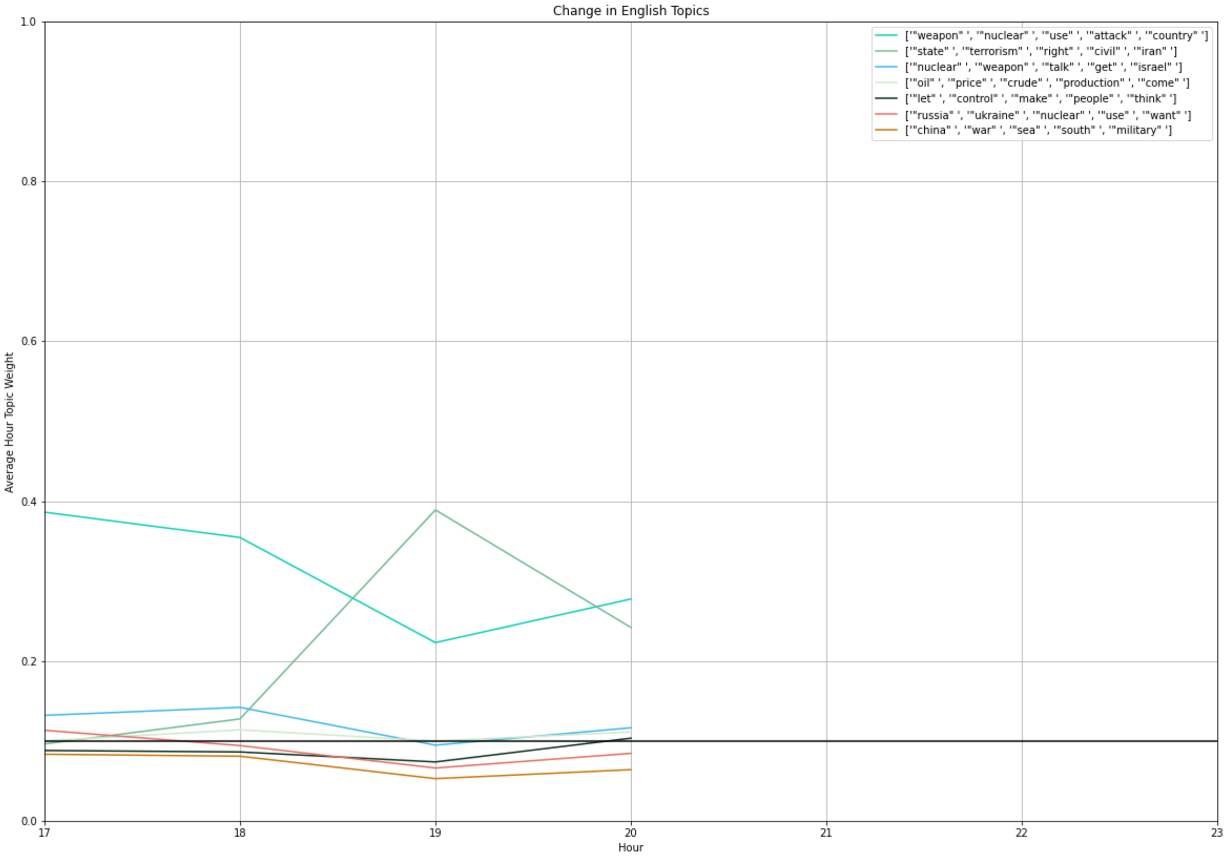


Figure 4. The English Topics that appeared and were tracked over time for June 1st, 2023. The solid black line is the median value for the average weights of the topic proportions across the three hours. We see that the Israel Nuclear Weapon topic becomes less relevant over the period. The other four topics (“oil price”, “Control”, “Ukraine War”, “South China Sea”) don’t spike in relevancy over the period. “Iran” and “Nuclear Weapon” emerge or stay relevant as described.

However, as Figure 5 below illustrates, while Iran is consistently a highly searched topic, the peak searches during this four-hour time window only surged after 10:00 PM, which is nearly three hours after the topic of "Iran" emerged on X / Twitter. Since the hourly Google Trends data can be quite noisy with potential variations between minutes, we include a polynomial trendline in orange to provide more context to the changes in search trends over time. The polynomial line facilitates a better comparison between Figure 4 and Figure 5, with Figure 4’s Iran line indicating when the topic was discussed on X / Twitter and Figure 5’s Iran polynomial line showing when the topic was searched on Google.

[11] Antony J. Blinken, Secretary of State. "Sanctioning Operatives Involved in Iranian External Plots."
[12] Shahrokhi, Sedighe. "Iran News in Brief – June 1, 2023."

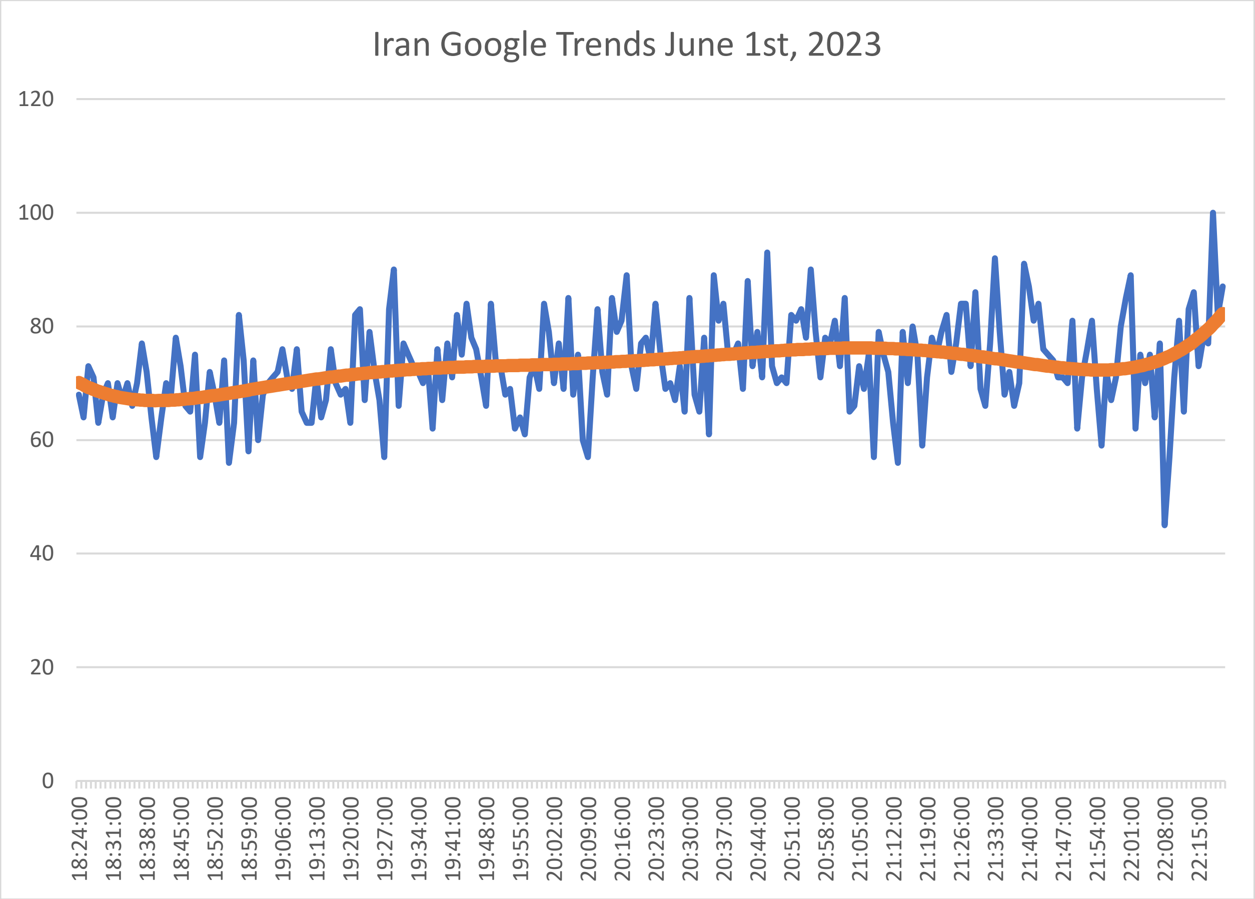


Figure 5. Google Trends tracking of "Iran" from 18:24 – 22:15 on June 1st, 2023. As Figure 4 shows, "Iran" emerged in relevance between Hours 18 and 19 (6 pm – 7 pm UTC) for X / Twitter, while Figure 5 shows how search queries from Google Trends should show queries started increasing at Hour 22 (10 pm UTC). Thus, for this example, the X / Twitter method found the emerging "Iran" topic three hours before the Google Trends query share spiked.

Appendix B shows another "Iran" topic that emerged in French tweets on June 6th, 2023, and a topic on French President "Macron" that emerged in Japanese tweets on June 6th, 2023.

Section 5.2. Daily

In contrast to our hourly analysis, we could compare the daily trends for both X / Twitter and Google on the same chart. We compare the trends on the same chart because while our X / Twitter method and Google Trends utilize different data types, they assess a similar metric: the popularity of a topic over time, thus providing comparable metrics. However, similar to our hourly analysis, we will only display results from one geopolitical topic in English, but it is also a topic that emerged across all seven languages. The results of this topic in other languages for the daily analysis can be found in Appendix C.

During our research period from February 4th, 2023, to March 23rd, 2023, North Korea launched multiple intercontinental ballistic missiles. The first launch occurred on February 18th, followed by two more missiles launched on February 20th, landing near the waters of Japan. Later, on March 9th and March 19th, North Korea conducted additional missile launches in response to joint military exercises by the US and South Korea. These missile launches triggered significant geopolitical tension and served as an ideal test case for our X / Twitter multilingual emerging topics. Figure 10 illustrates how the "North Korea" topic emerged in English during our study period and its relation to the response time of Google Trends at the daily level.

North Korea emerged as a topic on X/Twitter following four missile launches by the country. Although this does not encompass every missile launch from North Korea during this period (the total count is ten [34]), these trending topics on X/Twitter aligned with missile launches better than Google Trends' two peaks, which indicated increased search interest on February 20 and March 21. Furthermore, the X/Twitter topics appeared prior to the Google Trends topics, as evidenced by the first and third spikes in X/Twitter data, surfacing almost two days before the rise in Google Trends.

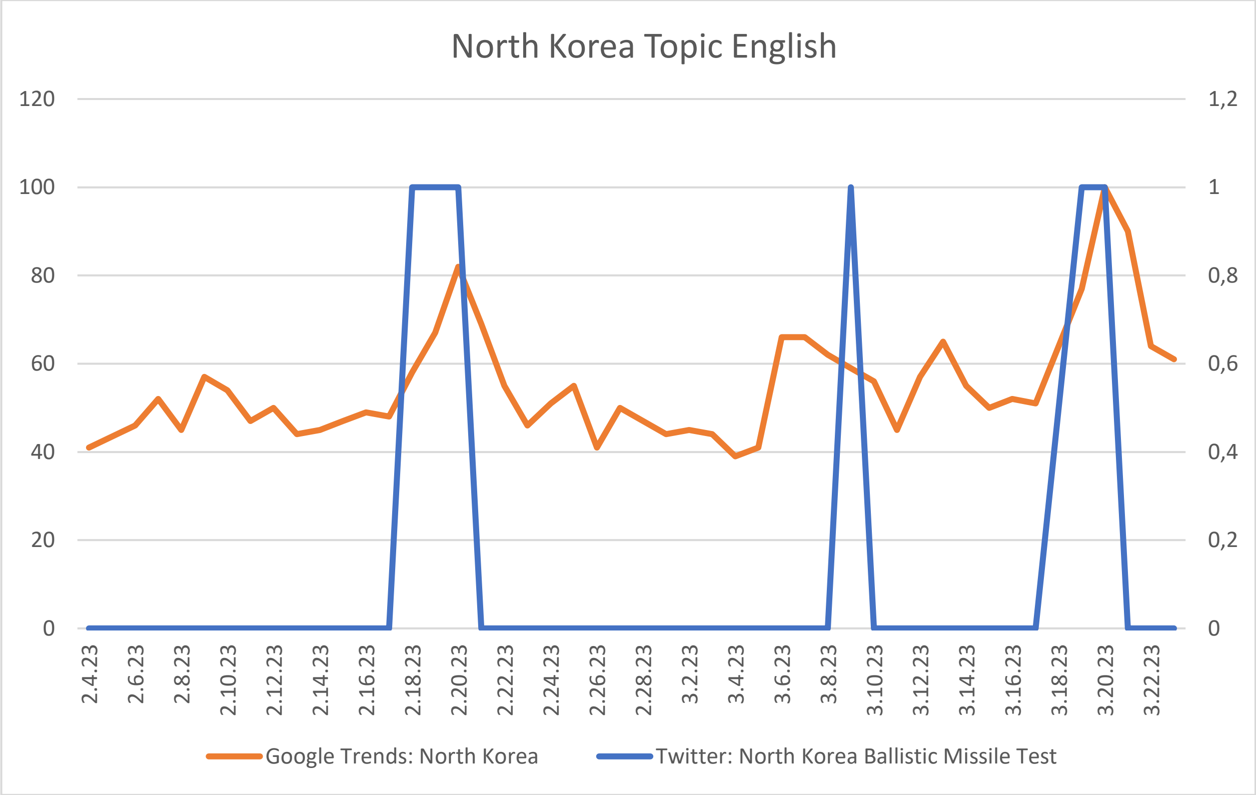


Figure 10. This compares our X / Twitter emerging geopolitical topics data for English and Google Trends search data for North Korea. In the legend, the Google Trends label is the criteria used to gather the Google Trends data. The X / Twitter legend label is the topic label used for the emerging topic on X / Twitter

Lastly, we have the results from the daily analysis of all topics generated and their respective counts, shown in the Tables in Appendix D. English had 41 unique topics, Spanish had 46 topics, French had 42 topics, Portuguese had 42 topics, Arabic had 52 topics, Japanese had 12 topics, and Korean had 19 topics. Many countries use English and French as a lingua franca thus there is an increase diversity from different parts of the world in the topics that emerged from during our study period. However, with Spanish, Portuguese, and Arabic, had many topics that focused on regional issues. Japanese and Korean had less diversity in topics, but an increase on “Crude Oil” and “Nuclear Weapons”.

4. Discussion

Our results align with Rill, et. al, 2014, who found “that new topics appearing in X / Twitter can be detected right after their occurrence. Moreover, we have compared our results to those of Google Trends. We observed that the topics emerged earlier in X / Twitter than in Google Trends.” (Rill, et. al, 2014, p. 1). We also matched the results from Lee, et. al, 2017, who showed that using X / Twitter produced more accurate and earlier results for influenza spread than Google Flu Trends. While we shared similar findings on the daily level for Rill, et al., we also went to a smaller time interval. We showed that X / Twitter can outperform Google Trends in capturing emerging topics hourly. Additionally, we found that topics that emerged before our time frame appeared, such as the Nova Kakhovka dam destruction on June 6th, 2023, would also appear in our analysis[13]. The explosion at the dam occurred before our three–hour interval for that day, but we could see that the topic was still popular on X / Twitter, as shown in Figure 17. This was also reflected in the Google Trends data where the search term remained high throughout the capture period as seen in Figure 18. There are a few potential reasons for this difference in response time to geopolitical events between X /Twitter and Google Trends. As described by Martin in 2023: “Twitter is the most popular social platform for news and current events. X /Twitter’s appeal has always been its short-form, real-time nature and that is still true today with 61.2% of people saying X / Twitter is where they go to stay updated with news and events” (Martin, 2023). As we are tracking emerging geopolitical events, it is possible that people first hear about the event on X / Twitter. Since

[13] Faulconbridge, Guy. "Nova Kakhovka Dam Breach: What Do We Know So Far?"

X /Twitter's micro-blogging nature does not provide much information, people might go to Google afterwards to learn more about the event. This would lead to our X / Twitter topic modeling program outperforming Google Trends as the information would get out on X / Twitter first. At the same time, Google Trends only captures search volume, which would only spike after people found out about the geopolitical event and started searching for more information.

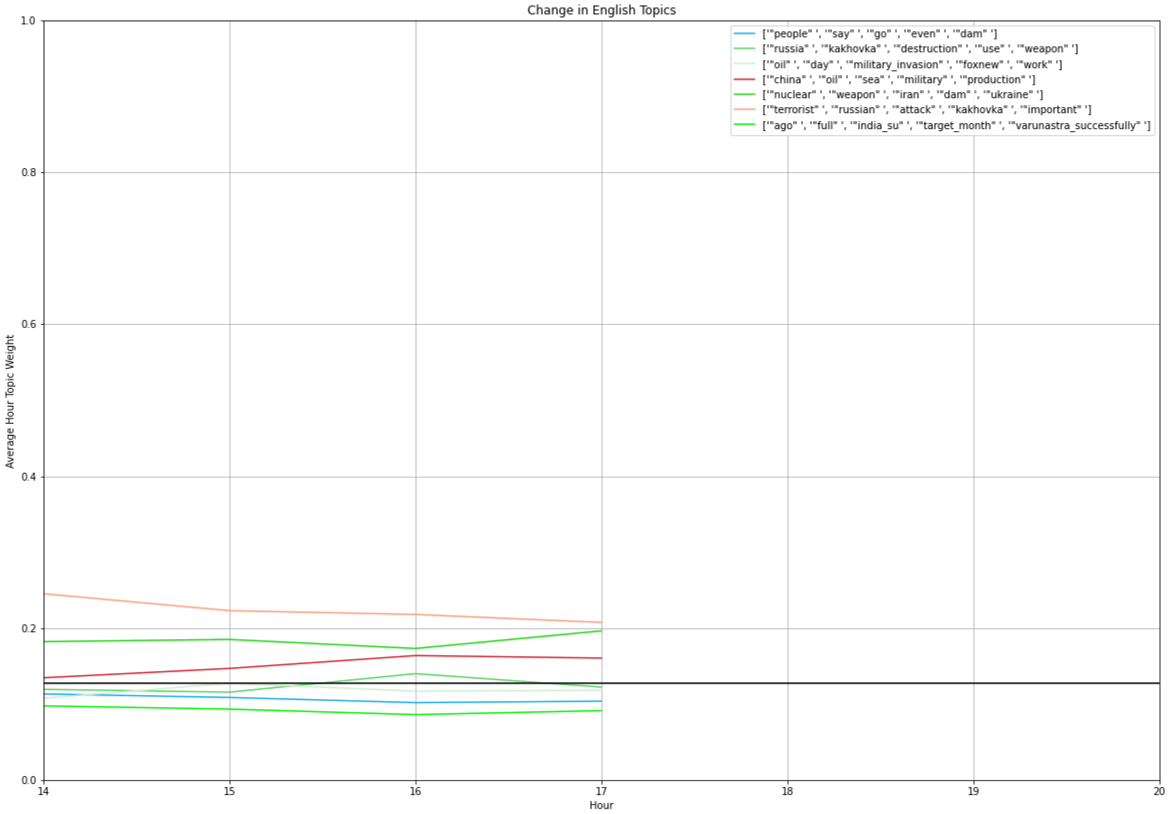


Figure 17. These are the English topics for June 6th, 2023, from 14:00 to 17:00 GMT, the Kakhovka Dam topic is represented by the Orange-Peach line. This line is far above the black median average weight line, showing that it is still a very relevant topic

Fukuhara et al. found that "in the context of a cross-lingual concern analysis, identifying common concerns across languages and recognizing unique concerns in a specific language are important… For unique (domestic or monolingual) concerns, cultural events in each country and domestic problems within a country might be included." (Fukuhara et al., 2005, p. 3). Many geopolitical topics in daily analysis appeared across all languages as "common concerns," such as the Ukraine War or North Korea. English, French, Spanish, and Portuguese had the Ukraine War as their most discussed topic during our study period. Unsurprisingly, Ukraine remains one of the most impactful geopolitical events of the 21st century. However, other topics emerged in more regional languages. For example, in English discussions, the "South China Sea" was the second most frequently mentioned topic during our study period. This reflects anxiety in English-speaking countries about how the South China Sea might develop into a potential geopolitical flashpoint soon, as tensions between China and the U.S. remain in flux14. Arabic featured various topics, mainly on the relationships between Arabic-speaking countries and other major nations. Given that the Arabic-speaking region has experienced geopolitical turmoil for decades and many countries influence outcomes, the topic diversity is not surprising considering geopolitics is significant in this context.

Crude oil was a major topic for Korea and Japan, as these subjects play a significant role in their economies. Japan imports nearly 97% of its oil[15] from the Middle East, while South Korea imports nearly all of its crude oil and natural gas[16]. Thus, it is understandable that most geopolitically focused tweets would center

[14] Hass, Ryan. "How Biden Could "Thaw" Us Relations with China.".
[15] "Total of 96.6% of Japan's March Imports of Oil Came from Arab Countries, Led by Saudi Crude."
[16] South Korea: 2021 Primary Energy Data in Quadrillion Btu."

on crude oil, a necessary commodity for nations. The table below presents the number of tweets collected for the daily study period, broken down by count for each language and their proportion of the total.

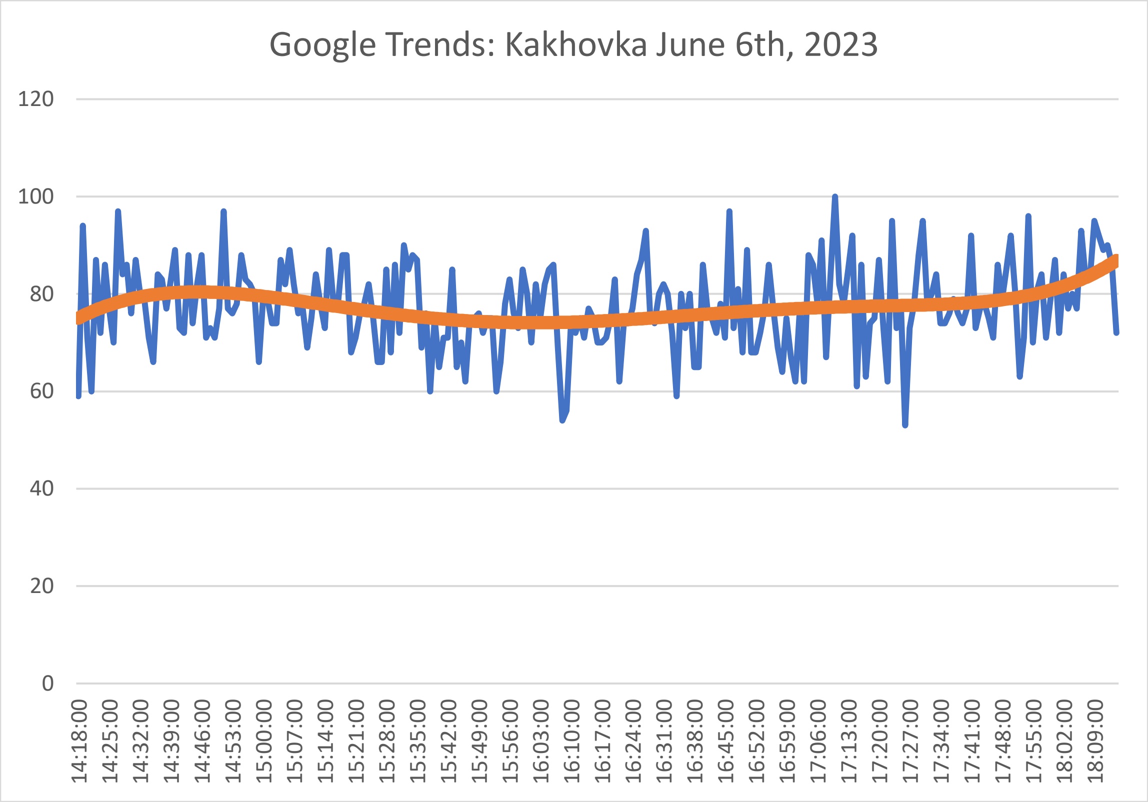


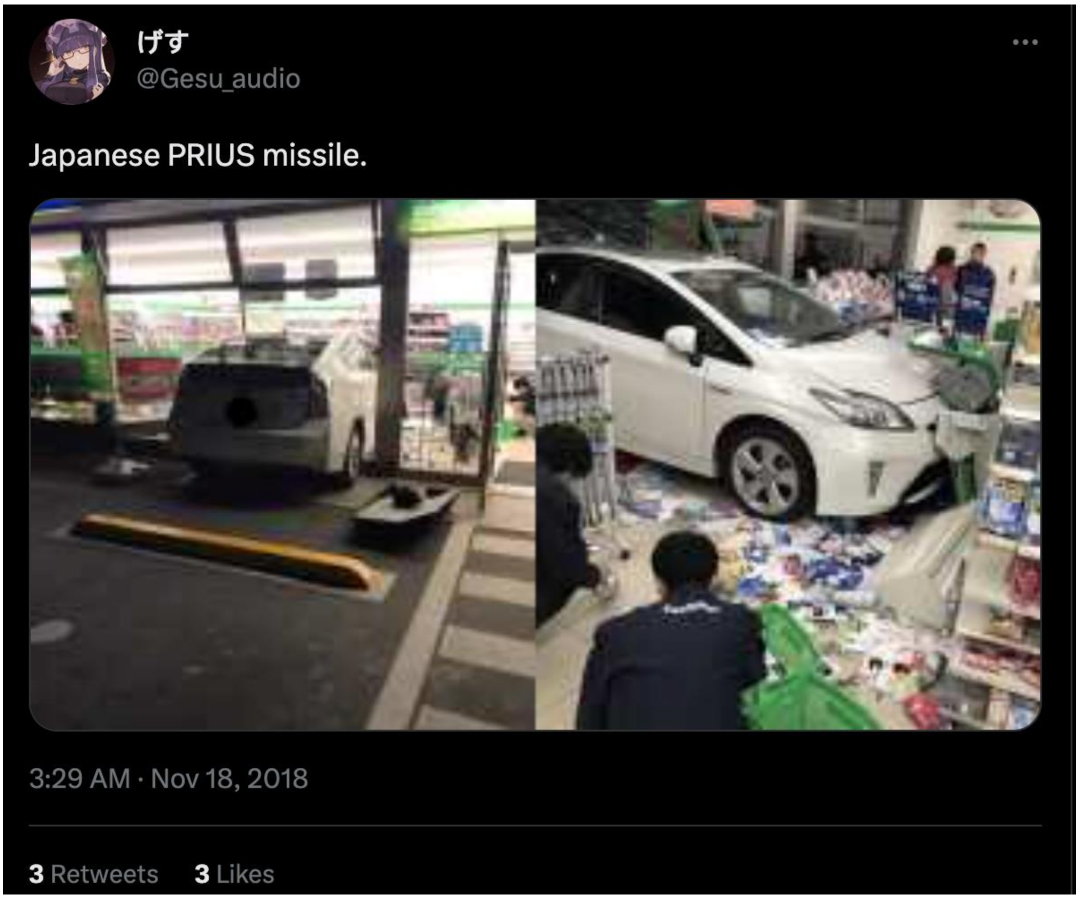
Figure 18. These are the Google Trend results for “Kakhovka” on June 6th, 2023. The Google search results stayed high, with the ratio above 60 for the entire timeframe, indicating increased interest in the geopolitical event.

Table 9 The breakdown by language of collected tweets for the daily analysis.

English	Spanish	French	Portuguese	Arabic	Japanese	Korean	Total
171,964	71,426	13,220	13,650	7,633	56,533	11,333	345,759
49.7%	20.7%	3.8%	3.9%	2.2%	16.4%	3.3%	100%

The top three languages collected were English, Spanish, and Japanese, which aligns with global usage rates (Vicinitas, 2018) and made up 86.7% of tweets collected. This helps support our methodology because if we had not separated the tweets by language, many of the topics generated by French, Portuguese, Arabic, and Korean would never have been noticed. The hourly collected tweets had similar proportions.

We also encountered a few notable data errors during our analyses. For example, in Korean, the term "재" (jae), which means "re" (as in a reply to a tweet), and "유가" (yuga), which means "oil price," occur frequently in our analysis. These terms combined as "재유가” (jaeyuga) often appeared in our topics and should translate to “re: oil price." However, this does not happen, as Google Translate leaves it as "jaeyuga," which is an error. We do not view this as a significant issue, since we can interpret “jaeyuga” as a topic involving oil. However, it is noteworthy that while Google Translate is very useful in our analyses, it can be inaccurate for some languages. One study showed that Google Translate had an 82.5% accuracy rate for Korean (Taira et al., 2021), indicating room for improvement. Another error arises from Japanese, with the terms “Prius” and “missile" co-occurring in many Japanese topics, which confused us. We found on X/Twitter that in Japan, the term “Prius missile” is a slang phrase used to describe lead-footed drivers who accidentally drive their cars into buildings while speeding, as shown in Figure 19 (Gesu_audio, 2018; Ryall, 2023). We also deemed this error not too detrimental, as Rill et al. explain: "with...search terms, some tweets without a political context have been collected. This occurs whenever a term is used with polysemous meanings" (Rill et al., 2014, p. 5). Rill et al. did not consider this a significant issue, nor did we, as it did not greatly affect our analysis.



げす
@Gesu_audio

Japanese PRIUS missile.

3:29 AM · Nov 18, 2018

3 Retweets 3 Likes

Figure 19. Tweet by Gesu_audio about the Japanese Prius missile that caused our data collection error

Overall, the main strength of our study lies in our ability to capture a wide range of geopolitical topics, from global to regional issues, as shown in Appendix D. This has not been accomplished before in this context. We tracked these topics from their emergence to periods of disinterest. Additionally, we conducted this analysis over multiple time frames, including real-time assessments with hourly updates. With this near real-time monitoring across various languages, it becomes possible to track geopolitical events as they emerge globally and observe how they transform and grow over time, potentially evolving from regional to global significance. However, our methodology also has some weaknesses. We were unable to obtain a complete list of geopolitical events. This is evident in the "North Korea" topics in Appendix C, where different languages yielded varying topics for "North Korea," and none captured all of the North Korean missile launches. Furthermore, our emerging geopolitical trends do not always identify trends before Google Trends. For instance, when Russia withdrew from the START treaty[17], our Portuguese emerging geopolitical topics detected it on the same day Google Trends did, as shown in Figure 20. Nevertheless, since we can identify some topics before they appear on Google Trends, utilizing topic modeling with X / Twitter to capture emerging geopolitical events remains a valuable source of information.

[17] https://www.reuters.com/world/europe/putin-russia-suspends-participation-last-remaining-nuclear-treaty-with-us-2023-02-21/#:~:text=MOSCOW%2C%20Feb%2021%20(Reuters),two%20sides'%20strategic%20nuclear%20arsenals.

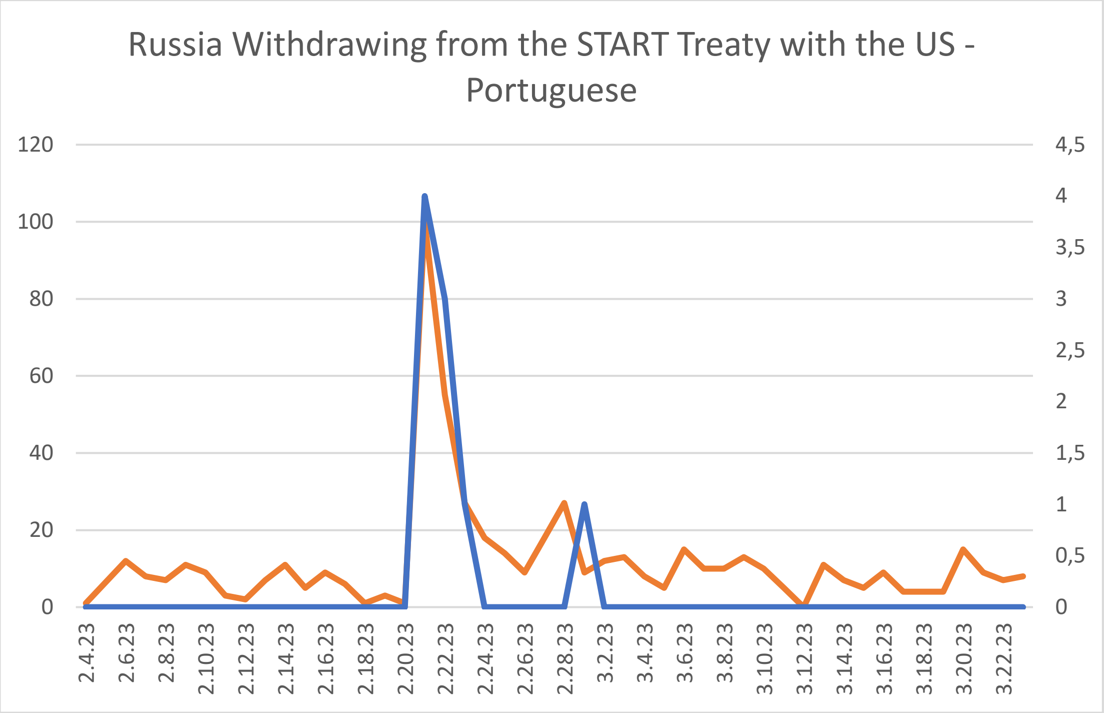


Figure 20. This is the Portuguese topic for Russia withdrawing from the START nuclear arms treaty with the U.S. As the chart shows, the Google trends spike at the same time as the X / Twitter emerging topics.

An additional weakness comes from the limitation on capping the maximum number of topics that could be generated in our algorithm. Thus, we could not capture all geopolitical events for a region. However, we believe that more topics could allow for more geopolitical events to be captured with our algorithm. We did this because we were optimizing for speed of our dynamic hyperparameter selection process as we were aiming to make sure it would complete within an hour for our analyses. Our methodology is computationally heavy; thus, our methodology could be improved by either greater computational resources or improving the dynamic hyperparameter selection criteria.

Lastly, it should be mentioned that the topic labels can potentially lead to some misinterpretation. Our methodology leaves the understanding of the labels to the reader of the label, thus there is some room for human error when trying to comprehend what they are referring to. Luckily, this issue is somewhat mitigated by having the time the topic was formed and the key words from the labels to cross-reference event referred to by the topic, however, it is not a hundred percent accurate.

5. Conclusion & Practical Implications

Our X / Twitter Emerging Geopolitical Topics algorithm provides new insights into the growth of geopolitical events over time. There are a few practical implications for our work. First, with a faster understanding of emerging events, governments, NGOs, and businesses can respond quicker, allowing for better response to the event. This could affect global stock markets (for business) or gathering local resources for events that affect specific regions. Second, our algorithm allows for pairing down of the massive amount of data produced on X / Twitter, this would be useful for media to pick up on relevant stories out of the noise of social media.

We see a few directions for our future work. One is improving a limitation of our study: expanding the number of languages we analyze. We did not include German, Russian, Chinese, or Amharic, as they were not as popular on X / Twitter. However, these languages can be filtered by the Filter Stream X / Twitter API. They could provide insights into different regions of the world that we did not investigate in our study. Additionally, we identified other geopolitical topics we chose not to include in this study, such as “Trade Agreements” from Klement or “Monetary Policy” from Baker, et al., that could provide insight into other geopolitical events we did not find. Lastly, we could increase the number of key bigrams for each topic, allowing for more tweets to be gathered for each topic and analysis below the hourly level.

The article attachments can be accessed from the link. https://doi.org/10.29329/jsomer.14

Statement of Researchers

Researchers' contribution rate statement:
John Corcoran Burns: Conceptualization; Data curation; Formal analysis; Investigation;Methodology; Project administration; Software; Validation; Visualization; Writing - original draft; **Tom Kelsey:** Funding acquisition; Project administration; Supervision; Writing - review – editing; **Carl Donovan:** Funding acquisition; Project administration; Supervision; Writing - review - editing
Conflict statement:
The authors declared no conflicts of interest.
Data Availability Statement:
Data can be made available upon request. Please contact the corresponding author.
Acknowledgements:
We would like to thank Irlanda Márquez Cruz for her Spanish translations of the 'Goldstein Index' bigrams for this study. Also, the translators at Gengo.com for their work for the French, Portuguese, Arabic, Japanese, and Korean.
Funding:
This work was supported by the School of Computer Science of the University of St Andrews.
Ethical Considerations:
Approved by the School of Computer Science Ethics Committee for the University Teaching and Research Ethics Committee (UTREC) at the University of St Andrews.
Use of AI Statement:
The authors submit that no generative AI tools or models were not used in any part of this study.

Author Biographies
John Corcoran Burns, PhD Graduate from the School of Computer Science at the University of St Andrews. Specializes in natural language processing, multilingual analysis, and geopolitical risk.
Tom Kelsey, Professor in the School of Computer Science at the University of St Andrews. Specializes in data-driven research, using advanced computational techniques to identify, retrieve, classify, and analyze data to generate hypotheses and develop models. His primary interest is the modelling of human fertility both for the healthy population and for survivors of childhood cancer.
Carl Donovan, Honorary Research Statistician and Lecture in the School of Mathematics and Statistics at the University of St Andrews and the Director and Consultant at DMP Statistical Solutions. Specializes in statistical modelling with an interest in Earth and Planetary Sciences.

6. References

2018 Research on 100 Million Tweets: What It Means for Your Social Media Strategy for Twitter. (2018). Vicinitas https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets#language.

Google Trends. (2025) Google https://trends.google.com/trends/?geo=US.

South Korea: 2021 Primary Energy Data in Quadrillion Btu. (2022). International, U.S Energy Information Administration. https://www.eia.gov/international/overview/country/KOR.

Total of 96.6% of Japan's March Imports of Oil Came from Arab Countries, Led by Saudi Crude. (2023). Arab News, Arab News, https://www.arabnews.com/node/2294816/business-economy.

The Use of Arabic Diacritics. (2023). International House Cairo https://ihcairoeg.com/arabic/the-use-of-arabic-diacritics/.

Ahmed, A., & Xing, E.P. (2010). Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010).* http://dx.doi.org/10.48550/arxiv.1203.3463.

Allahyari, M., & Kochut, K. (2015). Automatic Topic Labeling Using Ontology-Based Topic Models. *14th {IEEE} International Conference on Machine Learning and Applications, ICMLA 2015*, December 9-11, 2015, pp. 259–64. doi:10.1109/ICMLA.2015.88.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic Detection and Tracking Pilot Study Final Report. *Broadcast News Understanding and Transcription Workshop*, pp. 194-218. http://ciir.cs.umass.edu/pubfiles/ir-137.pdf.

Amara, A., Taieb, M. A. H., & Aouicha, M. B. (2021). Multilingual Topic Modeling for Tracking Covid-19 Trends Based on Facebook Data Analysis. *Applied Intelligence*, 51(5), 3052-3073. http://dx.doi.org/10.1007/s10489-020-02033-3.

A. J. Blinken, Secretary of State. (2023). Sanctioning Operatives Involved in Iranian External Plots. *U.S Department of State*, June 1, 2023. https://www.state.gov/sanctioning-operatives-involved-in-iranian-external-plots/.

Atefeth, F., & Khreich, W. (2013). A Survery of Techniques for Event Detection in Twitter. *Computational Intelligence*, vol. 0, no. 0, 2013, https://citeseerx.ist.psu.edu/viewdoc/download

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, vol. 131, no. 4, 2016, pp. 1593-636, http://dx.doi.org/10.1093/qje/qjw024.

Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding State Preferences with Text as Data: Introducing the Un General Debate Corpus. *Research & Politics*, vol. 4, no. 2, 2017, http://dx.doi.org/10.1177/2053168017712821.

Becker, H., Naaman, M., & Gravano, L. (2021). Beyond Trending Topics: Real-World Event Identification on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, 2021, pp. 438-41, doi:10.1609/icwsm.v5i1.14146.

Bekaert, G., Harvey, C. R., Lundblad, C. T., & Siegel, S. (2015). Political Risk and International Valuation (December 8, 2015). *Columbia Business School Research Paper*, No. 15-83, Available at SSRN: https://ssrn.com/abstract=2659257 or http://dx.doi.org/10.2139/ssrn.2659257

Besaw, C. (2021). Topic Modeling as Osint: Exploring Russian Presidential Speech Topics over Time. *medium.com* https://medium.com/the-die-is-forecast/topic-modeling-as-osint-exploring-russian-presidential-speech-topics-over-time-ad6018286d37.

Blei, D. M., & Lafferty J. D. (2006). Dynamic Topic Models. *ACM Press*, 2006. http://dx.doi.org/10.1145/1143844.1143859.

Blei, D. M, Ng A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993 - 1022, https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual Topic Models for Unaligned Text. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, June 18 - 21 2009. doi:10.48550/arxiv.1205.2657.

Bremmer, I., & Kupchan., C. (2023). Top Risk 2023, *Eurasia Group*, January 3rd, 2023, pp. 1-30. general editor, Eurasia Group, https://www.eurasiagroup.net/issues/top-risks-2023.

Brewer, B. J. (2016). Stats 331: Introduction to Bayesian Statistics. *Department of Statistics, University of Auckland,* 2016.

Caldara, D., & Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review*, vol. 112, no. 4, pp. 1194-225, http://dx.doi.org/10.1257/aer.20191823.

Chen, X., Liu, Z., Wei, L., Yan, J., Hao, T., & Ding, R. (2018). A Comparative Quantitative Study of Utilizing Artificial Intelligence on Electronic Health Records in the USA and China During 2008–2017. *BMC Medical Informatics and Decision Making*, vol. 18, no. S5, 2018, http://dx.doi.org/10.1186/s12911-018-0692-9.

Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. *Google Inc.,* 2009, pp. 1-23. https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf.

Choi, S. H., & Shin, H. (2023). North Korea Fires Two More Missiles into Its Pacific 'Firing Range'. *Reuters*, February 20th, 2023. https://www.reuters.com/world/asia-pacific/north-korea-fires-ballistic-missile-south-korea-military-2023-02-19/.

Contributors, Wikipedia. (2023) 1973 Oil Crisis. https://en.wikipedia.org/w/index.php?title=1973_oil_crisis&oldid=1161514686.

Contributors, Wikipedia. (2023) China–United States Relations. https://en.wikipedia.org/w/index.php?title=China%E2%80%93United_States_relations&oldid=1162607 472.

Contributors, Wikipedia. (2023) Cyberwarfare. https://en.wikipedia.org/w/index.php?title=Cyberwarfare&oldid=1160321593. A

Contributors, Wikipedia. (2023) Nuclear Warfare. https://en.wikipedia.org/w/index.php?title=Nuclear_warfare&oldid=1161701347.

Contributors, Wikipedia. (2023) Terrorism. https://en.wikipedia.org/w/index.php?title=Terrorism&oldid=1160928625.

Cordeiro, M. (2012). Twitter Event Detection: Combining Wavelet Analysis and Topic Inference Summarization. *DSIE'12, the Doctoral Symposium on Informatics Engineering*, https://www.researchgate.net/publication/281454053_Twitter_event_detection_combining_wavelet_analysis_and_topic_inference_summarization.

Culotta, A. (2010). Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. *Workshop on Social Media Analysis In conjunction with the International Conference on Knowledge Discovery & Data Mining (KDD 2010),* July 25, 2010. http://snap.stanford.edu/soma2010/papers/soma2010_16.pdf.

Danilak, M. M. (2021). Langdetect. *GitHub* https://github.com/Mimino666/langdetect. 2021.

De Guzman, C. (2023). North Korea Is Ramping up Its Missile Tests. How Worried Should We Be? World - North Korea. *Time* https://time.com/6266737/north-korea-ballistic-missile-tests-2023/.

Duarte, R. (2025). Top 35 Social Media Platform (2025) Exploding Topics https://explodingtopics.com/blog/top-social-media-platforms

Faulconbridge, G. (2023)/ Nova Kakhovka Dam Breach: What Do We Know So Far? *Reuters,* June 7th 2023. https://www.reuters.com/world/europe/what-is-kakhovka-dam-ukraine-what-happened-2023- 06-07/.

Flint, C. (2022). Introduction to Geopolitics. Routledge, *Taylor et Francis Group*.

Fukuhara, T., Utsuro, T., & Nakagawa, H. (2005). Cross-Lingual Concern Analysis from Multilingual Weblog Articles. *University of Tokyo* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=25b75750796c32780c044a791dfafc5105e73d5b.

Fukumasu, K., Eguchi K., & Xing, E. (2012). Symmetric Correspondence Topic Models for Multilingual Text Analysis. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, edited by F. Pereira and C.J. Burges and L. Bottou and K.Q. Weinberger. http://www.cs.cmu.edu/~epxing/papers/2012/Fukumasu_Eguchi_Xing_NIPS12.pdf.

Goldstein, J. S. (1992). A Conflict-Cooperation Scale for Weis Events Data. *The Journal of Conflict Resolution*, vol. 36, no. 2, 1992, pp. 369-85, https://www.jstor.org/stable/174480.

Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, vol. 101, no. Supplement 1, 2004, pp. 5228-35, doi:10.1073/pnas.0307752101.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, vol. 21, no. 3, 2013, pp. 267-97, doi:10.1093/pan/mps028.

Groontendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *arXiv*, https://arxiv.org/pdf/2203.05794

Hafezi, P. (2023). Iran Presents Its First Hypersonic Ballistic Missile, State Media Reports. *Reuters,* June 6, 2023. https://www.reuters.com/world/middle-east/iran-unveils-its-first-hypersonic-ballistic-missile-state-media-reports-2023-06-06/.

Hass, R. (2023). How Biden Could "Thaw" Us Relations with China. *Brookings*, May 23rd 2023. https://www.brookings.edu/articles/how-biden-could-thaw-us-relations-with-china/#:~:text=The%20year%202023%20was%20supposed,in%20the%20fall%20of%202023.

Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic Text on Social Media. *Heliyon,* vol. 7, no. 2, 2021, http://dx.doi.org/10.1016/j.heliyon.2021.e06191.

Hida, R., Takeishi, N., Yairi, T., & Hori, K. (2018). Dynamic and Static Topic Model for Analyzing Time-Series Document Collections. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. Volume 2: Short Papers, Association for Computational Linguistics, pp. 516–20. doi:10.18653/v1/P18-2082.

Hisano, R., Sornette, D., Mizuno, T., Ohnishi, T., & Watanabe, T. (2013). High Quality Topic Extraction from Business News Explains Abnormal Financial Market Volatility. *PLoS One*, 8(6) e64846, http://dx.doi.org/10.1371/journal.pone.0064846.

Hong, L,, & Davison, B. D. (2010). Empirical Study of Topic Modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics - SOMA '10, ACM Press, July 25, 2010. doi:10.1145/1964858.1964870.

Honnibal, M., & Montani. I. (2023). Spacy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. *Explosion*, https://spacy.io/usage.

Huang, T., Wu, F., Yu, J., & Zhang, B. (2014). International Political Risk and Government Bond Pricing. Journal of Banking and Finance, Forthcoming, *26th Australasian Finance and Banking Conference 2013*, 2014, http://dx.doi.org/10.2139/ssrn.2312188.

Hurtado, J. L., Agarwal, A., & Zhu, X. (2016). Topic Discovery and Future Trend Forecasting for Texts. *Journal of Big Data*, vol. 3, no. 2, 2016, http://dx.doi.org/10.1186/s40537-016-0039-2.

Johnson, J. (2023). North Korea Fires Short-Range Missile Ahead of U.S.-South Korea Military Exercises. *The Japan Times*, March 9th 2023. https://www.japantimes.co.jp/news/2023/03/09/asia-pacific/north-korea-missile-launch-march-9/.

Kapadia, S. (2019). Evaluate Topic Models: Latent Dirichlet Allocation (Lda), *Towards Data Science* https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

Karim, A. A., & Abandah, G., (2021). On the Training of Deep Neural Networks for Automatic Arabic-Text Diacritization. *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021, pp. 276-86, https://thesai.org/Downloads/Volume12No8/Paper_32- On_the_Training_of_Deep_Neural_Networks.pdf.

Klement, J. (2021). Geo-Economics: The Interplay between Geopolitics, Economics, and Investments*. CFA Institute*. Research Foundation Books.

Koren, V. (2022). Korean Tokenization & Lemmatization. *Medium* https://korenv20.medium.com/korean-tokenization-lemmatization-a741fc9939cc.

Kravariti, A. (2023) Machine Vs Human Translation: The Pros, Cons and When to Use Each. Blog. *translate plus* https://www.translateplus.com/blog/machine-vs-human-translation-pros-cons-use/.

Kulshrestha, R. (2022) A Beginner's Guide to Latent Dirichlet Allocation (Lda). *Towards Data Science* https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2.

Lakezina, V. (2023) War Zone Villagers Flee after Massive Ukraine Dam Destroyed. *Reuters,* June 6th, 2023. https://www.reuters.com/world/europe/russia-says-it-thwarts-another-major-ukrainian-offensive-donetsk-2023-06-05/.

Lee, H., & Song, J. (2020). Understanding Recurrent Neural Network for Texts Using English-Korean Corpora. *Communications for Statistical Applications and Methods*, vol. 27, no. 3, 2020, pp. 313-26, https://doi.org/10.29220/CSAM.2020.27.3.313.

Lee, J. J., Kwon, S. B., & Ahn, S. M. (2018). Sentiment Analysis Using Deep Learning Model Based on Phoneme-Level Korean. *Journal of Information Technology Services*, vol. 17, no. 1, 2018, pp. 79-89, https://doi.org/10.9716/KITS.2018.17.1.079.

Lee, K., Agrawal, A., & Choudhary, A. (2017). Forecasting Influenza Levels Using Real-Time Social Media Streams. *IEEE, 2017*. doi:10.1109/ichi.2017.68.

Lind, F., Eberl, J. M., Galyga, S., Heidenreich, T., Boomgaarden, H. G., Jiménez, B. H., & Berganza, R. (2019). A Bridge over the Language Gap: Topic Modelling for Text Analyses across Languages for Country Comparative Research. Working Paper, *Reminder*, November 2019. https://www.reminder-project.eu/wp-content/uploads/2019/11/D8.7.pdf.

Martin, M. (2023). 29 Twitter Stats That Matter to Marketers in 2023. Strategy. *Hootsuite*, https://blog.hootsuite.com/twitter-statistics/..

Massicotte, P., & Eddelbuettel, D. (2022). Gtrendsr: Perform and Display Google Trends Queries. 1.5.1 ed., *CRAN*, https://github.com/PMassicotte/gtrendsR.

Metzler, D., Cai, C., & Hovy, E. (2012). Structured Event Retrieval over Microblog Archives. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, *Association for Computational Linguistics*, pp. 646–55.

Nidhaloff, (2020). Deep-Translator. *GitHub*, https://github.com/nidhaloff/deep-translator.

Peersman, G., & Smets, F. (2002). "The Industry Effects of Monetary Policy in Euro Area." Working Paper Series, edited by *European Central Bank*, https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp165.pdf.

Pelgrim, R. (2023). "Arabic Nlp: Unique Challenges and Their Solutions." *Medium* https://towardsdatascience.com/arabic-nlp-unique-challenges-and-their-solutions-d99e8a87893d.

Psaledakis, D, (2023). "Us Slaps Sanctions on Iranian, Chinese Targets over Tehran's Missile, Military Programs." Reuters, June 6, 2023. https://www.reuters.com/world/middle-east/us-slaps-sanctions-iranian-chinese-targets-action-over-tehrans-missile-military-2023-06-06/.

Ranaei, S., Arho, S., Porter, A., & Carley, S. (2020). Evaluating Technological Emergence Using Text Analytics: Two Case Technologies and Three Approaches. *Scientometrics*, vol. 122, no. 1, 2020, pp. 215-47, http://dx.doi.org/10.1007/s11192-019-03275-w.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* ELRA, https://pypi.org/project/gensim/.

Reimers, N., & Gurevych, I. (2019). Sentence - BERT: Sentence Embeddings using Siamese BERT - Networks, *CoRR*, https://arxiv.org/pdf/1908.10084

Reuter, K., Zhu, Y., Angyan, P., Le, N. Q., Merchant, A. A., & Zimmer, M. (2019). Public Concern About Monitoring Twitter Users and Their Conversations to Recruit for Clinical Trials: Survey Study. *Computer Science Faculty Research and Publications*. 29. https://epublications.marquette.edu/comp_fac/29

Reuters. (2023). France Plans Major Police Presence for June 6 Day of Protest. Reuters, Reuters, June 4th, 2023. https://www.reuters.com/world/europe/france-plans-major-police-presence-june-6-day-protest- 2023-06-04/.

Revert, F. (2021). An Overview of Topics Extraction in Python with Lda. *Towards Data Science* https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early Detection of Emerging Political Topics on Twitter and the Impact on Concept-Level Sentiment Analysis. *Knowledge-Based Systems*, vol. 69, pp. 24-33, https://doi.org/10.1016/j.knosys.2014.05.008.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The Structural Topic Model and Applied Social Science. *Neural Information Processing Society*, https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf.

Ryall, J. (2023). Older Drivers in Japan Are Mixing up Their Accelerate and Brake Pedals with Fatal Results. East Asia. *South China Morning Press* https://www.scmp.com/news/asia/east-asia/article/2050472/older-drivers-japan-are-mistaking-stop-and-go-fatal-results.

Sakamoto, T. & Takikawa, H. (2017). Cross-National Measurement of Polarization in Political Discourse: Analyzing Floor Debate in the U.S. The Japanese Legislatures. *IEEE International Conference on Big Data (Big Data), IEEE*, doi:10.1109/bigdata.2017.8258285.

Seo, Y., Lendon, B., & Ogura, J. (2023). North Korea Says It Tested Icbm in Surprise Drill. *CNN*, February 18th, 2023. https://www.cnn.com/2023/02/18/asia/north-korea-missile-launch-intl-hnk-ml/index.html.

Shahrokhi, S. (2023). Iran News in Brief – June 1, 2023. *National Council of Resistance of Iran* https://www.ncr-iran.org/en/news/iran-news-in-brief-june-1-2023/.

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*, vol. 36, no. 11, pp. 3361-5, doi: 10.1007/s11606-021-06666-z.

Tetlock, P. C., Saar-Tsechansky, M., & MacSkassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, vol. 63, no. 3, pp. 1437-67, doi:10.1111/j.1540-6261.2008.01362.x.

Twitter. (2022). Filtered Stream. *Twitter.* https://developer.twitter.com/en/docs/twitter-api/tweets/filteredstream/integrate/build-a-rule.

Urquhart, A. (2018). What Causes the Attention of Bitcoin? *Economics Letters*, vol. 166, pp. 40-44, doi:10.1016/j.econlet.2018.02.017.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation Methods for Topic Models. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, ACM Press, doi:10.1145/1553374.1553515.

Wang, X., & McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *Conference on Knowledge Discovery and Data Mining (KDD), edited by ACM, ACM*. https://people.cs.umass.edu/~mccallum/papers/tot-kdd06s.pdf.

Wieringa, J. (2023). Ways to Compute Topics over Time, Part 1. https://jeriwieringa.com/2017/06/21/Calculating-and-Visualizing-Topic-Significance-over-Time-Part- 1/.

Yang, W, Boyd-Graber, J., & Resnik, P. (2019). A Multilingual Topic Model for Learning Weighted Topic Links across Corpora with Low Comparability. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics*, pp. 1243–48. doi:10.18653/v1/d19.

Yim, H. (2023). North Korea Fires Ballistic Missile as U.S.-South Korean Drills Go On. *Reuters*, March 19th, 2023. https://www.reuters.com/world/asia-pacific/north-korea-fired-ballistic-missile-towards-sea-off-east-coast-yonhap-2023-03-19/.

Yuan, M., Van Durme, B., & Boyd-Graber, J. (2018). Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages. *32nd Conference on Neural Information Processing System (NeurIPS 2018),* https://forest-snow.github.io/docs/2018_nips_mtanchor.pdf

Yuning, C., & Lianzhong, L. (2016). "Development and Research of Topic Detection and Tracking." *IEEE,* doi:10.1109/icsess.2016.7883041.

Zheng, L., Nie, T., Moriya, I., Inoue, Y., Imada, T., Utsuro, T., Kawada, Y., & Kando, N. (2014). Comparative Topic Analysis of Japanese and Chinese Bloggers. *28th International Conference on Advanced Information Networking and Applications Workshops, IEEE*, doi:10.1109/waina.2014.107.

Zosa, E., & Pivovarova, L. (2022). Multilingual and Multimodal Topic Modelling with Pretrained Embedding. *Proceedings of the 29th International Conference on Computational Linguistics*, ACL, October 12th-17th. https://aclanthology.org/2022.coling-1.355.pdf

@Gesu_audio. (2018). Japanese Prius Missile. *Twitter*, November 18th, 2018. https://twitter.com/Gesu_audio/status/1064073144355287041.