Project-11:

# Analysis of political events using Twitter

Marzie Faramarzzadeh

✓*GituHub page of the project*: https://github.com/aiefaramarzzadeh/Natural-Language-Processing-project.git

**Abstract**— Russia has invaded Ukrain in Feburary 2022 which led to an ongoing war between two countries. This war has stimulated many reactions in twitter. In this report, I retrieved 240k English tweets in March 2022 about the war to see how people feel and discuss about it. The first step in this study was categorizing tweets based on their sentiment score (i.e., valence state) into three negative, neutral and postive datasets. Then, I used Latent Dirichlet Allocation (LDA) and Empath clustering methods for detecting discussed topics in each of these three datasets. Also, the most frequent words and named-entities in each of these three datasets were visualized by word cloud and histogram. I found that 50% of tweets had negative sentiment score. Also, 15% and 35% of them were categorized in the neutral and positive datasets respectively. Thus, people majorly have reacted negatively to this war. In the positive dataset, prodominat topics were calling for peace and support from international organizations (e.g., United Nation). Also, some words representing resistance againt this invations (e.g., "stand") were frequent in co-occuring with "Russia" word in the positive dataset. I found that LDA and Empath have strong potential in detecting discussed topics. For example, Empath detected topics related to war, death, suffering, pain, etc. in the negative dataset. Having similarities with Empath, LDA clustered negative dataset tweets into a cluster having frequent words such as zone exclusion, suffering, radiation, etc. All defined specifications for the project were achieved.

**Keywords**— Natural Language Processing, NLTK, Latent Dirichlet Allocation, Empath Client categories, topic modeling, named-entity recognition

## I. INTRODUCTION

Russia invaded Ukraine in February 2022, and this country faced immediate backlash and condemnation from the world. The war led to humanitarian and refugee crisis in Ukraine. About 7.7 million Ukrainians have fled (https://en.wikipedia.org/wiki/2022_Ukrainian_refugee_crisis). Also, in the online space another war emerged to use social media for supporting Ukraine people and information warfare. Many studies focused on text data available in internet and used Natural Language Processing (NLP) and Text Mining techniques to understand how people feel and react to this war.

Chen et al. (2022) presented a collection of over 63 million tweets (https://github.com/echen102/ukraine-russia), in the first month of the war being maintained and regularly updated. They showed evidence of public engagement with Russian state sponsored media and other domains that are known to push unreliable information (Chen & Ferrara, 2022). Polyzos (2022) proposed the utilization of social media related to the war in Ukraine to understand public's perception of the progression of events. Using sentiment analysis on 42 million tweets , they found that currencies and markets in Europe experienced an immediate negative response. However, US stock markets was unaffected, and the US Dollar affected positively after the war (Polyzos, 2022).

Ukraine conflict roots back to 2014. Following the 2014 Ukrainian Revolution, Russia annexed Crimea, and seized part of the Donbas region (south-eastern of Ukraine). In another study in 2015, Makhortykh & Lyebyedyev focused on #SaveDonbassPeople hashtag which is an online protest campaign against the military operation of Russia in Donbas. This movement was on Twitter. They found that Twitter was used as a propaganda outlet to broadcast opposing views on the ongoing conflict (Makhortykh & Lyebyedyev, 2015). The invasion of Russia has opponent inside this country. In the social sciences, "framing" is to understand how individuals, groups, and societies organize, perceive, and communicate about reality. A study by Nikolayenko using "framing" approach studied the case of a Peace March held in Moscow on 21 September 2014 to examine how antiwar activists and their opponents framed a protest against Russia's intervention in Ukraine. Based on this study analysis of tweets on the eve of the march, peace activists present themselves as highly moral citizens with healthy dose of patriotism Therefore, they criticized the Russian government for attacking Ukraine. On the other hand, their opponents define themselves as real patriots, so they accused people who denied Russia's military presence in Ukraine as traitors (Nikolayenko, 2019).

In this study, I considered Kaggle Twitter data source (53.33M tweets) which is on Ukraine war. As the dataset is big (over 14 GB), so I took only a sample of 340k tweets in March 2022 (200 MB). Then, 240k English tweets were extracted. The Distribution of tweets frequency based on location is shown in Figure 1. As shown in this figure, 75% of tweets are from USA (27%), UK (10%), Germany (8%), Ukraine (7%), Italy (6%), Canada (5%), France (5%), India (4%) and Poland (3%). Interesting point about this figure is high contribution of Finland (2%) despite its low population. NLP goal is Knowledge Discovery in Databases (a.k.a. KDD) to present information and patterns from unstructured or semi-structured text data (Card et al., 2009). As mentioned in similar case studies, NLP is utilized

in interpreting text data. Therefore, here I utilized similar techniques such as histogram of most frequent words, sentiment analysis, topic modeling, frequent co-occurring words, and frequent named-entity to find how people all over the world reacted to Russia invasion.


| Country | Sum of Frequency | percentage |
|---|---|---|
| USA | 16052 | 27.14 |
| UK | 5925 | 10.02 |
| Germany | 4692 | 7.93 |
| Ukraine | 4356 | 7.37 |
| Italy | 3630 | 6.14 |
| Canada | 3046 | 5.15 |
| France | 2944 | 4.98 |
| India | 2443 | 4.13 |
| Poland | 1671 | 2.83 |
| Finland | 1139 | 1.93 |
| Netherlands | 923 | 1.56 |
| Venezuela | 897 | 1.52 |
| Belgium | 786 | 1.33 |

FIGURE 1. DISTRIBUTION OF THIS STUDY TWEETS FREQUENCY BASED ON LOCATION

## I. METHODOLOGY

Using retrieved tweets and based on sentiment analysis (using NLTK python package), I built negative (sentiment score < 0), positive (sentiment score > 0) and neutral (sentiment score = 0) datasets. Preprocessing is important to have cleaned dataset from symbols, non-meaningful words and stopwords. Then, using Latent Dirichlet Allocation (LDA) topic modeling from Gensim python package, I determined 10 topics based on 10 words in each positive, negative and neutral preprocessed dataset. Visualization of words in each of these three datasets is informative, so I visualized most frequent words in the tokenized preprocessed positive, negative and neutral datasets using histogram plot and word cloud. Also, Empath categorization technique to find relevant topics to each of three datasets was applied. Finally, I used Named-entity Recognition (NER) tool from Spacy python package to filter named-entities in each dataset and plotted the most frequent ones. The pipeline that I followed in this study is shown in Figure 2. More detail on each method is discussed in the following sections. Words are lowered in this project.

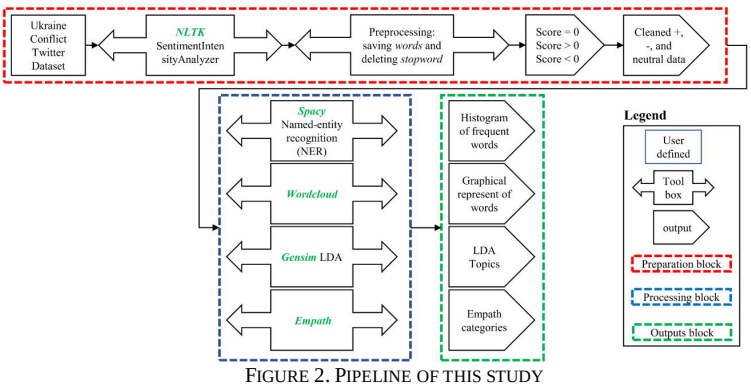


FIGURE 2. PIPELINE OF THIS STUDY

### A. *Sentiment analysis*

For determining sentiment score of each tweet, I used NLTK python package which contains `SentimentIntensityAnalyzer` class. Inside this class, there is a function named as `polarity_scores` which returns the sentiment score of input sentence. Output score of this function is in the range of -1 and 1 (more positive values represent positive valence of input text). For example, sentiment score of "VADER is VERY SMART, really handsome, and INCREDIBLY FUNNY!!!" is 0.94, but sentiment score of "The plot was good, but the characters are uncompelling and the dialog is not great." is -0.71 (for more detail see *ti.py* file in the GitHub page of the project: https://github.com/aiefaramarzzadeh/Natural-Language-Processing-project.git).

### B. *Latent Dirichlet Allocation*

Latent Dirichlet Allocation is for topic detection in a text corpus. It treats corpus as probabilistic distribution sets of words/topics. To calculate ranked lists of words associated with a given word $w_n$, set of topics generated by LDA is considered. Thus, for each word, the sum of the weight of each topic multiplied by the weight of given word in each topic must be calculated. For N topics the ranking weight for the word *i* is computed as:

$$w_i = \sum_{j=1}^{N} w_{ij} * w_{nj} \qquad (1)$$

where and $w_{ij}$ is the weight of the word *i* in the topic *j*. This representation is for calculation of a ranked list of words co-occurred with a given word based on their probability of co-occurrence in text corpus (Korzycki et al., 2017). In other words, LDA assumes each topic is a mixture of words, and each document is a mixture of topics (Figure 3). In this project, I detected 10 topics using 10 words in each topic (for more detail see *tii.py* file in the GitHub page). The output of this section is a csv file containing words in each topic and html file for representing results (see "final outputs" zipped file in the GitHub page).

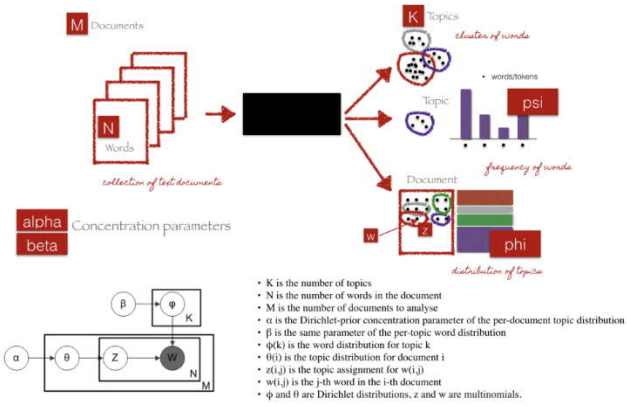


FIGURE 3. LDA TOPIC CLUSTERING PIPELINE (SOURCE: HTTP://CHDOIG.GITHUB.IO/PYTEXAS2015-TOPIC-MODELING/#/3/4)

### C. *Word cloud*

Word cloud is a visualization technique of frequent words in a text corpus. In word cloud, bigger words in font size represents more frequency of relevant word in the investigated corpus. I used wordcloud package in python to visualize frequent words in positive, negative, and neutral datasets (see *tiii.py* file in the GitHub page of the project)

### D. *Named-entity recognition*

In data mining, a named-entities can be recognized from other items in the text based on the similarity in attributes they have with other items. For examples, geographic locations, ages, addresses, phone numbers, organizations names, etc. can be mined in a text corpus for different initiatives such as marketing. In this project, I detected named-entities in each positive, negative and neutral datasets using Spacy package in python. Then, I plotted most frequent named-entities in each of these three documents separately (see *tiv.py* file in the GitHub page of the project).

### E. *Co-occuring words*

In NLP, ngram is a sequence of n words based on a given text corpus. In this project, I detected ngram of words with window size equal to 7 (shown as n in the below pseudo-code) in each positive, negative, and neutral dataset. I used `ngram` function from NLTK python package. Then, filtered those sequences in which word "Russia" was in the middle (three words on left and three words on right). Based on the filtered sequences, most frequent words histogram was plotted. I repeated this step by substituting keyword in the below pseudo-code (for example "Russia") with synonyms of "hate" and "like". To detect synonyms, I utilized `synsets` function in `wordnet` (from NLTK package) (see *tv.py* and *tvii.py* file in the GitHub page of the project).

```
keyword = "russia"
for text_corpus in [positive_set, negative_set, neutral_set]:
    list = ngram(text_corpus, n = 7)
    filtered = [elem for elem in list if elem[3] == keyword]
```

### F. *Empath categorization*

Empath is a tool that can generate and validate new lexical categories on demand from a small set of seed terms (like "bleed" and "punch" to generate the category violence). Empath draws connotations between words and phrases by deep learning a neural embedding across more than 1.8 billion words of modern fiction. Given a small set of seed words that characterize a category, (Fast et al., 2016). In this project, I used Empath python package to determine categories in each of positive, negative, and neutral datasets (see *tvi.py* file in the GitHub page of the project).

### G. *Finding tweets containing shall/must/need/wish*

In the last step of the project, I filtered tweets containing modal verbs (shall, must, and need) or "wish". Using new separately filtered tweets from positive, negative, and neutral datasets, I plotted most frequent named-entities and determined Empath categories for them (see *tviii.py* and *tx.py* file in the GitHub page of the project).

## II. RESULTS AND DISCUSSION

In this section, all outcomes after running the codes (python files) explained in the methodology section are discussed. The outputs figures and tables can be found in a zipped file (named 'final outputs') in the GitHub page of the project.

### A. *Sentiment analysis*

The context of retrieved tweets is about a war; thus, as expected, about half of the tweets have negative sentiment score (120k). Also, neutral, and positive datasets have about 40k and 80k tweets, respectively (Figure 4).

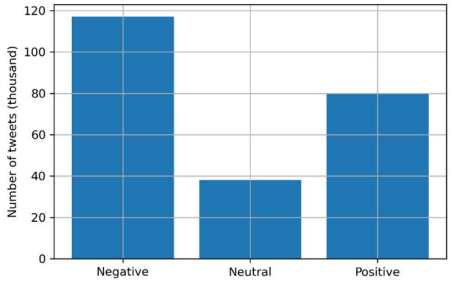


FIGURE 4. NUMBER OF TWEETS IN NEGATIVE, NEUTRAL AND POSITIVE DATASETS

### B. *LDA topics*

Topics discussed in negative datasets are shown in Figure 5 based on first and second principal components analysis (PCA1 and PCA2 in x and y axis) of words frequencies. As in shown this figure, topic 1 of negative dataset does not have any overlap with other topics. Also, topics 5 and 9 are in one side of clustering range and topics 2, 6, and 7 are on the other side. Also, topics 3, 6, 4, and 10 are in between of this range. In negative data set, first topic consists of 10 words as "zone", "exclusion", "agency", "state", "news", "radiation", "suffering", "via", "acute", and "syndrome". It seems that in this topic, people are discussing about war consequences such as suffering of people or probable radiations from Ukraine nuclear facilities. Also, exclusion (southern and eastern) parts of Ukraine in 2014 is remined in the tweets. On the other hand, in topics 2, 6, and 7, words "russia", "invasion", "already", "war", "see", "support", "military", "city", "sent", "region", "president", "former", "world", "fire", "footage", "least", "tank", "people", "many", "told", "anonymous", "piece", "hit", "recruitment", "applied", "worst" are frequent. It seems that taking action by government and calling for support is discussed in these

topics. In topics 5 and 9, words "year", "one", “old", "son", "vehicle", "every", "moreover", "house", "several", "said", "home", "fighting", "trucks", "part", "two", "munitions", "russia" are used more frequently. However, clustering of LDA method in different for positive and neutral dataset. For example, in these data sets, topics 1 and 2 are separate from all other topics (see Figure 5b and c). In the first topic of positive dataset, people are tweeting about “peace”, “standing” against “war”, and “nightingale” which is symbol of Ukraine. In the first topic of neutral dataset, words “make”, “china”, “world”, and “russia” are more frequent which can represent calling for action from world to react to the war (more detail in Appendix and html files in the “final output” zipped file in the GitHub page).

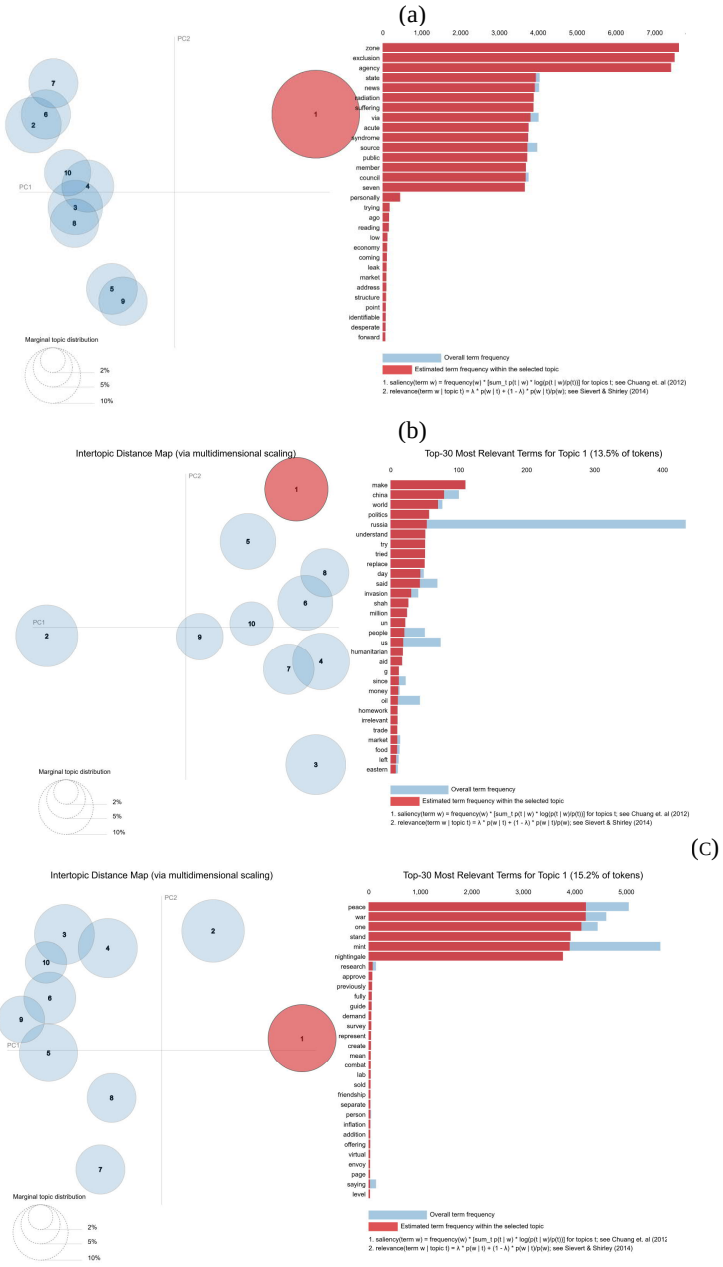


FIGURE 5. TOPICS AND WORDS IN A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATASETS

## C. *Word cloud visualization*

World cloud of negative, neutral, and positive data sets before and after preprocessing are shown in Figure 6. As shown in this figure, words “radiation” , “exclusion”, “zone”, “syndrome” is frequent in negative dataset. Also, word “russia” and “peace” are dominantly frequent in neutral and positive datasets, respectively.

The effect of preprocessing to achieve better results is obvious from Figure 6 by comparing part I and II. Some nonmeaningful words such as “co” and “https” are frequent in all datasets before preprocessing which are deleted by preprocessing step to have more informative word clouds. However, there exist similar word in both parts (e.g., radian in a-I and a-II).

## D. *Named-entity recognition*

In the negative dataset, most frequent name-entities are “seven” and “russia” which are labeled as “Cardinal” (representing cardinal numbers) and “GPE” (representing countries, cities, and states) by Spacy python package. In neutral entities, “russia”, “china” (as “GPE” entities), and “march” (as month entity) are more frequent than others. Finally, in the positive dataset, “one” (“cardinal” entity), “russia”, and “china” are more frequent named-entities (more details in Figure 7).


FIGURE 6. WORLD CLOULD OF A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA SETS AFTER PREPROCESSING (I) AND BEFORE IT (II)

## E. *Co-occuring words*

In the negative dataset, words “war”, “military”, “invasion”, and “security” are more co-occurring with ”russia” (based on ngram filtering of words in window size equal to 7). In the neutral dataset, “language”, “china”, “today”, “news”, and “war” are more co-occurring with “russia”. As expected, co-occurring words in the negative dataset with ‘russia’ have negative valence, but they are neutral in the neutral dataset (e.g., compare “invasion” with

"news"). Finally, "peace", as a positive word, is the most co-occurrent word with 'russia' in the positive dataset. This means that in positive dataset, people are hoping for peace reacting to the war news (more detail in Figure 8).

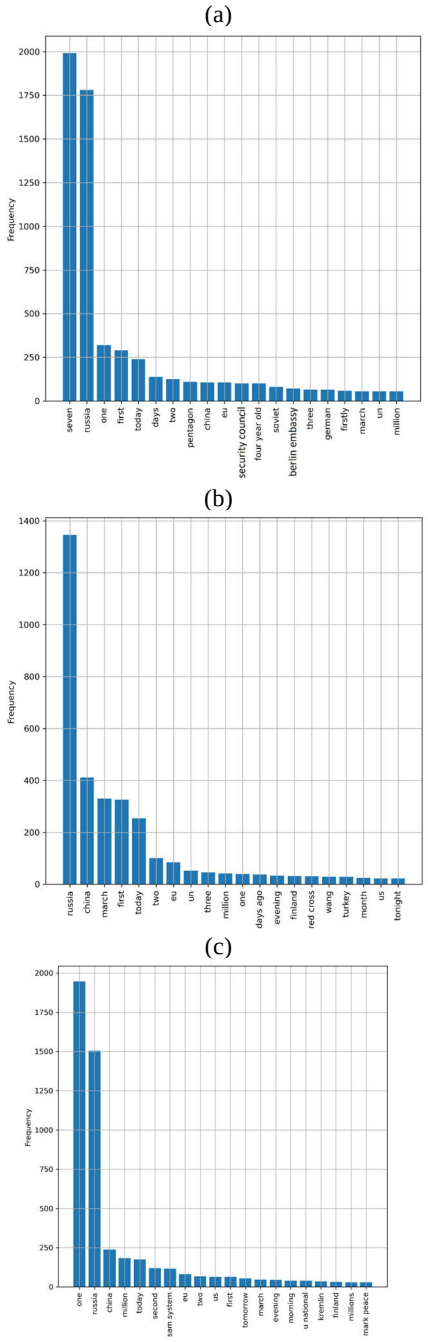


FIGURE 7. HISTOGRAM OF MOST FREQUENT NAMED-ENTITIES IN THE A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA SETS

In the negative dataset, words "shot", "world", "footage", and "drone" are more co-occurring with "like" and its synonyms. In the neutral dataset, "russia", "love", "follow", and "travel" are more co-occurring with "like" synonyms. The frequency of these words is low. It means that in the neutral dataset, word "like" and its synonyms are rare. Also, "know", and "winner" are the two mostly co-occurrent words with "like" synonyms in the positive dataset (more detail in Figure 9).

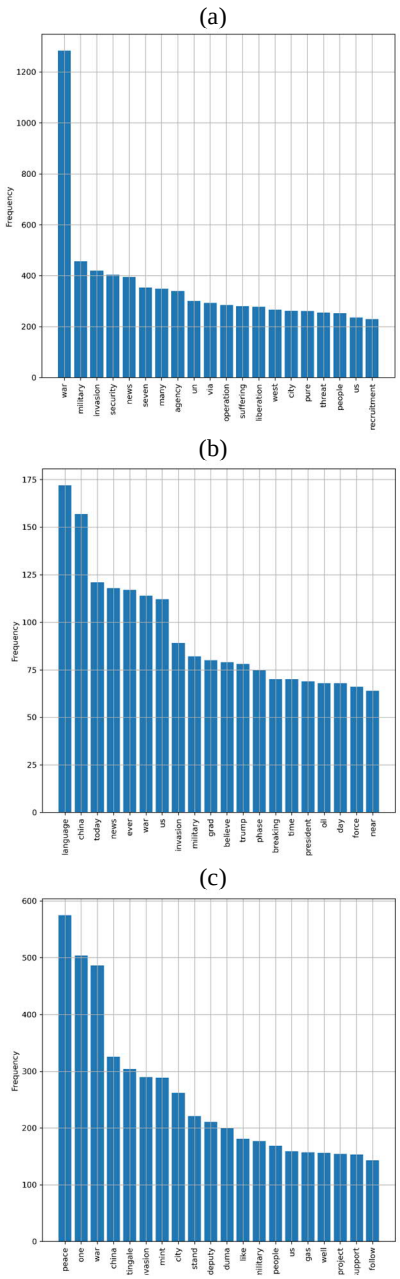


FIGURE 8. MOST FREQUENT CO-OCCURENT WORDS WITH "RUSSIA" IN THE A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA SETS

In the negative dataset, words "free", "violence", "music", and "lightheart" are more co-occurring with "hate" and its synonyms. It may mean that people are condemning violence in their tweets reacted to the war. Also, "lightheart" is a game company (in Helsinki) which has condemned the invasion of Russia strongly in its LinkedIn page. The frequency of these words is low which means that in the negative dataset, word "hate" and its synonyms are rare. Also, in the neutral dataset, there is no co-occurring word with "hate" synonyms (i.e., an empty set). Furthermore, "like", and "love" are the two mostly co-occurrent words with "hate" synonyms in the positive

dataset with a very low frequency equal to 2 time (more detail in Figure 10). The low frequency of co-occurring words inhibits any robust inference.

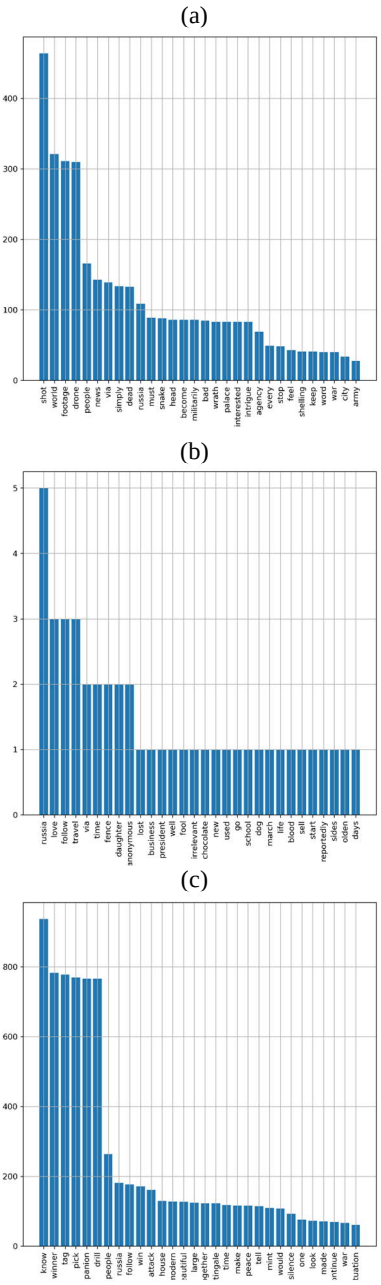


FIGURE 9. MOST FREQUENT CO-OCCURENT WORDS WITH "LIKE" SYNONYMS IN THE A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA

## F. *Empath categorization*

These are the most 15 relevant categories to the negative dataset based on Empath categorization technique: war, death, government, leader, fight, work, banking, violence, military, negative emotion, suffering, pain, dominant hierarchy, politics. This shows high potential of Empath categorization to detect relevant topics in a text corpus because context of the tweets is war, and such topics are expected to be discussed. For the neutral dataset, more relevant topics are war, government, military, travelling, musical, leaser, and vacation. As expected, the negative valence of topics in the neutral dataset is less than negative dataset. Finally, in the positive data set, Empath categorization has found these topics as mostly relevant: war, money, fight, smell, death, military, body, leader, help, and giving. Presence of giving help, and money in the topics of positive dataset show that people try to support victims by supporting them financially (see Table 4.S in Appendix).

## G. *Finding tweets containing shall/must/need/wish*

Most frequent named-entities in tweets containing modal verbs in negative dataset are "seven", "russia", "one", "first", "today", "days", "two", "pentagon", "eu", "security council", "four year old", "soviet", "berlin emabassy". More frequent entities in the negative, neutral, and positive datasets are shown in Figure 11. Excluding cardinal entities (e.g., seven), people require organizations such as United Nation Security Council, Berlin Embassy, Pentagon, etc. to react to the war. Empath categories for the same filtered tweets from negative dataset are majorly on government, war, leader, death, work, banking, fight, violence, pain, and suffering. Therefore, people are talking about victims of the war and their pain calling for action from governments and international organizations. Similar (not exactly same) Empath categories are detected for the neutral and positive datasets (more detail in the "Final output" zipped file in the GitHub page).

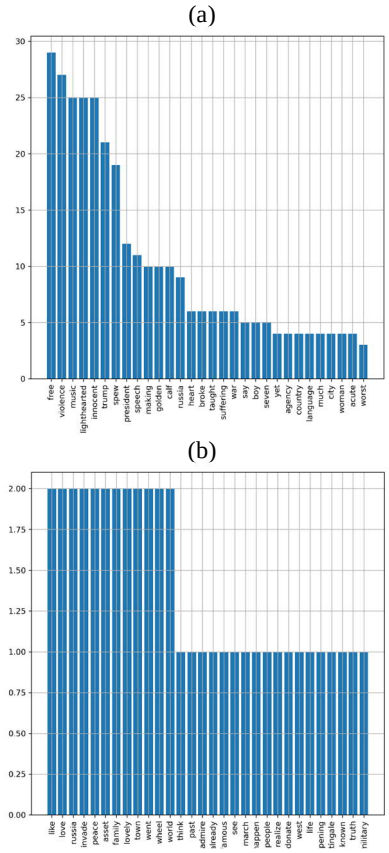


FIGURE 10. MOST FREQUENT CO-OCCURENT WORDS WITH "HATE" SYNONYMS IN THE A) NEGATIVE, AND B) POSITIVE DATA SETS

Most frequent named-entities in tweets containing “wish” in the negative, neutral, and positive datasets are “seven”, “russia", and “one” respectively (Figure 12). In the positive dataset, Empath categories are majorly war, money, fight, smell, death, military, body, giving, and party. While in the negative dataset Empath categories are government, war, leader, death, work, banking, fight, violence, suffering, and pain which have more negative valence than the positive dataset topics (more detail in the zipped file of outputs in the project GitHub page).

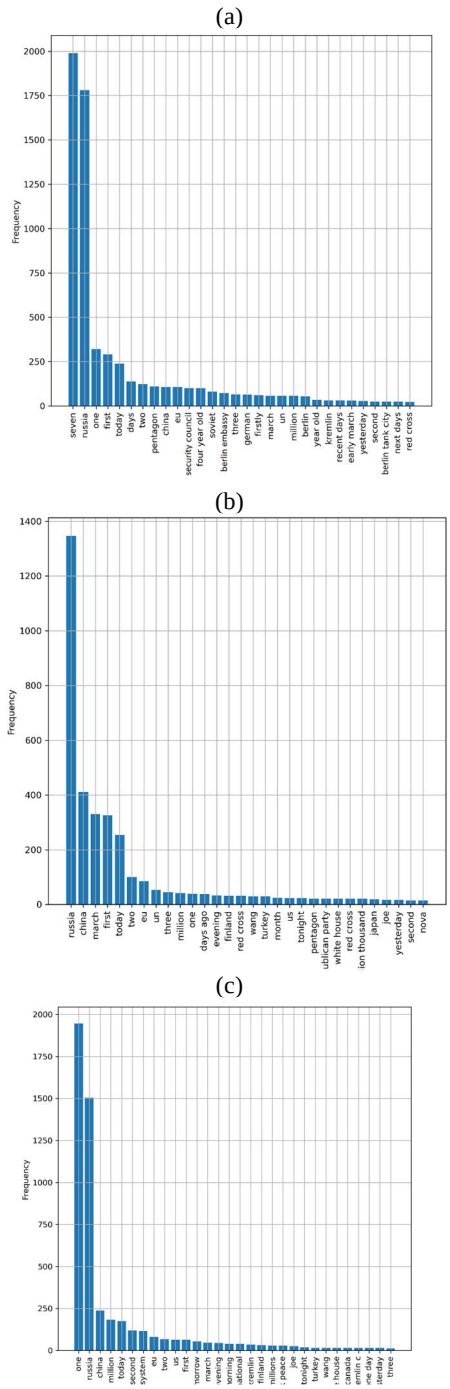


FIGURE 11. MOST FREQUENT NAMED-ENTITIES IN TWEETS CONTAINING MODAL VERBS (SHALL, MUST, AND NEED) IN THE A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA SETS

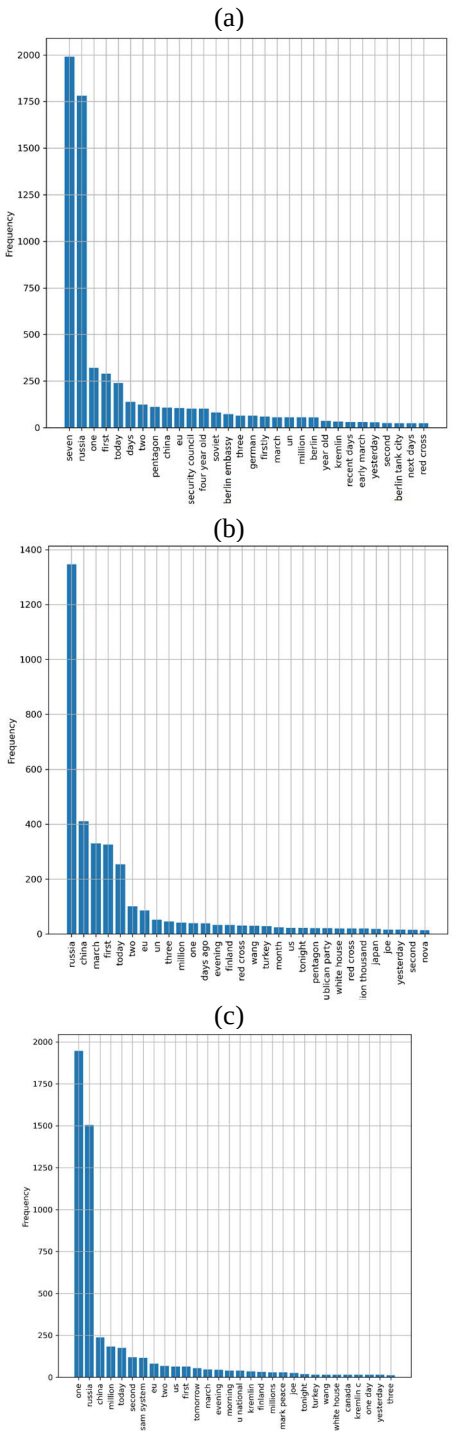


FIGURE 12. MOST FREQUENT NAMED-ENTITIES IN TWEETS CONTAINING “WISH” IN THE A) NEGATIVE, B) NEUTRAL, AND C) POSITIVE DATA SETS

In detecting the location of tweets, all of them did not have a meaningful tag. For example, some of tweet’s location were tagged as “somewhere”. Also, in some countries (e.g., Iran), twitter website is filtered. Therefore, the map and statistics given in Figure 1 faces some limitations and it does not cover all tweets processed in this project. In all three types of investigated datasets, there existed very frequent named-entities such as “seven” or “one”. Interpreting them need having access to the context and background of the date in which tweets are posted. For

example, if tweets are posted after seven days after Russia invasion, it makes sense that “seven” word is frequent. Investigated tweets in this project are for March 2022. I did not investigate background information of this month, so I had challenge to interpret them. Also, there existed limitations inside LDA and Empath categorization techniques. For example, there are overlapping topics distinguished by LDA, so interpreting them was hard. Detected categories by Empath for datasets containing modal verbs or “wish” was similar to what I had detected for whole of positive, negative and neutral data set. This make sense because tweets containing modal verbs was a subset of all tweets. I was not able to retrieve any specific information from categories using filtered tweets having modal verbs or “wish” more than what I interpreted using whole dataset.

## III. Overall discussion

Detected topics by LDA and Empath have similarities. For example, in the negative dataset, Empath python package suggest categories such as war, death, government, leader, and fight. Also, topic 1 clustered by LDA has frequent word such as zone, exclusion, agency, radiation, suffering, and acute for the same dataset. Also, in the negative dataset, topics 2, 6, 7 are clustered by LDA because of war related words such as “russia”, and “war”.

I recommend pursuing similar investigation on each positive, negative, and neutral dataset considering the location of tweets. Also, pursuing changes of overall sentiment score in time can be informative. I retrieved tweets in March 2022. There exist other tweets from March to November 2022 in the Kaggle dataset. Benefiting from the timeseries of tweets will help to take the context of time and location into account when analyzing results.

## I. Conclusion

In this study, I investigated sentiment score of about 240k English tweets posted in March 2022 about Russia invasion to Ukraine. I separated tweets into negative, neutral, and positive datasets based on sentiment score of them. Then, topics discussed in each dataset were studied. Also, most frequent words and name-entities of these three datasets were investigated separately. The results of this project showed that Natural Language Processing techniques such as Latent Dirichlet Allocation (LDA) and Empath categorization can be used to detect main topics in a big text corpus such as what was investigated in this research. However, investigating the frequency of named-entities cannot lead to informative outcome if temporal and spatial context of tweets are not considered in analysis.

## References


Card, S. K., Mackinlay, J. D., & Shneiderman, B. (2009). Text Mining. *Encyclopedia of Database Systems*, *3*, 3061–3065. https://doi.org/10.1007/978-0-387-39940-9_418

Chen, E., & Ferrara, E. (2022). *Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia*. https://doi.org/10.48550/arxiv.2203.07488

Fast, E., Chen, B., & Bernstein, M. (2016). Empath: Understanding Topic Signals in Large-Scale Text. *Conference on Human Factors in Computing Systems - Proceedings*, 4647–4657. https://doi.org/10.1145/2858036.2858535

Korzycki, M., Gatkowska, I., & Lubaszewski, W. (2017). Can the Human Association Norm Evaluate Machine-Made Association Lists? *Cognitive Approach to Natural Language Processing*, 21–40. https://doi.org/10.1016/B978-1-78548-253-3.50002-0

Makhortykh, M., & Lyebyedyev, Y. (2015). #SaveDonbassPeople: Twitter, Propaganda, and Conflict in Eastern Ukraine. *Http://Dx.Doi.Org/10.1080/10714421.2015.1085776*, *18*(4), 239–270. https://doi.org/10.1080/10714421.2015.1085776

Nikolayenko, O. (2019). Framing and counter-framing a Peace March in Russia: the use of Twitter during a hybrid war. *Https://Doi.Org/10.1080/14742837.2019.1599852*, *18*(5), 602–621. https://doi.org/10.1080/14742837.2019.1599852

Polyzos, E. (2022). Escalating Tension and the War in Ukraine: Evidence Using Impulse Response Functions on Economic Indicators and Twitter Sentiment. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4058364

## APPENDIX

TABLE 1.S. LDA OF THE NEGATIVE DATASET

| Topics | Words |
|---|---|
| 1 | 0.105*"zone" + 0.104*"exclusion" + 0.103*"agency" + 0.054*"state" + 0.054*"news" + 0.054*"radiation" + 0.054*"suffering" + 0.053*"via" + 0.052*"acute" + 0.052*"syndrome" |
| 2 | 0.032*"russia" + 0.028*"invasion" + 0.028*"already" + 0.023*"war" + 0.023*"see" + 0.022*"support" + 0.022*"military" + 0.021*"army" + 0.020*"go" + 0.019*"first" |
| 3 | 0.027*"army" + 0.024*"fighting" + 0.019*"think" + 0.019*"russia" + 0.017*"would" + 0.017*"people" + 0.016*"even" + 0.014*"used" + 0.013*"vehicle" + 0.013*"begun" |
| 4 | 0.066*"russia" + 0.036*"war" + 0.028*"use" + 0.025*"returned" + 0.022*"security" + 0.022*"un" + 0.019*"people" + 0.016*"west" + 0.015*"million" + 0.014*"offensive" |
| 5 | 0.049*"year" + 0.040*"one" + 0.039*"old" + 0.032*"son" + 0.027*"vehicle" + 0.026*"every" + 0.026*"moreover" + 0.024*"house" + 0.020*"several" + 0.018*"said" |
| 6 | 0.059*"city" + 0.033*"sent" + 0.032*"region" + 0.031*"president" + 0.023*"former" + 0.022*"world" + 0.019*"fire" + 0.019*"footage" + 0.019*"least" + 0.018*"tank" |
| 7 | 0.023*"people" + 0.023*"many" + 0.022*"told" + 0.021*"russia" + 0.021*"anonymous" + 0.019*"piece" + 0.019*"hit" + 0.018*"recruitment" + 0.016*"applied" + 0.015*"worst" |
| 8 | 0.087*"russia" + 0.030*"war" + 0.021*"us" + 0.020*"food" + 0.013*"china" + 0.013*"high" + 0.011*"new" + 0.011*"evidence" + 0.010*"world" + 0.010*"tell" |
| 9 | 0.028*"home" + 0.021*"fighting" + 0.015*"trucks" + 0.015*"part" + 0.014*"two" + 0.014*"munitions" + 0.014*"russia" + 0.013*"bread" + 0.013*"military" + 0.012*"serious" |
| 10 | 0.074*"war" + 0.039*"russia" + 0.024*"minister" + 0.024*"got" + 0.018*"another" + 0.018*"human" + 0.018*"cat" + 0.017*"woman" + 0.016*"days" + 0.016*"boy" |

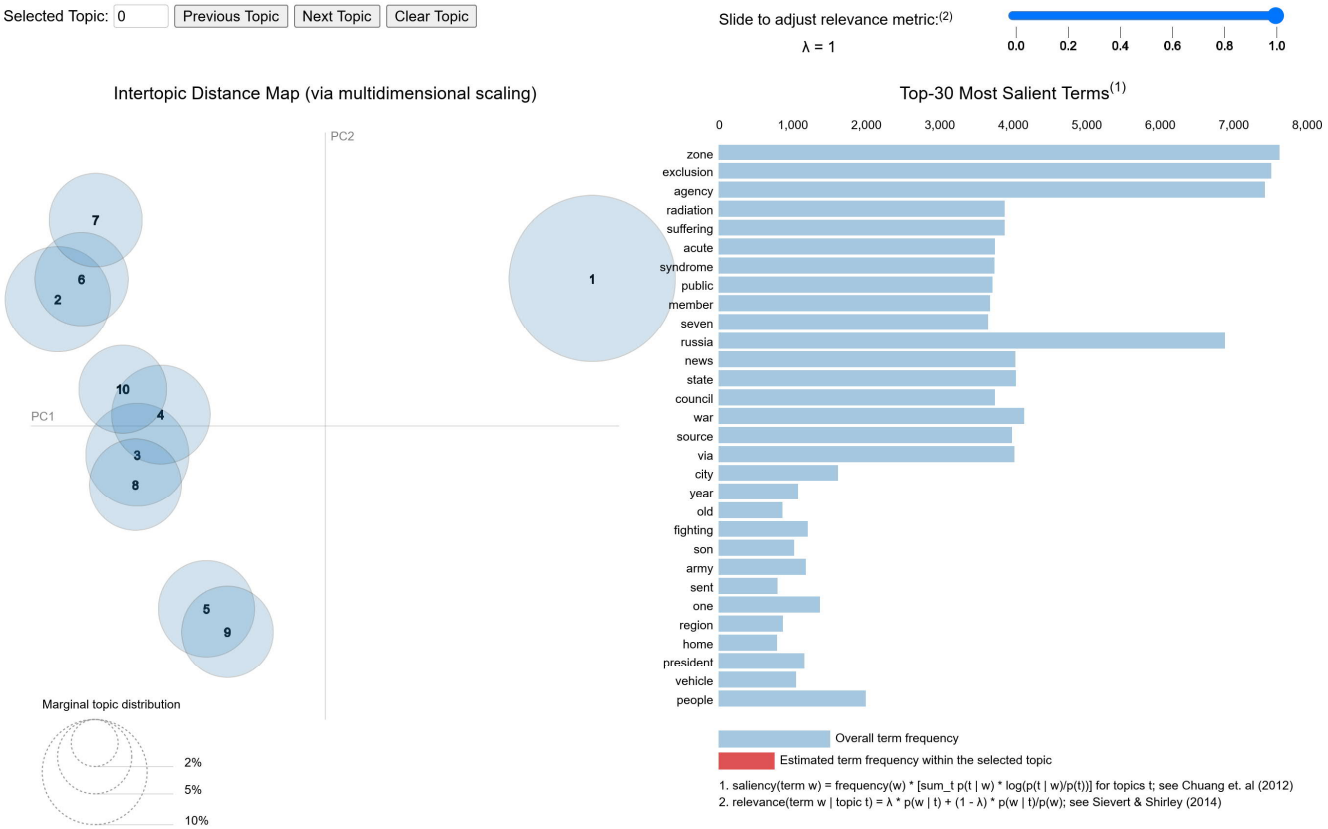


FIGURE S 1. VISUALIZATION OF LDA TOPICS IN NEGATIVE DATASET
(HTML AVAILABLE IN THE GITHUB PAGE: HTTPS://GITHUB.COM/AIEFARAMARZZADEH/NATURAL-LANGUAGE-PROCESSING-PROJECT/BLOB/MAIN/1.FINAL%20OUTPUTS.7Z)

TABLE 2.S. LDA OF THE NEUTRAL DATASET

| Topics | Words |
|---|---|
| 1 | 0.044*"russia" + 0.027*"people" + 0.026*"via" + 0.024*"trump" + 0.020*"city" + 0.019*"near" + 0.017*"military" + 0.016*"back" + 0.015*"breaking" + 0.014*"country" |
| 2 | 0.083*"russia" + 0.037*"today" + 0.025*"region" + 0.019*"days" + 0.017*"two" + 0.016*"war" + 0.014*"east" + 0.014*"little" + 0.013*"get" + 0.011*"many" |
| 3 | 0.040*"russia" + 0.036*"foreign" + 0.034*"minister" + 0.025*"president" + 0.019*"china" + 0.019*"even" + 0.018*"part" + 0.018*"ago" + 0.017*"per" + 0.017*"us" |
| 4 | 0.048*"russia" + 0.047*"anonymous" + 0.041*"situation" + 0.038*"breaking" + 0.034*"us" + 0.033*"map" + 0.025*"news" + 0.023*"approximate" + 0.018*"house" + 0.013*"anti" |
| 5 | 0.049*"russia" + 0.032*"video" + 0.022*"another" + 0.020*"hit" + 0.020*"time" + 0.018*"new" + 0.016*"gas" + 0.014*"right" + 0.014*"follow" + 0.014*"mortar" |
| 6 | 0.047*"army" + 0.041*"cross" + 0.040*"red" + 0.039*"building" + 0.035*"russia" + 0.035*"bombed" + 0.026*"south" + 0.026*"marked" + 0.016*"bank" + 0.015*"go" |
| 7 | 0.049*"unknown" + 0.046*"russia" + 0.030*"target" + 0.027*"use" + 0.027*"sam" + 0.027*"footage" + 0.026*"drone" + 0.026*"system" + 0.025*"location" + 0.024*"previously" |
| 8 | 0.077*"border" + 0.063*"complete" + 0.062*"ambassador" + 0.060*"closure" + 0.021*"lake" + 0.019*"real" + 0.017*"national" + 0.016*"russia" + 0.013*"guard" + 0.013*"photography" |
| 9 | 0.074*"make" + 0.053*"china" + 0.047*"world" + 0.038*"politics" + 0.036*"russia" + 0.034*"understand" + 0.034*"try" + 0.034*"tried" + 0.034*"replace" + 0.029*"day" |
| 10 | 0.072*"bam" + 0.059*"live" + 0.057*"new" + 0.051*"time" + 0.049*"first" + 0.047*"march" + 0.041*"way" + 0.040*"international" + 0.040*"spotted" + 0.038*"together" |

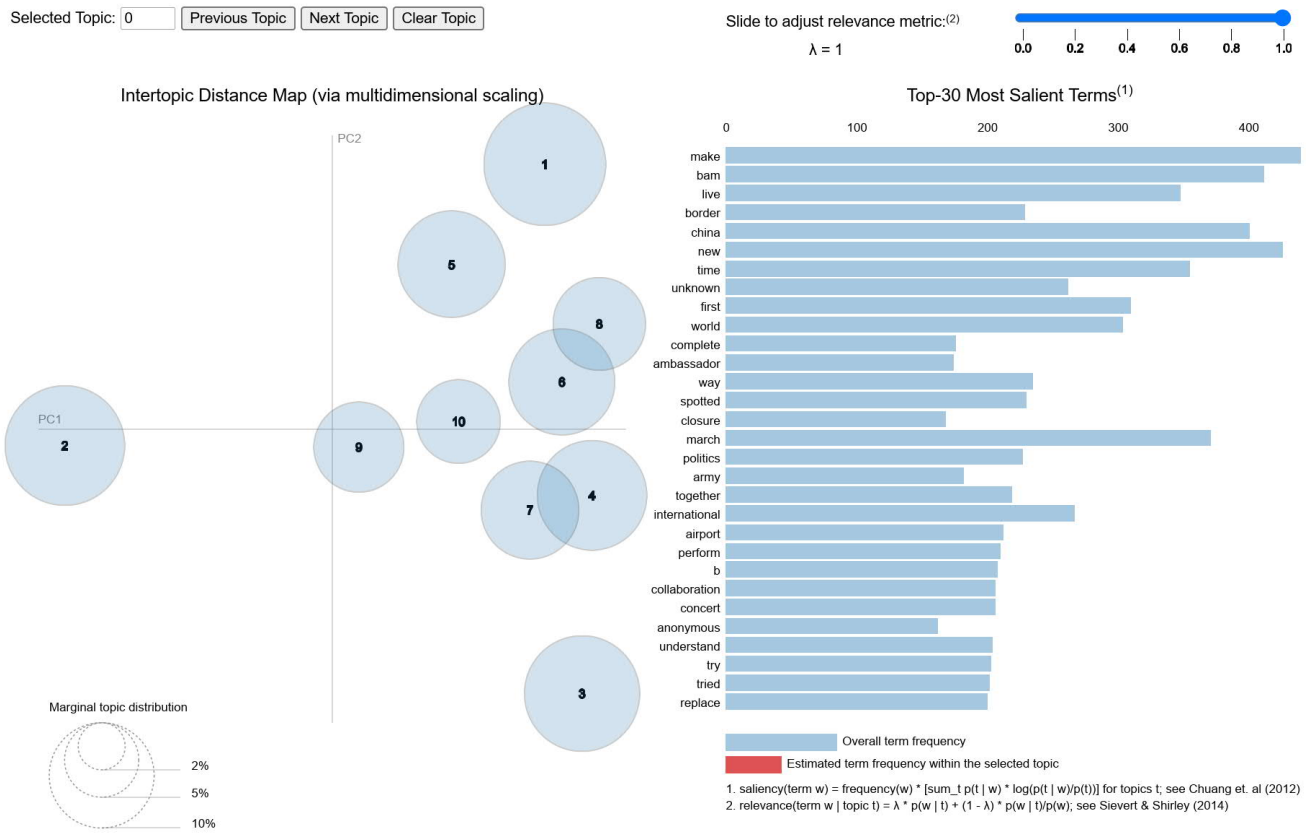


FIGURE S 2. VISUALIZATION OF LDA TOPICS IN NEUTRAL DATASET
(HTML AVAILABLE IN THE GITHUB PAGE: HTTPS://GITHUB.COM/AIEFARAMARZZADEH/NATURAL-LANGUAGE-PROCESSING-PROJECT/BLOB/MAIN/1.FINAL%20OUTPUTS.7Z)

TABLE 3.S. LDA OF THE POSITIVE DATASET

| Topics | Words |
|---|---|
| 1 | 0.039*"series" + 0.023*"used" + 0.022*"system" + 0.020*"interesting" + 0.020*"side" + 0.019*"battalion" + 0.017*"sam" + 0.017*"clearance" + 0.017*"russia" + 0.015*"nuclear" |
| 2 | 0.160*"peace" + 0.160*"war" + 0.156*"one" + 0.148*"stand" + 0.148*"mint" + 0.143*"nightingale" + 0.003*"research" + 0.003*"approve" + 0.003*"previously" + 0.002*"fully" |
| 3 | 0.029*"thank" + 0.023*"president" + 0.022*"military" + 0.021*"peace" + 0.018*"us" + 0.017*"million" + 0.017*"russia" + 0.016*"case" + 0.016*"list" + 0.015*"new" |
| 4 | 0.031*"k" + 0.028*"buy" + 0.025*"help" + 0.021*"get" + 0.020*"need" + 0.017*"one" + 0.017*"spirit" + 0.016*"new" + 0.016*"unbreakable" + 0.015*"possible" |
| 5 | 0.079*"mint" + 0.057*"like" + 0.049*"tag" + 0.047*"know" + 0.037*"day" + 0.034*"tomorrow" + 0.033*"let" + 0.033*"giveaway" + 0.030*"excited" + 0.030*"winner" |
| 6 | 0.041*"russia" + 0.031*"situation" + 0.029*"march" + 0.025*"latest" + 0.024*"defence" + 0.023*"update" + 0.022*"intelligence" + 0.022*"response" + 0.018*"media" + 0.015*"gas" |
| 7 | 0.048*"tell" + 0.041*"army" + 0.029*"neo" + 0.027*"truth" + 0.024*"russia" + 0.024*"want" + 0.019*"never" + 0.019*"show" + 0.017*"back" + 0.017*"world" |
| 8 | 0.084*"russia" + 0.029*"good" + 0.023*"morning" + 0.022*"project" + 0.021*"release" + 0.020*"peace" + 0.018*"form" + 0.018*"celebrate" + 0.018*"mon" + 0.018*"filling" |
| 9 | 0.048*"russia" + 0.031*"well" + 0.026*"eu" + 0.018*"army" + 0.018*"future" + 0.018*"military" + 0.018*"security" + 0.016*"drop" + 0.015*"promise" + 0.015*"belt" |
| 10 | 0.041*"please" + 0.030*"people" + 0.029*"support" + 0.028*"find" + 0.025*"world" + 0.022*"government" + 0.021*"think" + 0.020*"place" + 0.018*"thanks" + 0.018*"supporting" |

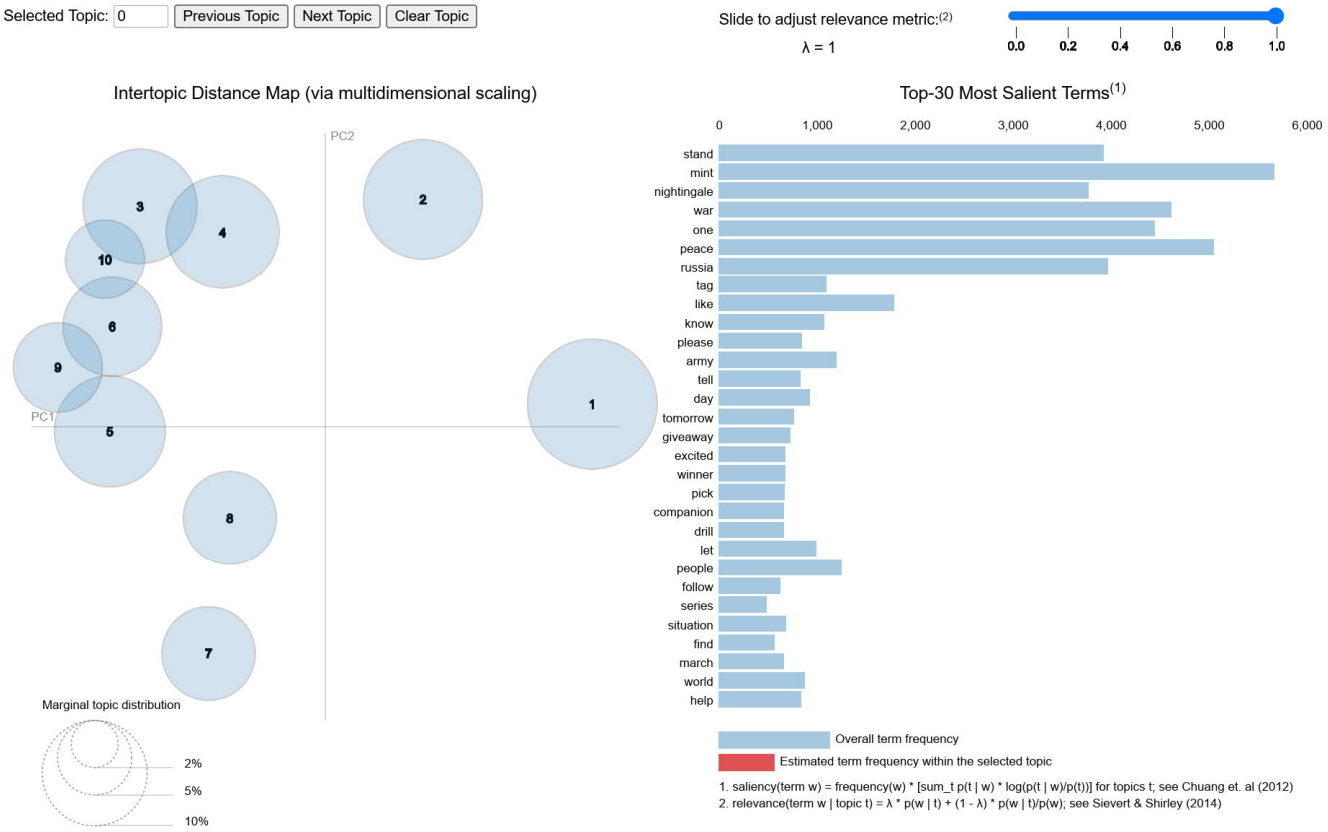


FIGURE S 3. VISUALIZATION OF LDA TOPICS IN THE POSITIVE DATASET
(HTML AVAILABLE IN THE GITHUB PAGE: HTTPS://GITHUB.COM/AIEFARAMARZZADEH/NATURAL-LANGUAGE-PROCESSING-PROJECT/BLOB/MAIN/1.FINAL%20OUTPUTS.7Z)

TABLE 4.S. FIRST FIFTHEEN EMPATH CATEGORIES IN EACH OF INVESTIGATED DATASETS

| # | NEGATIVE | NEUTRAL | POSITIVE |
|---|---|---|---|
| **1** | war | war | war |
| **2** | death | government | money |
| **3** | government | military | fight |
| **4** | leader | traveling | smell |
| **5** | fight | musical | death |
| **6** | work | leader | military |
| **7** | banking | music | body |
| **8** | violence | air travel | leader |
| **9** | military | vacation | giving |
| **10** | negative emotion | dominant hierarchical | communication |
| **11** | suffering | law | government |
| **12** | pain | politics | party |
| **13** | dominant hierarchical | speaking | help |
| **14** | politics | meeting | celebration |
| **15** | law | communication | business |