Emotions of War in Ukraine: Analyzing Public Sentiment from Twitter Data

Lev Paciorkowski and Khevna Parikh

DSGA 1015: Text as Data

Professor Andrew Halterman

May 11, 2022

ABSTRACT

Our core hypothesis is that aggregated Twitter data can be a reasonable proxy for overall public sentiment – in other words, that the general "mood" of a regional population can be accurately measured by analyzing specific patterns in the language coming from that population's tweets. Our research question is then what changes in mood can we detect from the Ukraine-Russia region by analyzing tweets before and after Russia's invasion on February 24, 2022? Indeed, by analyzing randomly sampled collections of tweets from this geographic area, we show that there is a drastic and clearly quantifiable change in tweet patterns immediately following 2/24.

We also note a phenomenon of public sentiment trends observed from our Twitter data that we call "persistence". Before 2/24, we can quantify numerous baselines in various tweet patterns from the Ukrainian and Russian population. After the "shock" of the invasion, these patterns deviate significantly from baseline, but then regress back towards their prewar levels. Some patterns regress slower or faster than others, while other patterns are not regressing at all, and remain persistently at their new post-war levels. We put forth some possible explanations for these pattern-specific differences and attempt to interpret their meaning.

Finally, we develop a formalized topic model based on Latent Dirichlet Allocation as a more robust way of discovering the most prevalent topics prewar compared to postwar.

Keywords: Ukraine, Russia, Twitter, Invasion

1. INTRODUCTION

The specific study of using tweet text for sentiment analysis is relatively recent, with one of the first publications being Go et al. (2009), in which the authors use machine learning techniques to classify tweets as either ”positive” or ”negative”. Soon afterwards, the field quickly grew in popularity; Kouloumpis et al. (2011) is today considered a landmark study where the authors add ”hashtag” text as one of their specific features as well as open a still-ongoing discussion about the unique challenges of using tweet text for sentiment analysis.

Indeed, as is well-documented in Zimbra et al. (2018), even state-of-the-art models have historically achieved relatively low accuracy in tweet sentiment classification. The main difficulty is in the vast range of topics that people tweet about, as well as the more informal language that is often used in the limited 280-character space.

This project ultimately aims to examine the emotions and thoughts of individuals from Ukraine and Russia as a result of the ongoing Russian-Ukrainian conflict. Rather than classifying individual tweets as positive or negative, we focus on the broad trends and patterns in word choice that are detectable from a large random sample of tweets, and what changes occur as a result of Russia's invasion. This requires the use of topic modeling, an important text analysis technique for discovering what natural groupings of words may exist in a corpus. In Section 3, we preface our formal topic model with an analysis of hand-picked topics that we think are most relevant. In Section 4, we develop a Latent Dirichlet Allocation topic model to more rigorously discover which topics are most relevant.

2. DATA COLLECTION

Twitter is a widely-known social media platform that allows users to blog in short messages, consisting of 280 characters or less. People often communicate their emotions, positive or negative, on this distinguished social networking site. For this reason, we collected text data using Twitter API. Using the computing programming language, Python, and one of its packages, Twarc, we were able to archive two sets of Twitter data from the Russia-Ukraine region. One set accounted for tweets a few months prior to Russia's invasion into Ukraine, serving as our baseline collection and another set constituting tweets after Russia's invasion to the present day. To maintain balance between both time periods, we only gathered tweets three months prior to Russia's invasion of Ukraine on February 24, 2022.

To generate the random sample of data using Twitter's API, we set our query parameters to the most commonly used stop words in the Ukrainian and Russian language as shown in **Figure 1**. In addition to collecting the text of each tweet, we collected the timestamp as well to further explore the change in sentiments over time.

```
stopwords = {
     'uk': 'але OR нею OR як OR в OR мають OR від OR сказати OR або OR він OR що OR це OR для OR вони OR іншого OR до OR з \
     OR твій OR на OR там OR був OR словом OR ви OR і OR їх OR про OR буде OR кожний OR не OR можна OR ми OR використати OR \
     який OR мав OR один OR всі OR вона OR я OR були OR робити OR коли OR бути OR є OR те OR якщо OR за',

    'ru': 'использовать OR когда OR в OR от OR как OR были OR был OR кто-то OR для OR это OR до OR их OR если OR там OR на \
    OR с OR его OR и OR делает OR или OR они OR не OR сказал OR что OR он OR все OR вы OR то OR она OR по OR слову OR такие \
    OR я OR есть OR быть OR сказать OR но OR каждый OR мы OR можете'
        }
```

Figure 1. List of Ukrainian and Russian stop words passed in our query to Twitter's API. The query would return any tweet in our specified geographic region that contains at least one of these words in either language. We assume this ultimately results in an approximately random selection of tweets.

Initially, we focused on gathering data from some of the largest cities in both Ukraine and Russia. We settled on the following cities based on highest populations - Moscow, St. Petersburg, Novosibirsk, Kyiv, Kharkiv, and Odessa. Consequently, we obtained a total of 12 data sets using the above query parameters, one for each city region and time period (pre-invasion/post-invasion). Remarkably, we were able to gather a total of approximately 2 million individual tweets.

Much to our dismay, our extraction consisted of an excessive number of duplicates. According to Twitter Developer Platform (Twitter, 2022), individual tweets can contain either a precise location, only if self-reported or a general area (e.g., a local coffee shop or city) generated by Foursquare. Even then, only 1 to 2 percent of tweets are tagged with Geo location. That is precisely why our model could not pinpoint an exact location and picked up the same tweet multiple times as it was searching across the entire Russia-Ukraine province. To combat this, we simply dropped duplicated tweets.

In the end, our data set consists of a total of 443,718 randomly sampled tweets from Ukraine and Russia, distributed evenly between November 24, 2021 and May 3, 2022 (an average of 2,773 tweets per day with no fewer than 2,530 and no greater than 2,909 tweets on any individual day). The text was cleaned and preprocessed with procedures for converting Cyrillic characters to lowercase, eliminating punctuation and stop words as well as word stemming.

3. MEASURING PERSISTENCE OF SPECIFIC PATTERN CHANGES

One of the first fundamental pattern changes that is readily noticeable is the increase in the proportion of tweets that are in Ukrainian after 2/24. According to the most recent estimates (Kemp, 2022a,b), there are roughly 2.95 million and 910,000 Twitter users in Russia and Ukraine, respectively. As a result, we would naively expect approximately one-fourth of tweets randomly collected from the two countries to be in Ukrainian. The true expected proportion is somewhat more complicated, as a certain percentage of Ukrainians speak Russian as their first language.

In our data set, we are able to label the language of each tweet based on the characters that appear in its text. There are four characters exclusive to Ukrainian that do not appear in Russian, as well as four characters exclusive to Russian that do not appear in Ukrainian. Twenty nine characters are then shared by both languages. Using this information, we can track the prevalence of each language in our data set over time, and show the results in **Figure 2**.

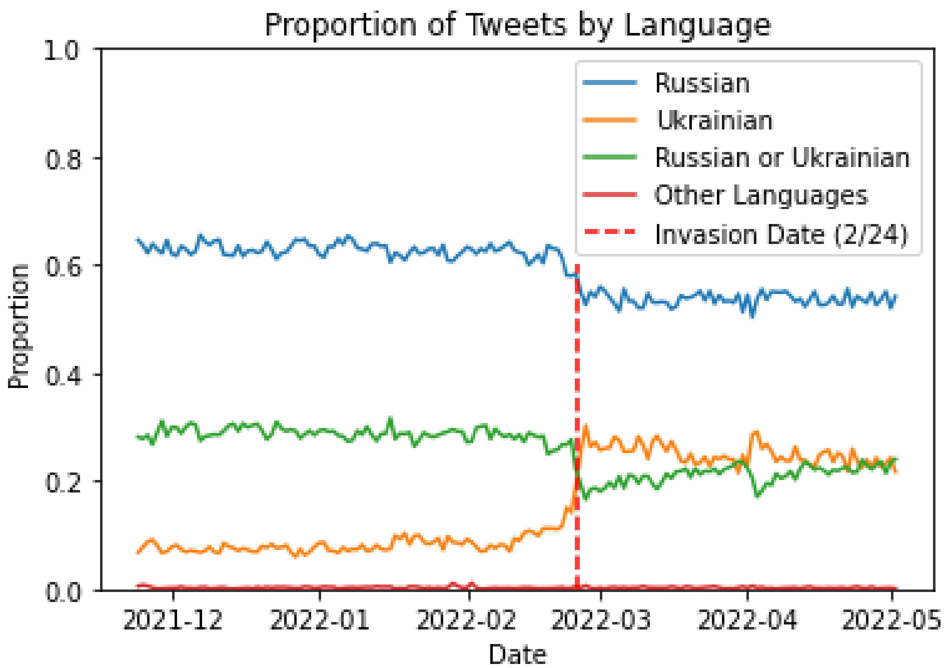


Figure 2. Daily proportion of tweets by language. The group "Russian or Ukrainian" are tweets that can not be distinguished between the two languages purely on the basis of character analysis.

If we consider the baseline Ukrainian proportion to be 7.8% - the average from the start of data collection through 01/31/2022 - then this statistic nearly quadruples after the invasion date, peaking at 30.2% on February 26. Considering the latest estimate to be the mean of the final seven days of data collection - 23.3% - we can say that the statistic has "recovered" roughly 30% of the way back towards baseline from 02/26 - 05/02. This amounts to a roughly 14% recovery per month, which we define to be this statistic's "recovery coefficient". In this way we can quantify the persistence of any measurable trend in our tweet data.

Using the following procedure, we measure the persistence of a measurable quantity ("statistic") observed in our data set:

1. Define *baseline* of statistic as its average from 11/24/2021 - 01/31/2022.
2. Define *extreme* of statistic as its highest (or lowest) value observed within the seven days after 2/24, and t_0 as the date of this observation.
3. Define *shock* as $extreme - baseline$.
4. Define *current* as the average of statistic in the last seven days of data collection (04/26/2022-05/02/2022), and t_1 as 05/02/2022.
5. Calculate recovery coefficient, i.e. the recovery towards baseline per month:

$$R = \frac{extreme - current}{extreme - baseline} * \frac{30}{t_1 - t_0}$$

Based on this definition, a smaller value of R indicates a more persistent trend, while a larger value of R indicates a more fleeting trend. In extreme cases, a negative value of R indicates a trend which has continued and amplified beyond the initial change after 2/24. We apply this procedure to measure the persistence of the following statistics, namely, the proportion of tweets mentioning declensions of each group of words (i.e. topics) in the column "Keyword(s)" in Tables 1-3.

Just from these hand-picked topics, we can see many noticeable shifts before and after the invasion, as well as several interesting differences between the reactions seen in Russian vs. Ukrainian tweets. There is a great deal that can be discussed here; for the sake of brevity we highlight only the most remarkable findings.

Expectedly, one of the greatest shocks is in the topic described solely by the word "war". However, the shock is drastically greater in Russian, which sees a more than ten-fold spike in the proportion of tweets mentioning the topic, compared to

Keyword(s)	Baseline	Extreme	Shock	Current	R
"war"	0.679%	8.03%	+1080%	3.13%	0.299
"tank"	0.625%	2.19%	+250%	0.973%	0.376
"happy", "cheerful" or "laugh"	4.53%	3.48%	-23.2%	4.87%	0.611
"help"	0.831%	2.34%	+181%	1.70%	0.192
"Ukraine" or "Ukrainian"	2.79%	18.8%	+572%	10.8%	0.235
"Russia" or "Russian"	4.46%	19.3%	+333%	10.5%	0.288
"fear", "afraid", "scary" or "worry"	1.19%	2.51%	+112%	0.978%	0.518
"Zelenskyy"	0.611%	2.18%	+257%	0.828%	0.392
"Putin"	1.42%	6.68%	+369%	2.87%	0.325
"Mariupol"	0.020%	0.899%	+4310%	1.82%	-0.506
"Donbass"	0.230%	1.51%	+554%	0.527%	0.343
"child"	0.835%	2.31%	+176%	1.49%	0.268
"love"	3.43%	0.888%	-74.1%	2.11%	0.216
"bravery", "courage" or "hero"	0.207%	0.817%	+296%	0.509%	0.229
"homeland", "fatherland" or "country"	1.28%	3.20%	+150%	1.82%	0.323
"Kyiv"	0.370%	7.08%	+1810%	1.56%	0.379
"Kharkiv"	0.102%	4.22%	+4040%	0.908%	0.389

Table 1. Persistence of shocks, Russian and Ukrainian language combined. Baseline: mean daily proportion of tweets that mention any word in Keyword(s), 11/24/2021-01/31/2022. Extreme: most extreme daily proportion in 02/24-03/03/2022. Shock: percentage increase baseline to extreme. Current: mean daily proportion from 04/26-05/02/2022. R: recovery coefficient.

Ukrainian's more modest +187% spike (Ukrainians also have a greater baseline for "war", which is interesting). Russians seem to be quicker to stop tweeting about war though - the topic's recovery coefficient for Russian is nearly twice as large as it is for Ukrainian. We can confirm this visually by inspecting the graph for prevalence of tweets mentioning "war" (**Figure 3**).

Both Russian and Ukrainian saw a large spike in the usage of "war" immediately following the invasion, but Russian has regressed faster towards its baseline compared to Ukrainian. In fact, Ukrainian mentions of "war" continued to increase well past the initial 2/24 shock, ultimately peaking at 8.03% on March 18.

Another notable observation is that prior to the invasion, Ukrainians tweeted about Zelenskyy (5.03%) and Ukraine (16.5%) much more frequently than Russians tweeted about Putin (1.66%) and Russia (5.19%). Both languages saw a spike in tweets mentioning "Ukraine" or "Russia" (with a greater relative spike in each language for its counterpart), but curiously, while Russians now tweet more about Putin than they used to, Ukrainians now tweet sharply *less* about Zelenskyy than they used to - in fact, this is one of the most persistent (and perhaps surprising) trends, with a recovery factor of just 0.106. Post-invasion, Ukrainians also briefly tweeted much more frequently about Putin, but that trend has since quickly regressed (with a factor of 0.456) back to its prewar baseline.

We believe this Ukrainian "Zelenskyy effect" could be demonstrative of Ukrainians' fierce nationalistic pride and their emphasized loyalty to their whole country over loyalty solely to their leader. While prewar baselines make it clear that Ukrainians tweet more frequently about their own country and leader than Russians do (by a factor of 3), the postwar shifts in Ukrainian tweets have drawn more attention to Ukraine itself (as opposed to Zelenskyy) as well as to major cities such as Kyiv, Kharkiv and Mariupol that have come under attack by the Russians.

In particular, the most persistent shock has been to mentions of Mariupol, which in Ukrainian tweets saw a sharp spike to over 8% prevalence on March 9 when the Russian Army bombed a maternity hospital in the city. In contrast, Russian mentions of Mariupol saw a much more muted increase on March 9. Nevertheless, Mariupol is one topic that has seen persistently elevated mentions in both languages as a result of the war.

Related to national pride, while both Russian and Ukrainian tweets saw a spike in mentions of "homeland", "fatherland" and "country", Ukrainian saw a greater spike in "bravery", "courage" and "hero" than Russian. Both languages have seen a persistent increase in mentions of "help", as well as a persistent decrease in mentions of "love". Mentions of "happy",

Keyword(s)	Baseline	Extreme	Shock	Current	R
”war”	0.749%	10.3%	+1280%	3.68%	0.311
”tank”	0.674%	2.03%	+200%	1.01%	0.362
”happy”, ”cheerful” or ”laugh”	4.94%	3.68%	-25.5%	5.23%	0.575
”help”	0.959%	2.58%	+169%	1.84%	0.222
”Ukraine” or ”Ukrainian”	2.03%	19.1%	+840%	11.0%	0.222
”Russia” or ”Russian”	5.19%	21.9%	+322%	12.3%	0.272
”fear”, ”afraid”, ”scary” or ”worry”	1.46%	3.57%	+145%	1.23%	0.497
”Zelenskyy”	0.269%	2.20%	+718%	0.693%	0.355
”Putin”	1.66%	7.94%	+377%	3.86%	0.291
”Mariupol”	0.011%	0.567%	+5140%	1.45%	-0.719
”Donbass”	0.227%	1.92%	+744%	0.465%	0.385
”child”	1.02%	2.37%	+134%	1.66%	0.253
”love”	4.09%	0.662%	-83.8%	2.52%	0.242
”bravery”, ”courage” or ”hero”	0.201%	0.920%	+357%	0.337%	0.380
”homeland”, ”fatherland” or ”country”	1.69%	4.32%	+155%	2.79%	0.272
”Kyiv”	0.241%	5.49%	+2180%	1.19%	0.378
”Kharkiv”	0.051%	3.56%	+6940%	0.753%	0.387

Table 2. Persistence of shocks, Russian language only.

”cheerful” and ”laugh” as well as ”fear”, ”afraid”, ”scary” and ”worry” saw noticeable changes immediately following 2/24, but have in both cases quickly regressed back towards their baseline levels.

In the following section, we expand upon this work and apply more formalized topic modeling to our data set.

4. TOPIC MODELING

In general, topic modeling is an unsupervised machine learning technique that helps us classify topical patterns across large collections of textual information. One such useful technique is Latent Dirichlet Allocation (LDA), which is based on the assumption that each document contains various topics, where documents with similar words are of the same topic and that the most recurring words together have the same topic. Consider the following list of words: ‘pizza’, ‘dog’, ‘tacos’, ‘cat’, ‘burgers’, ‘kangaroos’, ‘dog’, ‘cat’. Intuitively, we would group the food items ‘pizza’, ‘tacos’, and ‘burgers’ into one category, while designating ‘dog’, ‘cat’ and ‘kangaroos’ into another representing the topic, animals. We can further classify the recurring words (‘dog’ and ‘cat’) as household pets. Likewise, by grouping similar words and identifying the frequent occurrences of a collection of words (i.e., a topic), we are able to quantitatively characterize our randomly sampled collection of tweets.

To develop the topic model and examine changes as a result of the invasion, we took two subsets of the previously described sample of tweets: ”prewar”, containing 239,252 Tweets from November 24, 2021 to February 17, 2022; and ”postwar”, containing 182,458 Tweets from February 25 to May 3, 2022. We further separate the two subsets by language. While debatable, we assume tweets in Ukrainian pertain to individuals of Ukrainian descent while tweets in Russian represent individuals from Russia. However, due to the geographical proximity of the two countries and sharing various cultural similarities including language, it is plausible that individuals living in Ukraine speak Russian as their first language and vice versa.

In general, LDA trains by examining the frequency of a word across all documents. The model treats all words independently, regardless of its position in the document or the surrounding words. Thus, before moving on to building the LDA model, we began with some preliminary processing of the tweets. For each tweet, we lowercase all characters and removed punctuations, thus making the tweets more compliant for our analysis. Regardless, neither of these steps would have a significant impact on how the LDA model identifies words or topics.

Indeed, in any language, stop words like “the” or “a” are the most commonly used words and carry very little useful information. Evidently, **Figure 4** shows that the most frequent words that occur in our collection of tweets, are the common Ukrainian and Russian stop words. From here, it was deemed appropriate to remove such stop words. With the help of

Keyword(s)	Baseline	Extreme	Shock	Current	R
”war”	1.89%	5.42%	+187%	4.11%	0.166
”tank”	1.03%	3.68%	+258%	1.35%	0.432
”happy”, ”cheerful” or ”laugh”	5.57%	3.50%	-37.3%	5.03%	0.337
”help”	0.962%	3.53%	+267%	2.19%	0.241
”Ukraine” or ”Ukrainian”	16.5%	27.1%	+64.8%	17.5%	0.402
”Russia” or ”Russian”	6.10%	18.1%	+197%	10.9%	0.296
”fear”, ”afraid”, ”scary” or ”worry”	0.808%	2.09%	+159%	0.930%	0.418
”Zelenskyy”	5.03%	0.523%	-89.6%	1.56%	0.106
”Putin”	1.83%	6.02%	+229%	1.76%	0.456
”Mariupol”	0.174%	2.28%	+1210%	3.96%	-0.386
”Donbass”	0.798%	0.00%	-100%	0.829%	0.495
”child”	1.25%	3.34%	+167%	2.02%	0.307
”love”	1.50%	0.456%	-69.5%	1.27%	0.379
”bravery”, ”courage” or ”hero”	0.515%	2.89%	+461%	1.15%	0.334
”homeland”, ”fatherland” or ”country”	0.538%	1.81%	+236%	0.372%	0.507
”Kyiv”	2.11%	11.8%	+459%	3.17%	0.411
”Kharkiv”	0.652%	7.29%	+1020%	1.87%	0.395

Table 3. Persistence of shocks, Ukrainian language only.

Python’s package advertools, we were able to identify 391 stop words for the Ukrainian language and 264 stop words for the Russian language. Removing these uninformative words not only reduced the dimensionality of our data but also provided us with compelling insights.

Figure 5 displays the translation of the top 15 recurring words in Ukrainian and Russian tweets, respectively. It is worth noting that even after removing stop words like “and” or “is”, conjugates of such words still remain in our collection – one nuisance of text processing in non-English languages. Each language has its own unique composition from nouns to verbs. For that reason, this may not translate well across platforms and even excellent packages like NLTK could fail to process text with such precision. With regards to other most common words, we see that before the invasion, tweets in Ukrainian consisted high volume of words like “Ukraine” or “Zelenskyy”, emphasizing the above explained nationalistic pride. On the contrary, post-invasion, we see a decline in the usage of words related to “Zelenskyy” but a rise in words like “armed forces”. We also see a rise in retweets from news outlets, especially those reporting daily stories of the on-going war in Ukraine.

Moving on, LDA treats each text (tweets) as a set of words, otherwise known as a bag of words. It does not account for the surrounding context or the position of the word. Hence, we tokenized each tweet and stemmed all tokens into their root word. This was relatively easy to accomplish for Russian tweets by virtue of Python’s NLTK library. However, the Ukrainian language is not supported by any of the pre-existing packages. But on GitHub, we were able to find a Ukrainian stemming program written by a native,named Kyrylo Zakharov (Zakharov, 2014). From there, we converted the stemmed tokens into a dictionary object, essentially creating tuples of words and their relative word count. Finally, we employed the LDA model in hopes to identify the mixture of topics that could be present in a tweet. We kept our parameters simple and set the number of topics to 20. Top tokens in each topic and their respective English translations are present in **Figure 6**.

As evident in **Figure 7**, topic 1 dominates the number of tokens found in Ukrainian tweets post-invasion, accounting for 9.8% of the tokens. Additionally, topics 3 and 4 account for the next highest number of tokens, specifically 14.2% in total. At first glance, it’s quite difficult to categorize tokens from these two topics. However, a few words still seem strikingly significant such as Crimea, waters, histories, Moscow, and Ukrainian. Crimea could relate to the Crimean Peninsula near the Black Sea, which is predominately made up of Russian and Ukrainian minorities. Interestingly, this is where the Russian-Ukraine revolution began back in 2014. Note that prior to the invasion, topic 1 (9% of tokens) and topic 2 (7.7% of the tokens), also comprised of tokens like ’Crimea’, ’Putin’, ’troops’, ’Donbas’, ’tank’, ’how’, ’occupied’, ’Ukrainian’, and others. But, it also included other non-political words in the top two topics like ’faith’, ’truth’, ’great’. With further translation, we conclude that topic 1 for post-invasion tweets includes words such as ‘Zelensky’, ‘president’, ‘USA’, ‘support’, ‘war’, ‘Poland’, ‘people’, ‘UN’, and ‘right’. Although questionable, we could assume the Ukrainians are calling for the USA or UN for support.

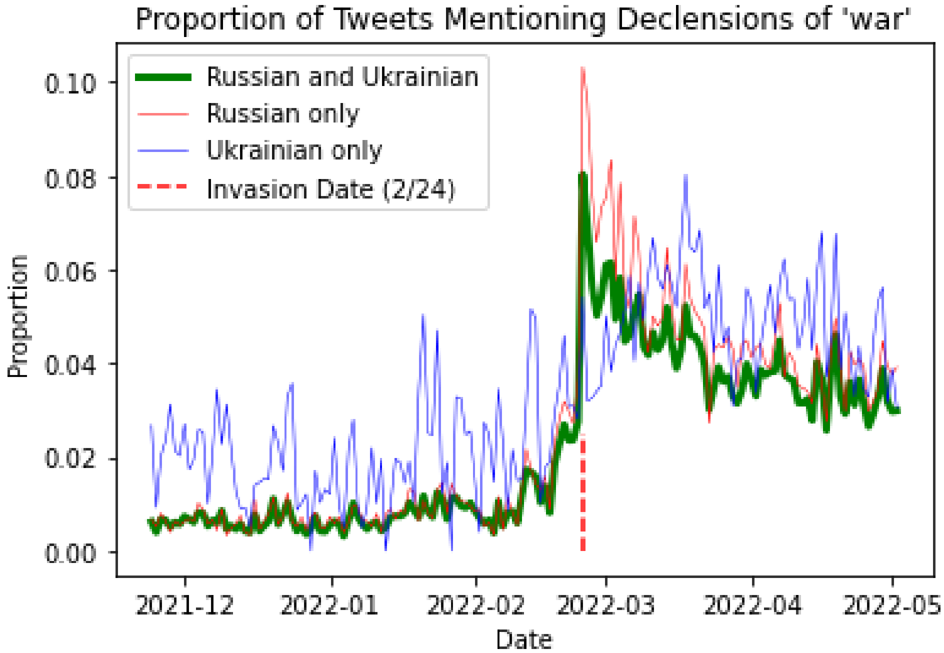


Figure 3. By language, the daily proportion of tweets mentioning ”war”.

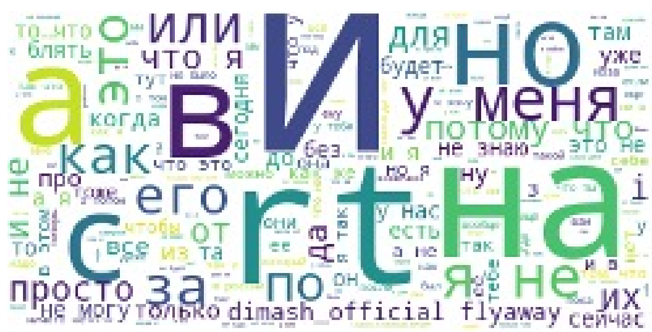


Figure 4. Word Cloud representing the most common words across tweets before any significant pre-processing such as removing stop words.

Whereas, top tokens in Russian tweets before for the invasion are dominated by political related words like Ukrainian’, ’Russian’, ”Putin’, ’military’, ’armed’, ’forces’, ’Zelenskyy’, ’territory’, ’missiles’, ’Ministry of Defense’ amongst others. This seems to indicate that tension between the two countries was already present, long before the 2/24 invasion. As shown in **Figure 8**, post-invasion Russian tweets are dominated by few topics: 3, 4 and 5, totaling 20.9% of the tokens. To our surprise, topic 3 (7.4% of tokens) does not explicitly state words related to government or politics. Meanwhile, we are able to distinguish tokens from topic 4 and 5 relating to the current situation between Russia-Ukraine. While debatable, we can assume emotions such as anger due to the curse words present, or fear or sadness from the mentions of words like bombs, kill, and weapons. Regardless of

Moving forward, we could possibly use the LDA model to further determine the distribution of topics in new tweets, as the current situation in Russia-Ukraine is heightening by the day. Even though our LDA model is capable of forming clusters of words that belong to each topic, it still has its limitations. We may never know the true context in which these words were spoken, especially because the model is limited to a bag of word structure rather than a sentence-based one. Furthermore, while speculating Figures 7, 8, 9 and 10, we can visualize topics that are of close proximity, hinting that they would be closely related in topical structure and content. However, when translated, it is quite difficult to group topics together. Limitations in our project call for further technological advances in building precise packages to process text in non-English languages as well as building a model for sentence-based text analysis.

REFERENCES

Go, A., Huang, L., and Bhayani, R. (2009). Twitter sentiment analysis. *Entropy*, 17.

Kemp, S. (2022a). Digital 2022: Russia. https://datareportal.com/reports/digital-2022-russian-federation.

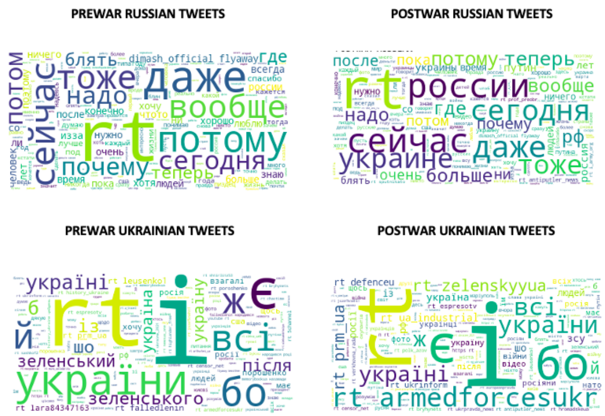
Kemp, S. (2022b). Digital 2022: Ukraine. https://datareportal.com/reports/digital-2022-ukraine.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Proceedings of the international AAAI conference on web and social media*, 5(1).

Twitter (2022). Filtering tweets by location. https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.

Zakharov, K. (2014). Ukraine stemmer. https://github.com/Amice13/ukr_stemmer/blob/master/ukr_stemmer.py.

Zimbra, D., Abbasi, A., Zeng, D., and Chen, H. (2018). The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29.



TOP 15 FREQUENT WORDS IN PREWAR TWEETS RUSSIAN

	Tokens	English Translation	Frequency
0	rt	retweet or russia today	1.000000
1	даже	even	0.220797
2	потому	because	0.215491
3	сейчас	now	0.195274
4	тоже	too	0.183398
5	вообще	generally	0.180365
6	сегодня	today	0.160907
7	надо	necessary	0.149662
8	почему	why	0.146630
9	потом	after	0.145492
10	теперь	now	0.143913
11	блять	fuck	0.142018
12	где	where	0.132605
13	dimash_official flyaway	dimash_official flyaway	0.131404
14	больше	more	0.129825

TOP 15 FREQUENT WORDS IN POSTWAR TWEETS RUSSIAN

	Tokens	English Translation	Frequency
0	rt	rt	1.000000
1	россии	Russia	0.207539
2	сейчас	now	0.206448
3	даже	even	0.172446
4	сегодня	today	0.150617
5	украине	Ukraine	0.133826
6	тоже	too	0.130384
7	теперь	now	0.128705
8	потому	because	0.128453
9	где	where	0.117790
10	больше	more	0.115775
11	вообще	generally	0.114600
12	рф	rf	0.112333
13	надо	must	0.111410
14	почему	why	0.109898

TOP 15 FREQUENT WORDS IN PREWAR TWEETS UKRAINIAN

	Tokens	English Translation	Frequency
0	і	and	1.000000
1	rt	russia today or retweet	0.452156
2	є	is	0.132519
3	україни	of Ukraine	0.109569
4	бо	for	0.098279
5	ж	Translation not found	0.087914
6	всі	all	0.084768
7	й	and	0.080326
8	україні	Ukraine	0.076994
9	б	Translation not found	0.074218
10	із	from	0.066074
11	шо	Translation not found	0.062558
12	зеленського	Zelenskyy	0.060152
13	після	after	0.055710
14	зеленський	Zelenskyy	0.053304

TOP 15 FREQUENT WORDS IN POSTWAR TWEETS UKRAINIAN

	Tokens	English Translation	Frequency
0	і	and	1.000000
1	rt	russia today or retweet	0.391470
2	є	is	0.139777
3	бо	for	0.102853
4	rt armedforcesukr	armedforcesukr	0.099227
5	всі	all	0.086251
6	україни	of Ukraine	0.081767
7	україні	Ukraine	0.081004
8	ж	Translaiton not found	0.076138
9	rt zelenskyyua	rt zelenskyy ua	0.074897
10	rt prm_ua	rt prm_ua (News Channel in Ukraine)	0.071272
11	й	and	0.069745
12	rt ua_industrial	Twitter account with news of war in Ukraine	0.067932
13	зсу	Armed forces of Ukraine	0.065833
14	шо	Translaiton not found	0.063925

Figure 5. After removing stop words, the above word clouds represent the most common words across our collection of tweets before Russia's invasion into Ukraine and after the invasion. We further extract the top 15 words occurring in Ukrainian and Russian tweets in terms of word frequency.

POST WAR RUSSIA

	Top Words: Topic 3	English Translation: Topic 3	Top Words: Topic 4	English Translation: Topic 4	Top Words: Topic 5	English Translation: Topic 5
0	блят	fuck	русск	Russian	rt	rt
1	дела	deeds	rt	rt	украин	ukrain
2	рад	happy	убива	kill	путин	putin
3	правильн	correct	уб	kill	войн	war
4	rt	rt	арм	arm	росс	ross
5	рассказыва	telling	украинц	Ukrainian	побед	victories
6	сук	bitch	девочк	girls	народ	people
7	нах	nah	земл	land	посл	last
8	заб	zab	корабл	ship	суд	court
9	сдела	done	мир	world	стран	countries
10	начина	beginning	фашист	fascist	российск	rossiysk
11	друз	druz	смерт	death	оруж	weapons
12	слишк	too	долж	duty	бомб	bombs
13	проблем	problems	путин	Putin	мир	world
14	нашл	found	херсон	Kherson	нача	beginning

POST WAR UKRAINE

	Top Words: Topic 1	English Translation: Topic 1	Top Words: Topic 3	English Translation: Topic 3	Top Words: Topic 4	English Translation: Topic 4
0	rt	rt	мов	languages	rt	rt
1	україн	ukraine	rt	rt	і	and
2	і	i	крим	Crimea	хоч	though
3	зеленськ	zelensk	дум	dum	бо	because
4	президент	president	повн	full	друг	friend
5	zelenskyyua	zelenskyyua	істор	histories	пан	gentleman
6	сша	usa	дитин	children	москв	moscow
7	ма	ma	росн	rosn	час	time
8	росі	rosi	вод	waters	зна	knows
9	підтримк	support	завд	zavd	хлопц	guy
10	війн	wars	ок	ok	хорош	good
11	польщ	poland	давайт	davait	робит	does
12	рад	rad	справ	sprav	момент	moment
13	народ	people	питанн	questions	ніч	night
14	рос	ros	головн	main	жит	live
15	оон	un	насправд	actually	твіт	tweet
16	прав	rights	українськ	Ukrainian	неймовірн	incredible
17	російськ	russian	40	40	всі	all
18	новин	news	кот	cat	нагад	reminder
19	володимир	Vladimir	депутат	deputy	медведчук	bear

Figure 6. Top 3 topics from Ukrainian and Russian LDA model are present along with the tokens and their English Translation that make up each of the topic are displayed above.

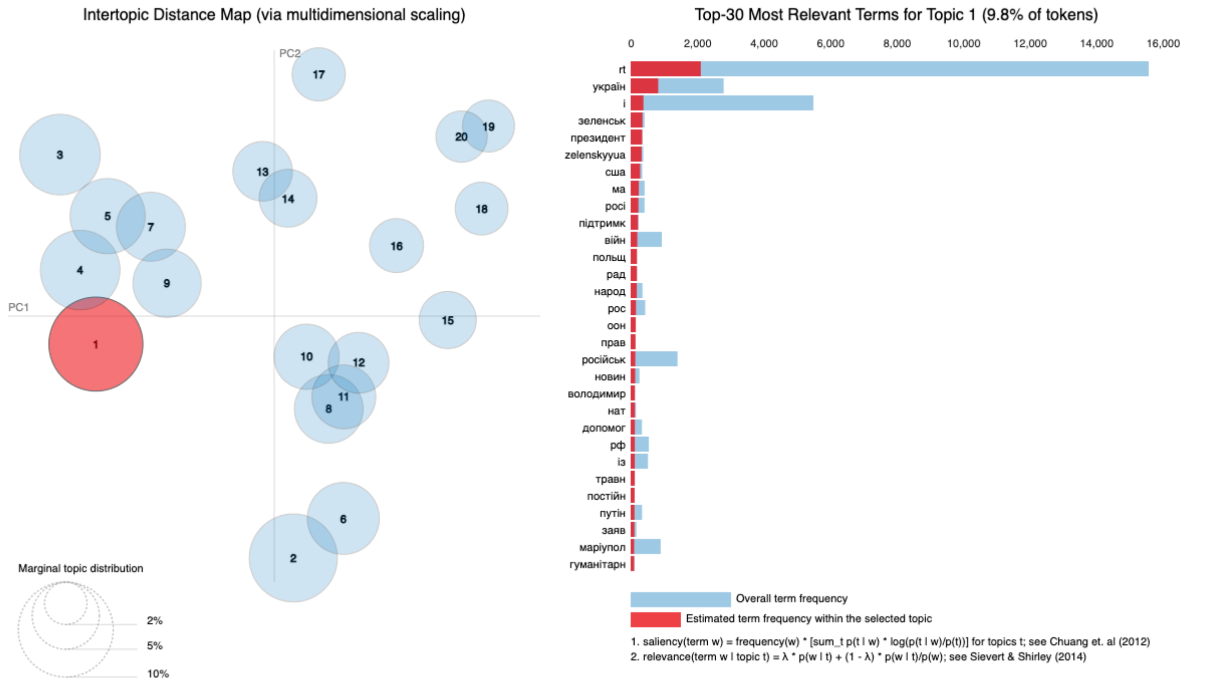


Figure 7. Distribution of topics in Ukrainian tweets post-invasion, along with the most relevant terms present in a particular topic. Please find the English translation of such tokens in Figure 6.

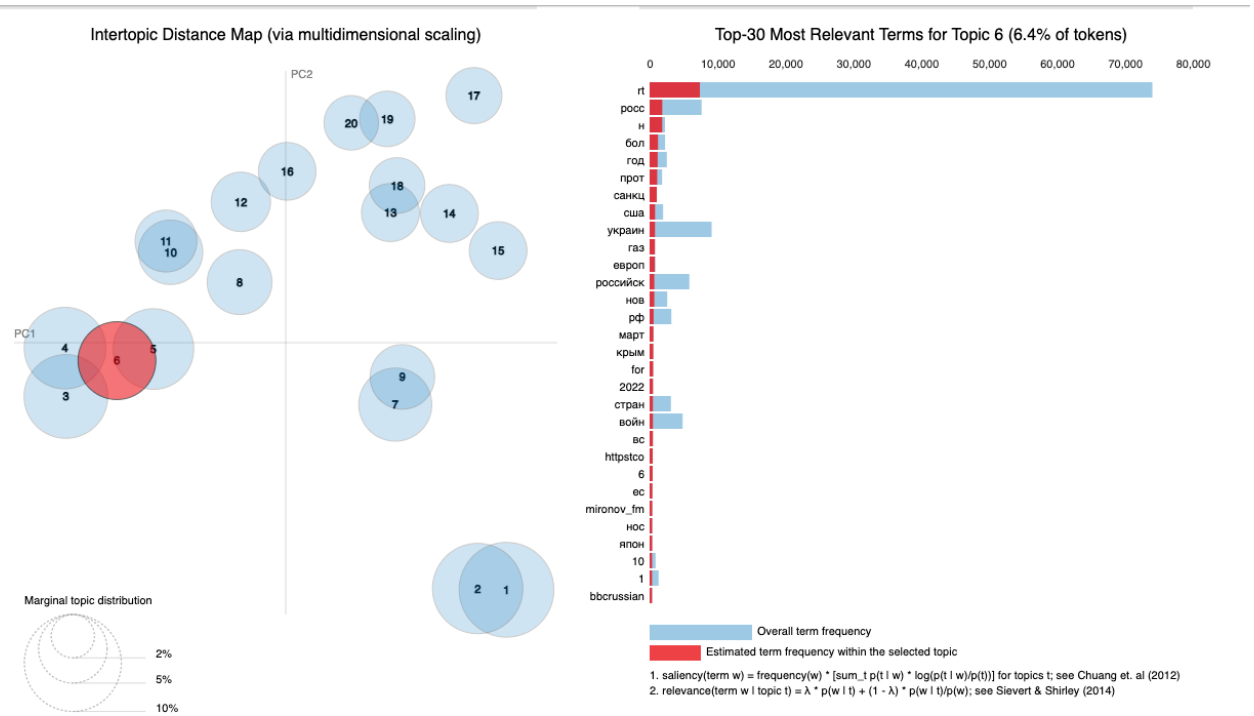


Figure 8. Distribution of topics in Russian tweets post-invasion, along with the most relevant terms present in a particular topic. Please find the English translation of such tokens in Figure 6.

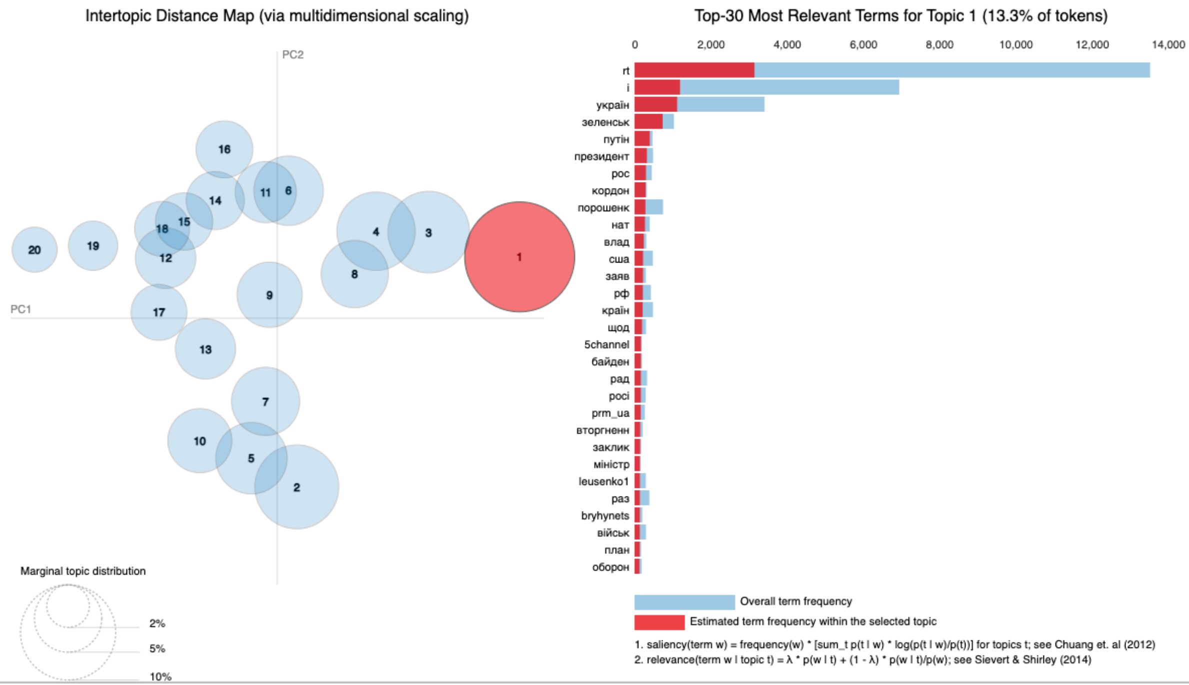


Figure 9. Distribution of topics in Ukrainian tweets pre-invasion, along with the most relevant terms present in a particular topic. Please find the English translation of such tokens in Figure 6.

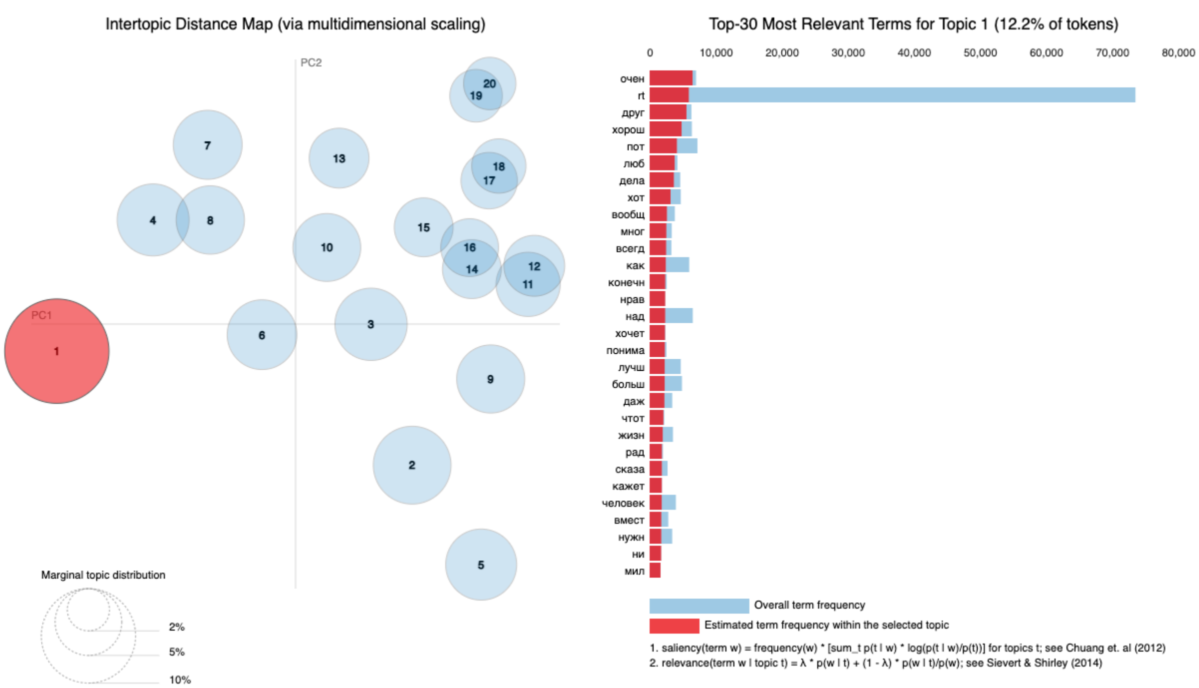


Figure 10. Distribution of topics in Russian tweets pre-invasion, along with the most relevant terms present in a particular topic. Please find the English translation of such tokens in Figure 6.