Clustering of Tweets Related to the Ukraine Crisis

Evan Ozaroff, Maya Webb

1. Introduction

On February 24th, 2022, Russia began a full scale invasion of Ukraine. It goes without saying that this crisis will be divisive and politically charged. In recent years, social media has become the main outlet for individuals to voice their opinions on current events. We would like to investigate the public conversation on the Ukraine crisis based on tweets related to the topic in recent days. Using unsupervised clustering techniques we would like to determine if there are any identifiable divisions in the public conversation surrounding the invasions.

Work on this project was done collaboratively, however Maya Webb had special focus on cleaning, transformation, and EDA, while Evan Ozaroff had special focus on clustering implementation and evaluation.

2. Related Work

The most popular implementation of clustering data from Twitter is to do a sentiment analysis of tweets pertaining to a specific topic. Our goals were somewhat different, as our primary focus was to use clustering to identify the varying conversations surrounding the crisis. Our goals are similar to topic modeling. An article we took inspiration from is "Topic Clustering of Tweets" [1] but instead of predetermining our topics we used the results from our unsupervised clustering to infer the topics of conversation present in the body of tweets we selected.

3. Background Information

The Russian invasion of Ukraine began on February 24th, 2022. Immediately following the news of the invasion, conversation erupted on social media platforms worldwide. Conversation on English-speaking platforms was largely characterized by condemnation of the invasion, as well as calls for a range of responses from the UN, western governments, NATO, and humanitarian organizations. During the early stages of the conflict, there was also significant public conversation regarding the possibility of implementing a "no-fly zone" over Ukrainian airspace.

4. Proposed Approach

The goal of our project is to implement unsupervised clustering techniques on a corpus of Tweets from the first week of the Russian invasion of Ukraine. We will then use the results of our clustering to remark on the various topics of conversation occurring on Twitter.

Prior to clustering, significant cleaning and pre-processing of the textual tweet data was required. The first step in our cleaning was dropping non-english language tweets so that the full dataset was understandable. Then, operating solely on the textual data from the tweets (discarding all other metadata), we performed tokenization. Following tokenization, we began extensive cleaning of the tweets. Cleaning operations included the dropping of punctuation, URLs, Twitter

user handles, and words under two characters in length. Finally, we performed lemmatization, the reduction of words to their grammatical base form. Cleaning and preprocessing steps were aided by the use of Python's NLTK (Natural Language Toolkit) library.

The next step was to convert the tokenized tweets to a Term-Frequency Inverse Document-Frequency (TF-IDF) vocabulary matrix. Computing TF-IDF scores allows us to retain the most characteristic words in the corpus and discard words that are common across a large number of the tweets. We limited our vocabulary size to 10,000 words. In order to run clustering efficiently, it was necessary to reduce the dimensionality of our large, sparse input data set. To do so, we employed sklearn's TruncatedSVD, a fast randomized SVD solver commonly used for Latent Semantic Analysis. We reduced our vocabulary of 10,000 words to 100 latent semantic dimensions.

Finally, we were able to use our reduced data to perform clustering. We implemented two models, k-Means and Gaussian Mixture Modeling (GMM), using sklearn's MiniBatchKMeans and GaussianMixture libraries respectively.

5. Experiments Section

We were able to retrieve a dataset from Kaggle containing tweets pertaining to the Ukraine crisis [2]. The creator of the dataset has three jupyter notebooks that run every fifteen minutes 24/7 to scrape the tweets. Roughly 9,000 tweets are scraped in each of the fifteen minute intervals. The criteria for scraping is a predetermined list of hashtags related to the topic. Each day was its own gzipped csv file, we merged all seven csvs to create our final working dataset which initially contained 2,984,341 rows. After dropping tweets that were not in English there were 1,961,759 rows in our final working dataset and and 18 columns labeled as:

['userid', 'username', 'acctdesc', 'location', 'following', 'followers', 'total tweets', 'tweet id', 'retweet count', 'text', 'hashtags', 'language', 'coordinates', 'favorite count', 'extractedts']

We only focused on the 'text' column in our clustering models.

Our initial step was to perform an exploratory data analysis (EDA) on the text and hashtag columns of the dataset. Figure 1 showcases a word cloud visualizing the words with the highest frequency in the dataset (prior to cleaning and tokenizing). Figure 2 showcases the top 10 most frequent hashtags in each tweet. The visualizations helped us see if there were any obvious conclusions we can predict before running our clustering.

Figure 1: Word Cloud

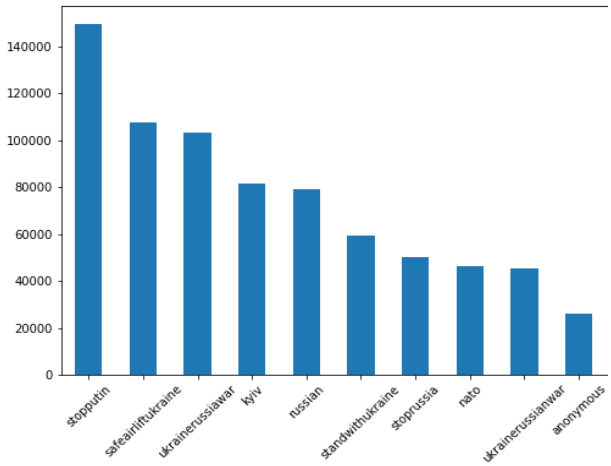

Figure 2: Top 10 Hashtags

Following EDA, we were able to begin our clustering process. The first step was to select an appropriate number of clusters. To do so, we plotted the SSE for k-values in the range of 1-10 for kMeans clustering and determined that the elbow value k = 3 was the optimal number of clusters. To allow comparison of our kMeans and GMM models, we selected k = 3 for our GMM model as well.

The results of our clustering are visualized by the tSNE plots shown in Figures 3 and 4. Immediate visual inspection of these plots tells us that in neither case were the clusters well separated, however the clusters produced by kMeans do appear more coherent than those produced by the GMM. In both kMeans and GMM, there was a significant imbalance in the number of tweets falling into each cluster. For kMeans, about 92% of the tweets fell into Cluster 0, while only about 2% and 6% fell into Clusters 1 and 2 respectively. For GMM, 47% of tweets fell into Mixture 0, 49% into Mixture 1, and 4% into Mixture 2.

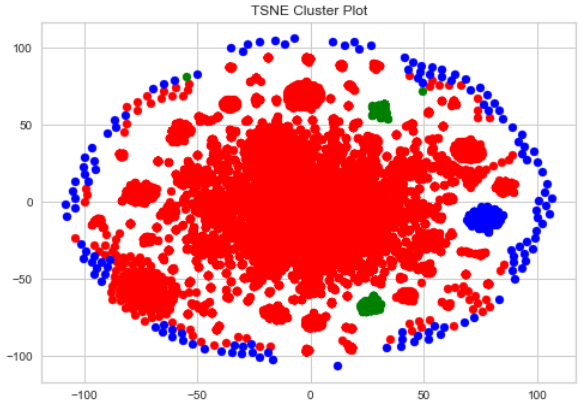

Figure 3: tSNE for k-Means

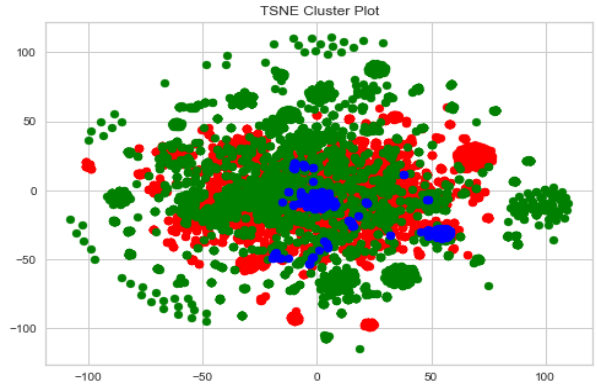

Figure 4: tSNE for GMM

To understand the features defining each cluster of tweets, we generated word clouds displaying the most characteristic words in each cluster. This was done by averaging the TF-IDF scores for

all tweets in each cluster, and returning the highest average TF-IDF value words. The results are displayed in Figure 5.

Figure 5: Word Cloud for Each Cluster/Mixture

We were pleased to see some identifiable relationships between the most characteristic words in each cluster. In the case of kMeans, Cluster 0 seems capture more general, non-specific tweets about the war in Ukraine ('russian', 'ukraine', 'war', '#ukrainerussiawar'). We have referred to Cluster 0 as the "General" cluster. Cluster 1 seems to capture those tweets from the first week of the invasion that were most concerned with the humanitarian impact of the war ("civilian", "innocent", "humanitarian", "defend"). We have referred to Cluster 1 as the "Humanitarian" cluster. Finally, Cluster 2 seems to capture those tweets related to the No-Fly zone discussion occurring during the first week of the war ("security", "council", "close", "sky"). We have referred to this cluster as the "No-Fly Zone" cluster. Interestingly, the mixtures determined by our GMM seem to closely mirror the clusters defined by our kMeans clustering. Mixture 0 appears to be the "Humanitarian" mixture, Mixture 1 the "No-Fly Zone" mixture, and Mixture 2 the "General" mixture.

Finally, we evaluated the quality of our clustering using silhouette score. Computation of the silhouette score on the full data set of 1,961,759 items and 100 dimensions was prohibitively expensive, so instead silhouette score was computed on a subset of the data reflecting the original cluster distribution. Figure 6 plots the silhouette score for points in each of the clusters generated by kMeans. The average silhouette score for points in Cluster 0 (General) was lower than for points in Clusters 1 (Humanitarian) and 2 (No-Fly Zone), suggesting that Clusters 1 and 2 were more compact and coherent than Cluster 0.

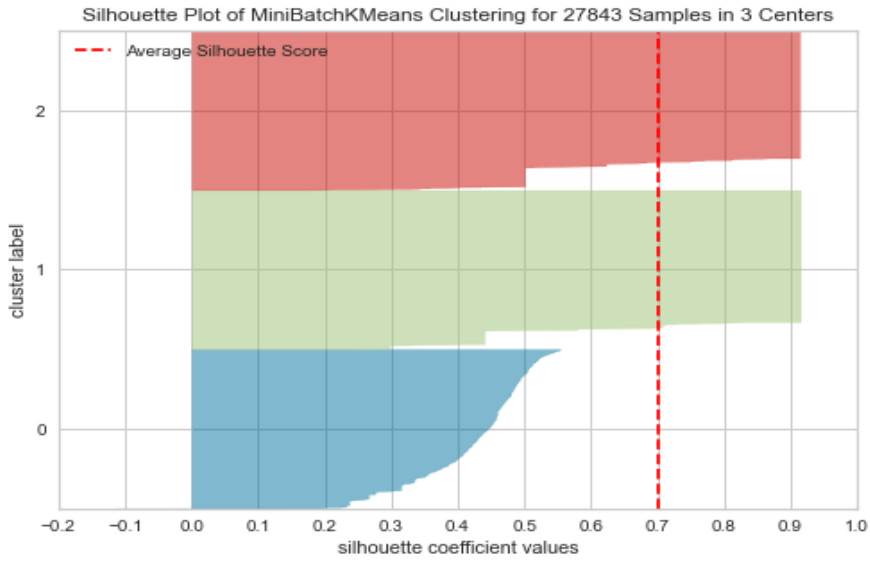

Figure 6: Silhouette Plot

6. Conclusions and Future Work

Our project was successful in performing clustering of tweets to identify varying public conversations surrounding the war in Ukraine. Entering the project, we expected to have difficulty interpreting the results of our clustering, as we did not expect the twitter data to be well separated. This expectation was confirmed when viewing the tSNE plots for both kMeans and GMM. However, our clusters did appear to identify understandable and relevant conversations from the first week of the invasion, specifically surrounding the humanitarian consequences of the conflict and the potential enforcement of a no-fly zone. Furthermore, the silhouette scores for these clusters suggests that they are compact in the latent semantic space. In other words, our clustering was successful in identifying tweets engaging in similar conversations surrounding Ukraine. As an increasing amount of our public conversations occur online, methods to understand and generalize the endless amount of conversational input are increasingly valuable. Our results show that clustering is a viable method for using large textual datasets to understand what conversations are being had in relation to a broader topic on social media.

If we had more time, our results could likely have been improved in numerous ways. Firstly, we understand that public conversations are not static, but rather evolve over time. By considering tweets as a time series, rather than static corpus, we could use clustering to identify how the conversations change over time. Additionally, hierarchical clustering could be considered, which could potentially allow us to identify sub-topics present within larger conversations.

7. References

[1] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, & Robert Frederking. (2011). *Topical Clustering of Tweets.* Available at: <https://www.researchgate.net/publication/267805712_Topical_Clustering_of_Tweets>

[2] Kaggle.com. 2022. *Ukraine Conflict Twitter Dataset (27.59M tweets)*. [online] Available at: <https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows>