Narrative

Ken Krmoyan

12/11/2020

Brief substantive background / goal

Armenia and Azerbaijan have been involved in a decades-long, frozen conflict over a disputed area known as Nagorno-Karabakh. The conflict traces its roots back to the beginning years of the Soviet Union, when the mostly Armenian-populated region of Nagorno-Karabakh was to be placed under the control of the Azerbaijani Soviet Socialist Republic (SSR). However, as the USSR was weakening in the late 1980s, ethnic violence began in both countries and became a full-scale war. Ultimately, Armenia annexed the disputed region by 1994 but it was still internationally recognized as Azerbaijani territory. Thus, on the one hand, Armenians fight for their right to self-determination, pushing the international community to recognize the region as an independent state or part of the Republic of Armenia. On the other hand, Azerbaijan pushes for territorial integrity, arguing that Armenia illegitimately occupied Azerbaijani land.

The conflict was never formally concluded and the ceasefire of 1994 was constantly violated by both sides. A full-scale war resumed on September 27, 2020 and lasted until November 10, 2020. However, the geopolitical balance of power in the region changed vastly as Azerbaijan gained massive amounts of wealth through its oil reserves, and Turkey started supporting Azerbaijan by providing military aid and holding joint military exercises. In fact, Russian and French intelligence have confirmed that Turkey is sending Syrian fighters from jihadist groups to fight against Armenia. The conflict ended through a Russia-brokered ceasefire agreement, rseulting in a large portion of Nagorno-Karabakh being transfered to Azerbaijan.

This projects looks at another novelty of 21st-century warfare: online rhetoric by warring parties. Specifically, I scrape tweets from Armenia and Azerbaijan's Ministry of Foreign Affairs (MFA) accounts from the period of the last outbreak of war (Sep 27 - Nov 10). I use distinct words and sentiment analysis to detect patterns about the kind of rhetoric each party uses to characterize itself and its enemy and whether those patterns changed throughout the period in question. Understanding the large patterns in rhetoric is important because the MFA Twitter accounts are used to share official statements with both citizens and the outside world.

Collecting data

I collect the tweets from the two accounts (@MFAofArmenia and @AzerbaijanMFA) and filter the dataframes to include the time period of interset. I store the two dataframes (`arm` and `aze`), which include the author (`screen_name`), the date (`created_at`), and the text of the tweet (`text`). These dataframes are available as `.csv` files in the "Data" folder.

Cleaning / pre-processing data

To clean/pre-process the data, I write four functions to facilitate the process: (1) `clean_up()`, (2) `about()`, (3) `split_about()`, (4) and `preprocessing()`.

First, I write the function `clean_up(i, country)` to clean the text of the tweets, where i is the number of the tweet and *country* is the dataframe. The function removes URLs, mentions, ampersands, emojis &

special characters, and splits hashtags (e.g. "#StopAzerbaijaniAggression" becomes "# Stop Azerbaijani Aggression"). Afterwards, I apply the function to all tweets within each dataframe and replaced the old texts with the cleaned output in the same column `text`. The mapping of the function to the dataframe was trickier than usual because my function had two inputs. After some research, I figured out that I had to put the second input (`arm`) after the function, like below:

```
arm <- arm %>%
  mutate(text = map_chr(1:nrow(arm), clean_up, arm)) # applying fn to Arm
```

Second, I write the function `about(i, country)` to categorize the tweets as either talking about Armenia or Azerbaijan, which was the most challenging part of the assignment. The intuition behind this comes from the fact that most tweets by the MFAs of Armenia and Azerbaijan talk either about themselves or the enemy. For instance, Armenia's 66th tweet is clearly about Azerbaijan:

> "Azerbaijan's policy of crimes against humanity ; ethnic cleansing is doomed to failure ; will encounter the resolute resistance of the people of # Artsakh, which will be carried out through all the means necessary for self-defense."

Another example is Azerbaijan's 16th tweet, which is clearly about Armenia:

> "In a blatant violation of international law, including international humanitarian law during almost 30 years, # Armenia commits armed #aggression against # Azerbaijan. Press Release of the Press Service Department of the M F A of of Azerbaijan"

Based on my observation of many tweets in the dataframe, the country that is mentioned first in a tweet is usually what the tweet is about. In the examples above, *Azerbaijan* is mentioned first in Armenia's 66th tweet, while *Armenia* is mentioned before *Azerbaijan* in Azerbaijan's 16th tweet. The function `about()` (1) takes each tweet, (2) splits it into words, (3) records the index position of the earliest appearing Azerbaijan-related and Armenia-related word through the `match` function, and (4) returns the categorization of either `arm` or `aze` based on which one appeared first through an `if` function. If either of the words were not present (like in the case of Armenia's 66th tweet above), the function automatically takes the other one to be the subject of the tweet. If none of them is present, the function codes it as `NA`. This logic does not always work perfectly (see example below, where the function would code is as `aze` but the statement is not necessarily focused on Azerbaijan), but it is the best I could do to automate the categorization of the tweets based on what they talk about. Just like in the case of `clean_up()`, I then apply the function to both dataframes.

> "Statement by the President of the Republic of Azerbaijan, the Prime Minister of the Republic of Armenia and the President of the Russian Federation"

Third, I write the function `split_about(author, subject)` to produce new sub-dataframes based on who the tweet is talking about. I created four separate dataframes: (1) tweets by Armenia talking about Armenia, (1) tweets by Armenia talking about Azerbaijan, (3) tweets by Azerbaijan talking about Azerbaijan, and (4) tweets by Azerbaijan talking about Armenia. I also combine them into a list called `all` for preprocessing.

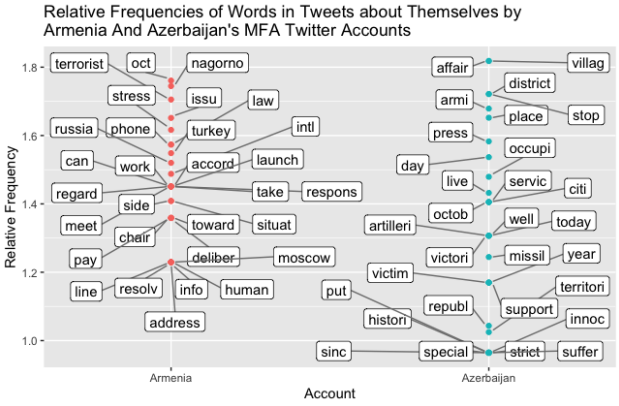
Fourth, I write the function `preprocessing(author_about_subject)` so that each dataframe would get preprocessed for text analysis (i.e. removing stop words, numbers, punctuations, making all words lower case, and stemming). I then map the function to `all`.

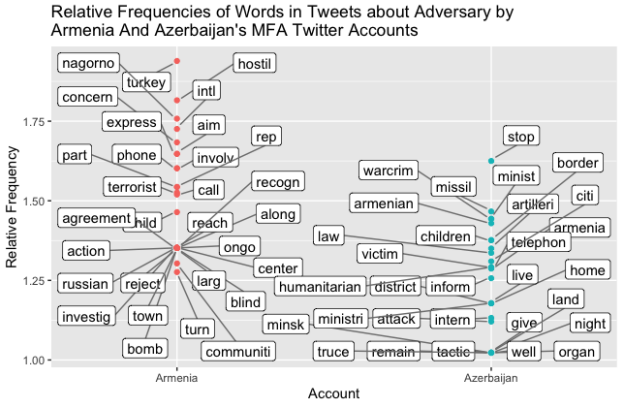
Analysis and visualization

I carry out two broad types of analysis: frequent/distinct words and sentiment analysis.

First, I write a function `word_cloud(i)` which creates a wordcloud for the four sub-dataframes created in the cleaning stage. The i in the function denotes the component in the list `all`. I then create a wordcloud for each sub-dataframe.

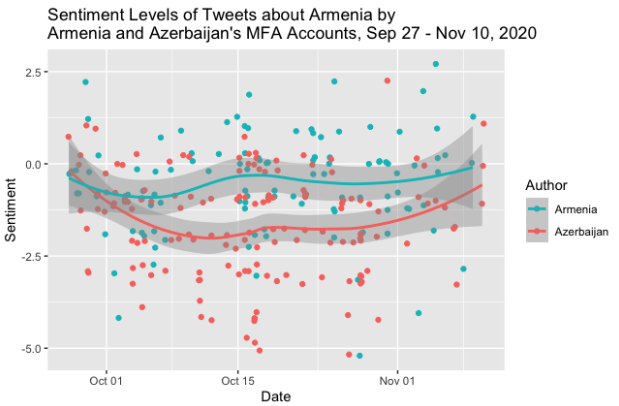
For distinct words, I first create a function `freq(doc1, doc2)` which takes two sub-dataframes, combines them into one dataframe, preprocesses it, splits back into two, calculates sums of columns across documents (tweets), calculates the average rate of word use in each, sorts them in descending order, and outputs the scores. I then apply the function to calculate scores for two categories: (1) tweets talking about themselves and (2) tweets talking about the adversary. I create two dataframes, each with the top 30 words obtained by Armenia and Azerbaijan. I visualize the results by grouping the observations based on whether they were written by Armenia or Azerbaijan.

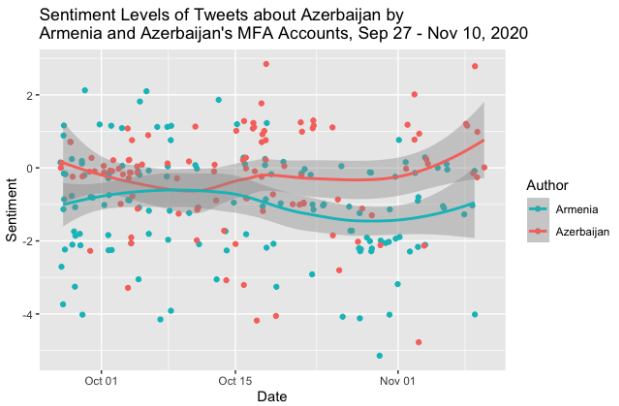


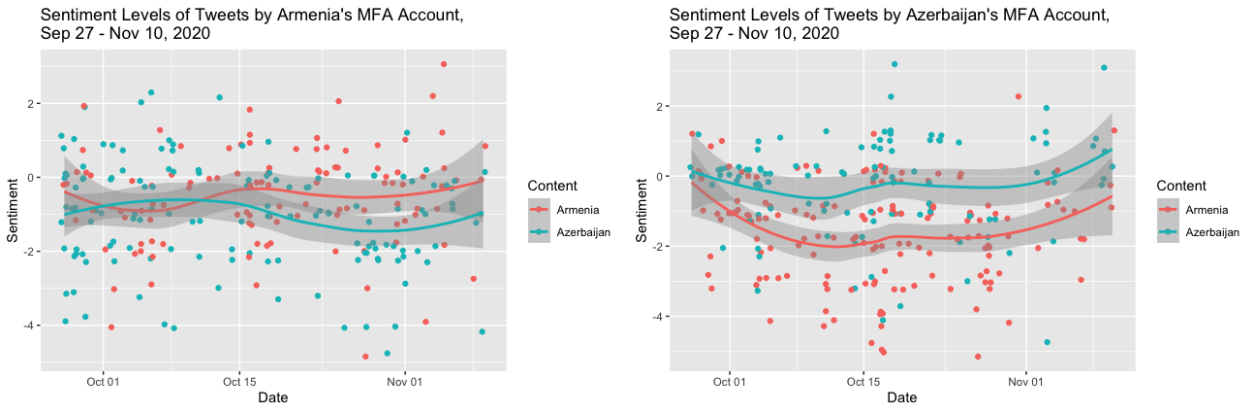


While some of the words in the graphs above are not very insightful, there are some interesting patterns I noted. In tweets about both themselves and the adversary, Armenia used the words "Turkey" and "terrorist" a lot, presumably to publicize the threat to Armenians. Armenia also used "Russia", "accord", "Moscow", "phone": words that are associated with a diplomatic solution to the problem. This is not surprising given that it would be in Armenia's interest to end the conflict peacefully since it was in possession of the region. From Azerbaijan's side, we see the use of the words "district", "occupation", "territory", "land", "border": words associated with territorial integrity.

Second, I carry out sentiment analysis. I first load the "bing" dictionary and assign scores to positive and negative words. Afterwards, I write the function `sentiment(author_about_subject, i)`, which calculates the scores and assigns it to the sub-dataframe `author_about_subject` provided. I then apply this function to all four sub-dataframes, combine it with the original `arm` and `aze` dataframes, and combine those two into one called `dat`. I visualize the results by plotting the graph of all tweets and graphs segregated by different dimensions across the time period in question. See below for the segregated versions:







From the graphs, we can see that there is not a lot of variation across time in terms of sentiments. However, one interesting observation is that tweets about Armenia tend to be significantly more negative when they come from Azerbaijan, but not as much the other way around. This is evident when comparing the first pair of graphs, with the one on the left showing a larger distance between the two lines and CIs. A similar trend is apparent in the next second pair, where the graph on the right shows Azerbaijan using a much more negative sentiment toward Armenia than the other way around.

Future work

To continue investigating patterns in online rhetoric, I would try to improve/add two aspects.

First, I would try to refine the methods of categorizing tweets based on whether they are about themselves or the adversary. Given the limitation of the "who-comes-first" approach, I would try to devise an alternative method. I would also work on categorizing the tweets based on whether it talks about itself in the active or passive voice. It'd be interesting to see whether there is a significant difference between the sides in whether they say "We were attacked" or "They attacked". This approach could illuminate aspects of self-perception, specifically about whether the sides implicitly think about themselves or the other as defensive or offensive.

Second, I would extend this analysis beyond Twitter. One limitation of relying on tweets is that there is a lot of repetition resulting from the need to convey information in 280 characters. So, both sides end up just conveying the main message along the lines of "the other side is bad, we are good." To make the analysis more nuanced, I would try to extend it to official press releases on the MFA websites or by country leaders.