Assignment 3

Jinman Xing, Tongye Liu, Vanessa Neeff

December 7, 2020

Which conclusions can we draw about global conflict events from Twitter data?

1 Peace and Conflict around the World

Our study project aims at analysing contemporary tweets about global events around conflict and peace events in different regions of the world. The time-based visualisations of our content analysis and topical categorisation will help us understand how and specifically when different countries correspond to specific conflicts and how subject matters change over time.

Twitter as a micro-blogging service is suitable as a source of opinion data since a vast amount of citizens around the globe express their views on economical developments and politics on the social media platform. Moreover, both regional and global happenings are most often followed by concurrent tweets and postings, thus making Twitter a good medium to explore past and real-time incidents (A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle).

Besides topical categorisation, word frequency and sentiment analysis, Twitter also offers the possibility to gauge regional differences in how incidents affect political opinion making and news propaganda.

Our project focuses on a rather smaller subset of Twitter data from September, October and November 2020 to explore the possibilities and limitations of topical categorisation using Python Jupyter Notebooks. With our exploration, we would like to suggest a starting point for further analyses with larger data sets, which in future could also include results in real time. Analysing and observing trends over a longer period of time will offer

us the possibility to draw conclusions whether it is possible to predict large conflict incidents ahead of time through emerging local Twitter trends. Furthermore, we want to inform media and peace scholars about contemporary international conflicts that are trending in different regions around the globe.

2 Dataset

For our content and geographical analysis, we concluded Twitter data from 27th-30th September (539 tweets), 1st-6th October (700 rows) and 1st-11th November (538 tweets). Our data set is stored as a comma-separated values (csv) file and comprises of various columns, few of them which will be used in the final outcome visualisation: timestamp, keyword search, tweets, replies, retweets, likes, username, location, friends, followers, and verified status.

Furthermore, our data is based on Twitter API search queries containing the key words 'peace', 'conflict' and 'war'. Our data was not filtered by tweet language or localisation and included both verified users such as news agencies as well as private users. We only processed tweets with given localisations, however in later cleaning stages not all of them were geocodable and thus dropped.

In order to gain larger data sets or real-time data, it is possible to retrieve more information using a professional Twitter Developer Account https://developer.twitter.com/en and a Python library for the Twitter API such as https://www.tweepy.org/.
(?, ?, p. 1023).

3 Data Collection

This section discusses future directions for data collection if larger data sets are desired for further analysis. Tweepy is a Python library for scraping Twitter data through the Twitter API. In this brief introduction, we will mention the necessary parameters to call for further analysis based on our findings.

After registration for a Twitter Developer Account, you will receive 'access token', 'acess secret', 'consumer key' and 'consumer secret'. In Jupyter Notebooks, the Tweepy package is installed via pip install Tweepy. Authen-

tication is then required following the steps describes in the official documentation: http://docs.tweepy.org/en/latest/getting_started.htmlintroduction.

The most essential attributes for a dataset are described here:

1. timestamp: the unix timestamp of each tweet
2. text: the tweet text
3. user: username
4. location: location of each tweet

If more attributes are desired, verification status, hashtags, tweet ID and followers can be called.

4 Data Cleaning

To understand the following analysis of twitter data, it is crucial to understand that tweets are short posts consisting of a maximum of 280 characters. Besides words, these messages can also contain 'Hashtags' marked with '' to refer to specific topics, usernames, URLs, images, videos, emojis, special characters, punctuation and white spaces. To remove these for proper corpora analysis, we used regular expression operations. Tweet duplicates and locations without location information were dropped, too. Detailed cleaning steps are described in below sections.

In our study project, we took several cleaning steps to make sure that our tweet content is properly cleaned:

1. Lowering cases of all strings to normalise words for further analysis, especially for tokenisation and lemmatisation
2. Removal of special characters such as 'amp' for '', 'b"', '+' via regular expression characters
3. Removal of URLs via regular expression characters
4. Removal of '' and selection of hashtag words into a new column
5. Removal of any punctuation such as '.' and '?'

4.1 Data Preprocessing

4.1.1 Word Tokenisation

The next step in our data cleaning process is the preparation of twitter data for sentiment analysis, word frequency count and other relations between the tweet corpora itself and other user metadata.

To make sense of the corpora, we tokenised each word within a tweet using nltk's word_tokenize, which means that each string is subdivided into its single word fregments. The word 'token' refers to the technical term 'for a sequence of characters ... that we want to treat as a group' (Bird, http://www.nltk.org/book/ch01.html).

This step is crucial since each tweet in our dataset consists of a string with multiple words separated by white spaces, which Python cannot distinguish before tokenisation. Tokenisation is considered one of the most important steps in data preprocessing for Natural Language Processing (An introduction to Twitter Data Analysis in Python Vivek Wisdom, Rajat Gupta) since it prepares our data for further cleaning such as removal of irrelevant words.

4.1.2 Stop-Word Removal

Since we aim at providing insights about trending tweet topics dealing with conflict and peace, we want to focus on relevant, informative words only, that might give us a hint about ongoing tweet tendencies.

For that reason, we made use of a function described as 'Stop-Word removal' which removes trivial, frequently used words such as 'and', 'or', 'to', 'what' as these words usually do not contain any relevant information or lexical context. We imported a corpus of stopwords from nltk.corpus which includes a set of above mentioned english words (https://www.nltk.org/book/ch02.html, Bird).

4.1.3 Categorising and Tagging

Next, we matched tokenised words according to their part of speech (POS). This step is necessary to retrieve the correct word stem afterwards. Based on each token's label, for example 'verb', 'noun' or 'adjective', different word stems will be matched to the token (http://www.nltk.org/book/ch05.html). For instance, the verb in the sentence 'she cares' could also be associated

with the word stem ‘car’ instead of ‘care’, if the token was not labelled as a verb beforehand.

These ‘lexical categories’ are matched using nltk pos_tag and wordnet.

4.1.4 Corpus Normalisation

In order to normalise text further and go beyond lowering all characters, we made use of nltk’s WordNetLemmatizer. This function returns the word origin of a substring which can be found in a common dictionary by removing any affixes. This process is also referred to as ‘stemming’ in the literature (https://www.nltk.org/book/ch03.html). We decided to use WordNetLemmatizer instead of PorterStemmer since it includes an extra matching step with a dictionary and is best suitable for compiling multiple tokens.

5 Data Analysis

5.1 Word Frequency

After cleaning and preprocessing our data corpus, a basic analysis of key values provides some interesting facts about our dataset. First, we looked into the frequency of words from our initial datasets provided for assignment 2 which consisted of a few hundred rows. Using small datasets first allowed us to quickly explore trends and iterate quickly with our analysis process. Noticeably, among our initial data sets, most frequent words were ‘armenia’ closely followed by ‘azerbaijan’. There was also some evidence for topics around the presidential election and the US, Yemen, COVID-19 and Turkey (see graph 1 below).

5.2 Topical Modeling with Latent Dirichlet Allocation

To test our observations made from a subset of our data as described in 5.3, we used topic modelling with our larger data set from November. Latent Dirichlet Allocation (LDA) is an unsupervised method used to identify a topic per document and to cluster specific words within that topic. The figure below shows that some of our findings were indeed represented in both the smaller test data sets and the larger data set. Due to time and computational limitations, we only run LDA on one of our data sets from November.

We decided to focus on all topics explored in 5.3 including Africa from 5.3. Moreover, the word frequency as shown in graph 2 shows clear evidence that terms around Armenia and Azerbaijan are significantly trending among conflict tweet corpora.

5.3 Topical Labeling

In order to observe the temporal change of topics and map respective tweets to their physical location, we decided to label each tweet with their topical context. For that purpose, we decided to associate each topical trend discovered in and with related words that appeared in those tweets as well (graph 3 showing words related to Azerbaijan, Armenia etc.).

In total, we included 6 subject matters related to COVID-19: the Azerbaijan-Armenia war, US, Yemen, COVID-19, Turkey and Africa. Although there was not much evidence in our initial data set that conflicts related to Africa were trending on Twitter, we wanted to include this region as a label since we found some evidence for it in 5.3 as well as in contemporary news media where various humanitarian crises were listed in recent weeks.

5.4 Color coding

Below (graph 4), you can find a color coding legend that matches labels to colors. This step was necessary for our data analysis to distinguish between labels such as US and Covid-19. Since Folium mapping does not allow us to directly associate specific colours to specific column values 'Label', we appended an additional column with with label specific colors which are then called through iteration in the Folium map application (graph 5 color coding code for further explanation).

5.5 Geocoding with geopy

Geomapping is a helpful visualisation tool to demonstrate significant regional changes and differences among opinions and events. Moreover, if combined with real-time data scraping, location data might even uncover events before they reach global news coverage.

Thus, we decided to use Python's geopy package to translate available physical addresses and city names from our dataset into latitude and longitude data to create interactive maps. In particular, we leveraged geopy's

function geocoder Nominatim to geolocate tweet origins to coordinates. Nominatim (latin for 'by name') makes use of the free wiki map OpenStreetMap to return the coordinates for every address search query. The documentation for geopy can be found here: https://pypi.org/project/geopy/.

Since geopy's Nominatim is only suitable for a limited number of requests, we also used geopy's RateLimiter to iterate through larger datasets.

For further visualisation purposes, we dropped all rows without geocodable addresses.

5.6 Geomapping with Folium and Plotly

For the last step of our project, we want to visualise our results using interactive maps over a specific time span. We decided to visualise a subset of our data for time complexity reasons. We only took into account tweet corpora that fell into one of the label categories ['Armenia_Azerbaijan', 'covid19', 'US', 'Yemen', 'Turkey', 'Africa'] with geocodable latitude and longitude values.

Since our overall goal is to explore the conclusions we can draw from frequently mentioned topical key words, we mapped the labeled tweets according to their source of origin, each with a different color as described above. We used mapping techniques from the Leaflet.js library via Folium (https://pypi.org/project/folium/) due to its interactive features, built-in tilesets with OpenStreetView and suitability with the Python Environment.

Furthermore, we also utilised Plotly's mapping functionality to animate Twitter conflict location hotspots over specific time periods.

5.7 Suggestions for Future Work

Our code for sentimental analysis and frequency count was initially implemented upon a small subset of data and can might be modified to run smoothly with larger data sets. Our code is also suitable to run a real-time analysis using Twitter's developer API to display immediate trend changes over time. Our analysis does not include any vectorisation of words using methods such as term frequency-inverse document frequency (tf-idf) or any supervised or semi-supervised machine learning yet. However, if further clustering with larger data sets is desired, we suggest using scikit's Spectral Clustering methodology. This will ensure that new data sets are trained

based on existing models to reduce biases in categorisations when relying on bigrams and word frequency count only. Moreover, another common method in topic analysis besides topic classification is topic modeling.

6 Findings

We decided to visualise our data set based on three different time periods: September, October and November. This allows us to compare Twitter hotspots and the topical change of tweet corpora.

6.1 Temporal Overview of Conflict Tweets

The first map shown below in figure 1 portrays the overview of the world map in September. Noticeably, most markers are displayed in green which refers to events related to the Armenia-Azerbaijan conflict. The map also indicated that most tweets are found in Western Europe, in particular in countries such as Germany, France and the United Kingdom and also the US, especially in the East Coast. There is also a significant amount of tweets in South Asia around India, the Middle East and a few are found in Africa.

Figure 1: Topical Overview of the Conflict Tweet Map. This figure displays the topical change across the world at different time periods in September, October and November. The legend on the right indicates the topical context of each marker color.

In comparison to September, we can observe a clear trend towards issues dealing with the US in October, as the majority of markers in the US are blue. Although the number of tweets around US conflicts are significantly increasing in Western Europe as well as shown in the second map of figure 1, conflict tweets around Armenia-Azerbaijan remain popular. Notably, there are some more COVID-19 related tweets emerging in all continents except

for South America.

Finally, we can observe that tweet topics change drastically in November. Whereas the vast majority of tweets originating from the US deals with the US itself, US conflicts are also the overall number one trend in Western Europe. Interestingly, Africa mainly shows tweets dealing with African conflicts and COVID-19 in September and October. However, in November there are equal amounts of tweets dealing with African conflicts and US conflicts, indicated in pink (African conflicts) and blue (US conflicts). The emerging trend around African conflicts is also noticeable in Western Europe and in some American states.

6.2 Topical changes in the US

To draw more conclusions from our data set, we decided to focus on specific areas of interest. We decided to analyse tweet trends in the US since the majority of Twitter data was found in that region. In the US maps figure shown below, we can track changes over three distinct time periods. However, it is crucial to know that the data is only indicative and does not represent the whole of September, October or November but only selected periods due to time complexity limitations on Jupyter Notebooks.

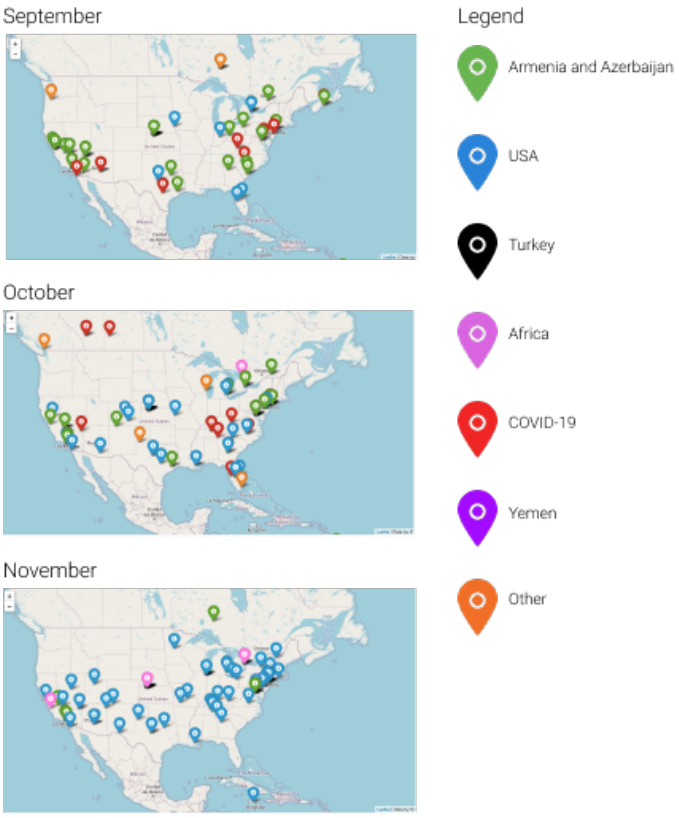

Figure 2: Tweet Context Development in the USA. This figure displays the topical change across the USA at different time periods in September, October and November.

Interestingly, the first map from September in figure 2 shows us clearly that users from the East and West Coast as well as from Texas mostly talk about events related to the Armenia-Azerbaijan conflict (green) and Covid-19 (red). Moreover, there are a few tweets related to US conflicts in Central and Eastern US.

The second map from October in figure 2 displays more tweets marked in blue, indicating that US conflicts are increasingly trending among the country, especially at the East Coast, in the South West and in Central US.

Finally, the third map in figure 2 shows us that tweets regarding US conflicts (blue) have taken over the whole country, with only a little number of tweets still addressing conflicts in Armenia and Azerbaijan (green) as well as Africa (pink).

6.3 Topical changes in Western Europe

Another Twitter hotspot was found among Western and Central European countries such as France, Germany and the United Kingdom. As shown in figure 3 below, most countries' tweets demonstrate a relation to the Armenia-Azerbaijan conflict (green) in September.

Figure 3: Tweet Context Development in Western Europe. This figure displays the topical change across Western Europe at different time periods in September, October and November. Legend: Green = Armenia and Azerbaijan, blue = USA, black = Turkey, pink = Africa, red = COVID-19, purple = Yemen, orange = other.

In October, subject matters evolve in terms of conflict related tweets. Especially in Belgium, Germany, the United Kingdom, Poland, Austria and Slovenia, Twitter users tweeted about conflicts related to the US (blue).

Similar to the US, topical trends change drastically in the month of November. While Eastern European countries still deal with conflicts around Armenia and Azerbaijan (green), Western European countries report mainly about US conflicts (blue) and conflicts related to Africa (pink).

6.4 Topical changes in Africa

Although there were only a small number of tweets detectable among African countries, the topical change over time is still noticeable as shown in figure 4. While tweets mainly address Africa related conflicts (pink), COVID-19 (red) and the Armenia-Azerbaijan conflict (green) in September, trends change in October. While no tweets related to Armenia and Azerbaijan are detectable, topics around Africa increase and more US conflict tweets emerge (blue).

Figure 4: Tweet Context Development in Africa. This figure displays the topical change across Africa at different time periods in September, October and November. Legend: Green = Armenia and Azerbaijan, blue = USA, black = Turkey, pink = Africa, red = COVID-19, purple = Yemen, orange = other.

Interestingly, the same trends in US and Western Europe apply to the African continent in the month of November: Africa and US related conflict tweets increase especially in Central African countries such as Kenya, Ethiopia South Sudan, Kamerun and Nigeria.

6.5 News Coverage and Related Events

Through initial analyses of small data sets and using the LDA method with more data, we identified a set of conflict issues that we want to follow over three specific time periods in September, October and November. The two major events that were addressed by Twitter users the most among our analysis were conflicts dealing with terms such as Armenia, Azerbaijan and the US. In order to understand why these issues emerged over time and why the trends might change, we have briefly analysed official news sources that have reported about conflict events during relevant periods of time. Notably, we only looked at English speaking media agencies which might be biased towards specific global events of interest.

6.5.1 Armenia and Azerbaijan Conflict Timeline

On 27th September, it was reported that the ongoing conflict between Armenia and Azerbaijan broke out once again (`https://dr.ntu.edu.sg/bitstream/10356/144825/2/CO20188.pdf`). The fighting included both civilian and military agents. Most of the incidents have taken place along the 200 km battle line marking the Nagorno-Karabakh region as displayed in figure 5. On 28th September, Azerbaijan announced further military operations. (`https://www.crisisgroup.org/content/nagorno-karabakh-conflict-visual-explainer`).

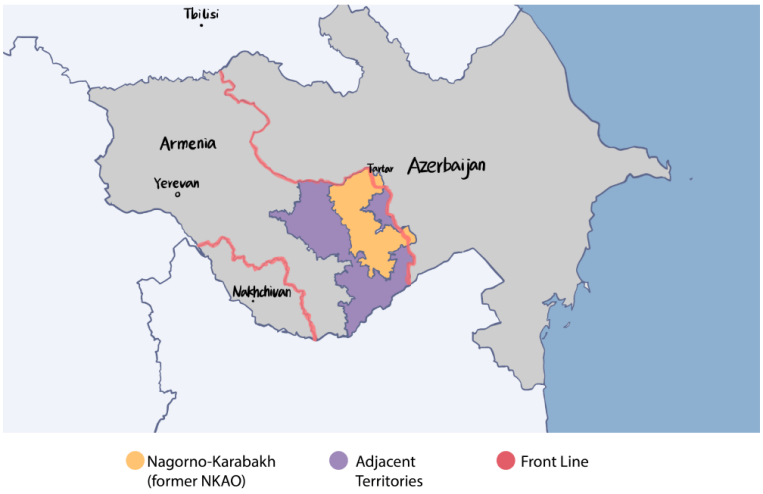

Figure 5: The conflict zone Nagorno-Karabakh. This figure illustrates the disputed Nagorno-Karabakh region and the conflicted front line. Content is adopted from https://www.crisisgroup.org/content/nagorno-karabakh-conflict-visual-explainer

It has also been suggested, that Turkey plays a big role in the war due to its strong economic and cultural ties with Azerbaijan (Infromation retrieved from https://www.jstor.org/stable/resrep26446?seq=1metadata_info_tab_contents,Erdogan Seeks to Upend Kremlin-Backed Status Quo in Nagorno-Karabakh). 6 weeks later, Russia demanded a ceasefire to stop the mass deaths for the time being. It is the longest conflict in the former Soviet Union and has its beginnings in 1988, when ethnic Armenians called for the original Nargorno-Karabakh Autonomous Oblast (NKAO) to be granted to Armenia. The war between 1992 and 1994 claimed many victims and many had to flee their homes. There are still ongoing conflicts today as displayed above ((https://www.crisisgroup.org/content/nagorno-karabakh-conflict-visual-explainer).

6.5.2 US Conflicts

The largest political event in the US in November was the 46th presidential election, held on 3rd November 2020. Competing candidates were the current US president Donald Trump from the Republican Party and Joe Biden from the Democratic Party. This year's presidential election went down in history for various reasons, including the prevailing COVID-19 pandemic, which permitted the postal vote (Mass casualty event scenarios and political shifts: 2020 election outcomes and the U.S. COVID-19 pandemic). The election has been controversial due to Trump's denial of the election outcome and further misconceptions and false information about alleged election fraud.

More background information on ongoing debated and conflicts about analysed tweet corpora can be found on our website or on crisesgroup.org.

Future directions

Here we introduce some vectorisation and Term Frequency/Inverse Document Frequency (TD/IDF) for further processing like using for support vector machine model by training data with a unigram approach. If we have bigger dataset and strong computation, some machine learning can be introduced.

Also we can use k-nearest neighbors algorithm (KNN) Classification. The way that the classification algorithm will work is that for a given tweet in the test dataset (d), we will compute Euclidean distance between d and every sample in the training dataset (D). We will then choose k samples that are nearest to d, i.e. those samples which have the smallest distances from d. From among these k samples, we will extract out the class that is assigned to a majority of the samples and assign that label to the test instance. F1 Score is a common metric used for the evaluation of Natural Language based tasks. It is often said to be 'The Harmonic Mean of Precision and Recall', or conveys the balance between precision and recall. It expresses the balance between the precision and the recall. As accuracy only gives the percentage of correct results of the model but does not show how adept the model is at finding true positive results, both measures have merit, depending on the need.

How to find the credibility of tweets? The number of retweets, likes and replies is also useful information.

With the help of natural language processing (NLP), machines are able to break down human language and analyze it for powerful insights. Topic modeling a text analysis technique that uses unsupervised machine learning to process text by topic or subject and cluster and categorize similar words and phrases, without the need to create defined tags (categories) ahead of time. With topic classification or topic extraction, on the other hand, you must pre-define the tags, use them to train a classification model, and the model will then categorize your texts (tweets) into the tags you set up.

References