Narrative

Kanika S Sisodia

11/24/2019

Brief substantive background / goal

On August 5th, 2019 the BJP government in India revoked Article 370 of the Indian Constitution. Thereby, revoking the special status which was given to the state in 1947. December 8th, 2019 marks the 125th day since Kashmir has been in lockdown and communications blackout imposed to prevent protests during which thousands of people, mostly young men, have been detained in Jammu and Kashmir (the India-administered portion of the disputed territory of Kashmir). There has been an exponential surge in the online activity following the government's. decision especially on social-media where people have both supported the government's decision to abrogate Article 370 while others have called out to end atrocities and return the valley to normalcy. For my final project submission I decided to perform sentiment analysis using one of the social-media platforms. I decided to scrape data from the Twitter website, mainly because of the format of the platform. Twitter employs a message size restriction of 280 characters or less compelling most users to stay focused on the message they wish to communicate. This very characteristic of Twitter makes it so interesting, and an ideal platform to pull data from for Sentiment Analysis, especially about recent political events happening all over the world.

For this project, I scraped approximate 10,000 tweets on December 8th, 2019. I will demonstrate how to analyze sentiment score based on what people tweeted about Kashmir on that particular day. My code is divided into following parts; 1.Extracting tweets using Twitter application. 2.Cleaning the tweets for further analysis. 3.Plotting word frequencies and Word Cloud. 4.Sentiment Analysis

Main Challenges and Limitations

The main challenge that I faced while working on this project was to scale down on the data sets that I had created after scraping tweets from the Twitter website.

My initial proposal was to scrape tweets from 18 twitter handles (nine accounts of people from the current BJP government and nine accounts of people that I randomly selected, who were tweeting against the abrogation of Article 370) and store the tweets in two datasets. I scraped tweets from November 3rd up until November 20th and saved them into .csv files. The idea was to merge the data sets, clean and process the tweets and perform Sentiment analysis from the merged data set. The second thing that I wanted to accomplish was to translate Hindi tweets and incorporate them into the data as well. This I had planned to do using either a Hindi stemmer or Google Translate. The main problem that I encountered was the data set I had collected was too large for me to analyse. Each .csv file contained 26,000 tweets and I had difficulty in processing the data in R. I tried to increase the RAM size and downloaded the quanteda package to help me process the large data. Unfortunately, because of lack of time and my inexperience with working with large data sets I could not work on the data I had collected. Lastly, I could not find a good stemmer for Hindi which would help with processing of the tweets in Hindi and again because of time constraints I could not fully explore Google translate.

Solutions

For this project I scraped 10,000 tweets in a day and imported the tweets into a dataframe and converted it into a corpus, and thereafter into a Document Term Matrix. After that I created a function (Textprocessing) to clean the tweets, such as removing mentions, hashtags, urls and whitespace. After cleaning the data I

ordered the frequencies to list the most common words and plotted the ten most used words in a barplot. I also created a wordcloud of the 150 most used words.

For the sentiment analysis, I used the 'nrc' dictionary which has eight different emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and their corresponding positive and negative valence from the nrc dictionary. The get_nrc sentiment function from the package syuzhet compares all the tokenized words with the word sentient EmoLex which contain a large number of words with different emotions. I created a second barplot mapping the sentiment score from the Kashmir.df.

Results

My conclusion after performing a sentiment analysis on tweets I collected on December 8th, is that even after 125 days of Kashmir being put under lockdown, the people are largely supporting the decision of the government on Twitter. The conclusions I drew were that the people trusted the government with their decision and were positive about the centre revoking the special status of Kashmir.

Future Work

I plan to write my Masters thesis on the current crisis in Kashmir. While working on this project I have realised that it's always better to work on small data sets and work up from there. And, if faced with the challenge of working with large data sets on R I would be looking at alternative approaches like Google Cloud and Amazon web services, so that I'm not hindered by RAM size and other obstacles that I faced while working on this project.