

Captioning the Discourse : A Proposed Approach to Detect Topics in Tweets concerning the Nirbhaya Rape Case 2012

Background:

In December of 2012, a woman was brutally gang raped and murdered in Delhi, India. The incident shook the country and sent waves of panic among the citizens. As a result, a discourse arose in social media with some demanding justice while others debating on women's safety issues and yet others blaming the government.

Aim:

The aim of this project is to identify the various topics that emerged from the discourse that unfolded on X (prev. Twitter) from the date of the incident to one year later.

Data:

The data was collected from X (prev. Twitter) using the following query : "nirbhaya rape case 2012 until:2014-02-13 since:2013-02-13". Since X has banned the free extraction of data from its website, I had to use an external API twitterapi.io which gave me a limited number of tweets for free.

For the cleaning and pre-processing I used Regex to remove extra spaces, emojis, usernames, links, punctuations and dates and numbers. For the removal of stop words, I used Python's NLTK library. I also removed duplicate tweets which the API had fetched.

Next I visualized the data using a wordcloud.

After basic cleaning, I moved to vector representation of the text. I used the tf-idf method with the help of the TfidfVectorizer class from the ScikitLearn library.

Algorithm:

For the topic modelling algorithm, I had the choice between LDA and a simple KMeans clustering. Due to limited availability of data and the short average length of sentences I went with a KMeans clustering. I first plotted the WCSS curve (or the elbow curve) to determine the optimum number of clusters and finally ran the KMeans model. The output was plotted using Matplotlib and top words for each cluster was figured out.

Limitation:

Unavailability of data is our main limitation. Since the dataset is very small (only 52 rows), the optimum number of clusters is 50, according to the Elbow Curve. Therefore, this is just a proposed approach to modelling such a dataset.

Future Directions:

I intend to complete this work with more data and find out the different topics of discussion that come up during such an incident. Furthermore, I want to use this to develop a model for automatic detection of victim blaming during rape cases.

