Spatial Hotspot Identification and Sentiment Strength Visualisation (SI – SSV) – A Twitter based View Into India's CAA & NRC Debate

Durga Toshniwal[1], Soumya Somani[2], Shruti Patil[3]

[1] Indian Institute of Technology Roorkee, Uttarkhand, India
durgafec@iitr.ac.in
[2] Symbiosis Institute of Technology, Symbiosis International University, Pune, India
soumya.somani@sitpune.edu.in, shruti.patil@sitpune.edu.in

Keywords: Sentiment Analysis, LSTM, RNN, Neural Network, Geocoding, Post Event Analysis

Abstract

Our everyday life has always been influenced by the thought processes of other people. Their ideas and opinions have a great effect on our own opinions and the decisions that we make. The explosion of the internet has led to great increase in Social Networking and microblogging. As a result, a large amount of data is being generated. There has been a lot of interest for mining these vast repositories of data for opinions. This field is called Sentiment Analysis (SA). Thus, SA refers to the use of computational techniques for determining opinions, sentiments and subjectivity of text resources.

In this paper, we have tried to analyze the public sentiment towards the recent CAA & NRC acts using Twitter data. An analysis of this data can reveal the prevalent public sentiments about these topics. A location based landscape has also been generated.

1 Introduction

With the emergence of popular social networking sites like Facebook, Twitter colossal amounts of public data is being generated which can be collected, processed and used to perform analyses to solve a vast variety of problems. Twitter has millions of users who share their opinion regarding a wide range of topics, making it a valuable platform for analyzing the sentiment regarding the topic being tweeted on. This analysis can provide insights and help in decision making in various domains. Due to the ease of availability of twitter data and possible applications of twitter sentiment analysis a lot of research is underway in applying this knowledge to various domains effectively.

The Citizenship Amendment Act (CAA) was passed on 11 December 2019 and its implementation began on 20 December 2019. This act along with National Register of Citizenship (NRC) faced criticism for discriminating on the basis of religion and led to nationwide protests which led to police intervention. This generated widespread outrage and led to people voicing out their opinion for/against these topics on social media. Thus, a large amount of data was generated.

An analysis of this data could reveal the prevalent public sentiments about these topics. Location based hotspots of the sentiments could also be generated to determine the parts of the country (and also the world) in support of or against these acts. This information could be very vital to the government.

In this paper, tweets pertaining to the topic have been collected and a sentiment based, location segmented view (SI – SSV) has been presented.

The rest of this paper is organized as follows: Section II presents the literature review. Section III gives the details of the proposed methodology. Section IV comprises of the dataset details, experiments performed and the results. Section V gives the conclusion and future scope of this paper. The last section gives the references.

2 Literature Review

In this section, some of the existing research works in this field have been briefly presented.

Not much work has been done for the analysis of twitter data for post-event analysis to present a global sentiment strength visualistion of the event. Post event analysis works thus far have focused on natural disasters in a specific country. A research by Tae H. Kim, et al [2005] focuses on the 2003 Southern California wildfires.

Research by Alves, et al [2016] provides a spatial sentiment analysis approach about an event in Brazil but is limited to Potuguese tweets from the country of Brazil. Another approach is given by Zhang and Gelernter [2014] is limited to geocoding location expressions in Tweets.

Research on the topics of the CAA & NRC has been focussed on examining the provisions of the act against the centizenship provisions of the Indian Constitution. These works discuss the act itself against the backdrop of the Indian Constitution and do not provide a view of the sentiments of the public towards the act.

Research work by Chandrachud [2020] argues about the legality of the Act in the present conditions. Another work by Jayaram and Rahul [2020] focusses on the protests and the government's retaliation against the protests in certain parts of the country like Uttar Pradesh. It does not include the sentiments of the people in support of the act.

Based on this detailed literature survey, it can be observed that the existing research on spatial sentiment analysis is limited to the scope of a particular nation. Furthermore, most of the existing reasearch in the field of CAA & NRC focusses on the legality of the act and its opposition. There is a lack of research that provides a consolidated spatial sentiment based view encompassing al the sentiments (positive, negative or neutral).

This paper aims to address these limitations of the existing works.

3 Proposed Work

In order to identify the sentiment hotspots and generate a visualisation of the sentiment strengths the following methodology has been proposed.

3.1 Proposed Framework

The proposed framework for this reseach work is illustrated in Fig. 1.

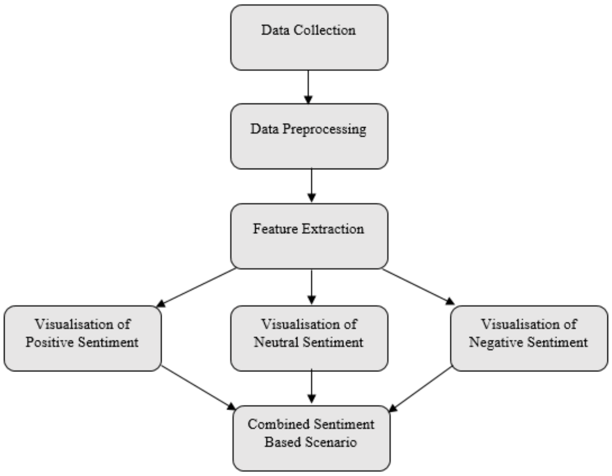
Data Collection
Data Preprocessing
Feature Extraction
Visualisation of Positive Sentiment
Visualisation of Neutral Sentiment
Visualisation of Negative Sentiment
Combined Sentiment Based Scenario

Fig. 1: Proposed Framework.

Data Collection

The data was collected from the twitter live stream using the twitter API. It was done over a period of 4 weeks from 22 December 2019 to 18 January 2020. The tweepy library for Python allows us to connect to the twitter live stream via the twitter API and open a pipeline to deliver selected data to us.

Different keywords and hashtags pertaining to the topic were used for the dataset collection. To have maximum search coverage using hashtags, top trending hashtags on the topic were added to the search keywords every day.

About 12 million tweets were collected to form the corpus.

Data Biases

Only English language tweets were collected. This was done to avoid noisy data that would be generated by the translation of non English tweets. Thus, this work does not reflect any expression in non english languages. No thresholding was applied to the volume of tweets from any location.

Data Preprocessing

Not every tweet has the location attribute enabled. For tweets not having the location attribute, the user profile location (if available) was used. This was done since the profile location is same as the tweet location for most of the tweets. This way, the location tagged tweets are filtered out for further analysis. Tweets having null or garbage values as their location values were dropped. The tweets were futher processed using Google's Geo-coding API to geolocate the exact location from where the tweets were posted. The corpus was then segmented based on the tweet locations.

The tweet text was converted to lowercase. All non – Unicode characters, hexal characters, punctuation marks, new lines, URLs, usernames, stop words and hastags were removed from the tweet text. Noisy tweets i.e tweets containing non english words, very few words etc were filtered out.

Feature Extraction and Model Construction

Feature extraction is an extremely essential task in any machine learning approach. The first step in any approach for determining sentiments is the conversion of the given text to a feature vector.

Word Embedding refers to a set of language modelling or feature engineering techniques that are used to map the words in a vocabulary to vectors of real numbers. They are used to represent the words in the inputs of various NLP (Natural Language Processing) tasks such as sentiment analysis.

In this work, the 200 dimensional GloVe (Global Vectors for Word Representation) word embeddings provided by Stanford we used. GloVe is an "off the shelf" word embedding which has been trained on word to word co-occurrence statistics.

A LSTM Model was created and trained for sentiment based text classification. The standard "Sentiment-140" corpus of labelled tweets from kaggle was used as the training dataset. The 200 dimensional GloVe twitter word embedding was used for the purpose of embedding weights.

The training dataset consists of 1.6 million tweets which have been annotated as 0 (negative), 2 (neutral) and 4 (positive) for the purpose of detecting sentiments. The model was trained over 10 epochs achieving an accuracy of 80.13% over the validation data.

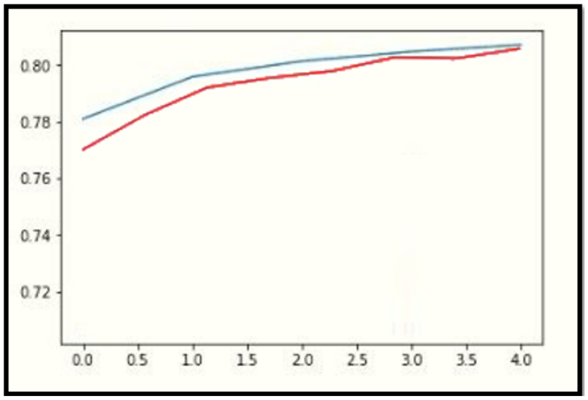

Fig. 2: Training and validation accuracy of the model.

The model was used to predict the sentiments of the tweets in the corpus. The tweets were then segmented based on the tweet location and its sentiment. This way, a count of tweets of all sentiments from each location was generated. The sentiment strengths for each location were calculated as:

$$I_{l,s} = \frac{T_{l,s}}{T_{l,1} + T_{l,0} + T_{l,-1}}$$

Where, $I_{l,\,s}$ is the strength index of sentiment s from location l, $T_{l,\,s}$ is the tweet count of tweets having sentiment s from location l, $T_{l,\,1}$ is the tweet count of tweets from location l having positive sentiment, $T_{l,\,0}$ is the tweet count of tweets from location l having neutral sentiment and $T_{l,\,-1}$ is the tweet count of tweets from location l having negative sentiment.

The sentiment strengths were used for hotspot identification i.e. locations having the highest value of strength index were marked more prominently in the visualization map. This was a Spatial Hotspot Identification and Sentiment Strength Visualisation (SI – SSV) was achieved.

4 Experiments and Discussion

The following section contains the dataset description and presents the results obtained from this work.

4.1 Dataset Description

The data was collected from the twitter live stream using the twitter API. It was done over a period of 4 weeks from 22 December 2019 to 18 January 2020. About 12 million tweets were collected to form the corpus.

Sr. No.	Attributes	Values
1	Duration	December 2019 – January 2020
2	Total Tweets Collected	12,813,038
3	Total unique Tweets (with locations)	1,663,100
4	Total Number of Locations	10,178

Table 1: Dataset Description

4.2 Discussion

This section presents the experiments and results of this project. These results give us a location based sentiment strength visualisation both on a global as well as national (India) level.

Visualisation of Positive Sentiment

The given figure (figure 3) presents the global landscape for the positive sentiment i.e. the tweets in suppprt of CAA & NRC.

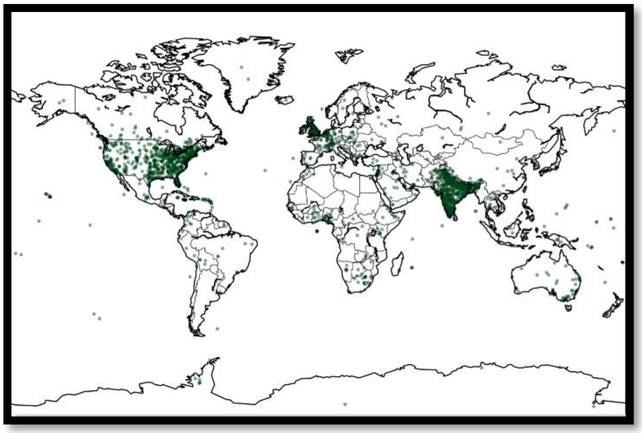

Fig 3: GLOBAL visualisation (positive sentiment)

Tweets in support of CAA & NRC were received from various nations (apart from India) around the globe such as the USA, UK, Spain, France, Germany and Belgium to name a few.

Three nations comprised the majority of tweets. They were the United States of America (USA), the United Kingdom (UK) and India.

Sr. No.	Nation	No. of Tweets (positive)
1	USA	81,018
2	UK	14,992
3	India	3,45,381

Table 2 GLOBAL visualisation (positive sentiment)

A national landscape (figure 4) was also generated. It depicts the major locations (within the nation) in support of the given topic.

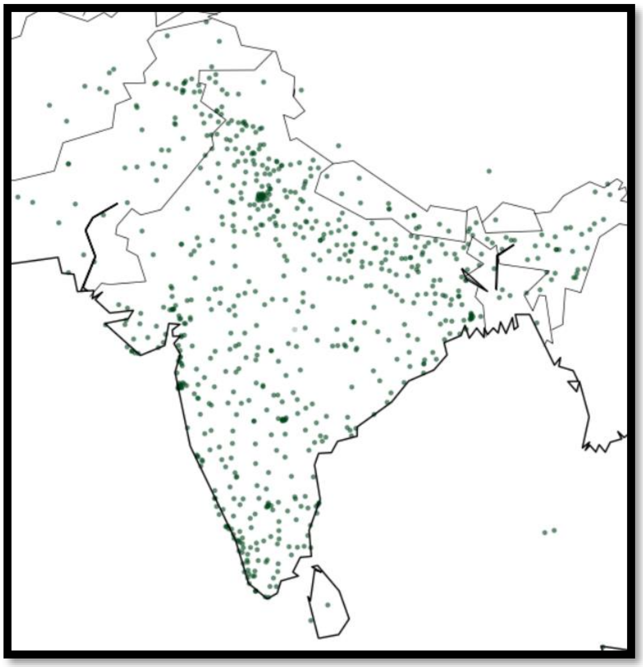

Fig 4: NATIONAL visualisation (positive sentiment)

Sr. No.	City	No. of Tweets (positive)
1	New Delhi	80,852
2	Mumbai	42,771
3	Bengaluru	19,664
4	Hyderabad	12,780
5	Chennai	10,921
6	Kolkata	8,112

Table 3: Major cities (positive sentiment)

The largest positive support hotspots (for CAA & NRC) in India are the cities New Delhi, Mumbai, Hyderabad, Bengaluru, Chennai and Kolkata.

The size and intensity of the colour of the location markers are based on how strong the sentiment at that location is. Thus, these landscapes also provide us with a visual representation of sentiment strength.

Visualisation of Neutral Sentiment

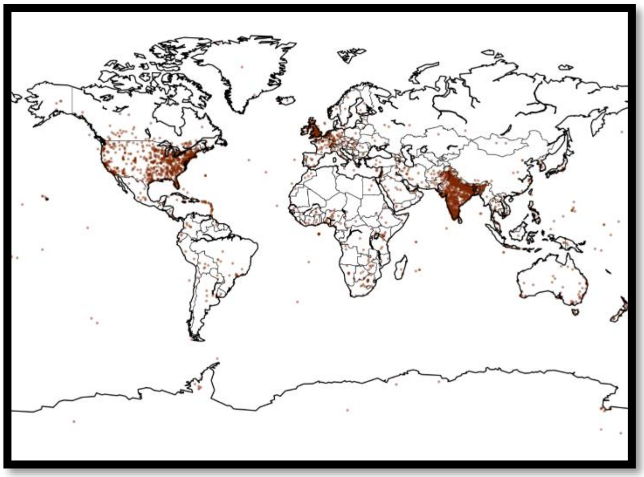

Fig 5: GLOBAL visualisation (neutral sentiment)

The above figure depicts the global landscape for the neutral sentiment i.e. tweets that cannot be classified either as positive or negative.

These tweets were received from various nations around the globe. Some of them are the USA, UK, Italy, France, Germany, the Netherlands and Belgium in addition to India. As with the positive Tweets, predominantly three nations namely the USA, UK and India comprised the majority of the tweets.

Sr. No.	Nation	No. of Tweets (neutral)
1	USA	1,05,228
2	UK	19,592
3	India	7,08,050

Table 4: GLOBAL visualisation (neutral sentiment)

A national landscape (figure 6) was also generated. It depicts the major locations with neutral sentiments.

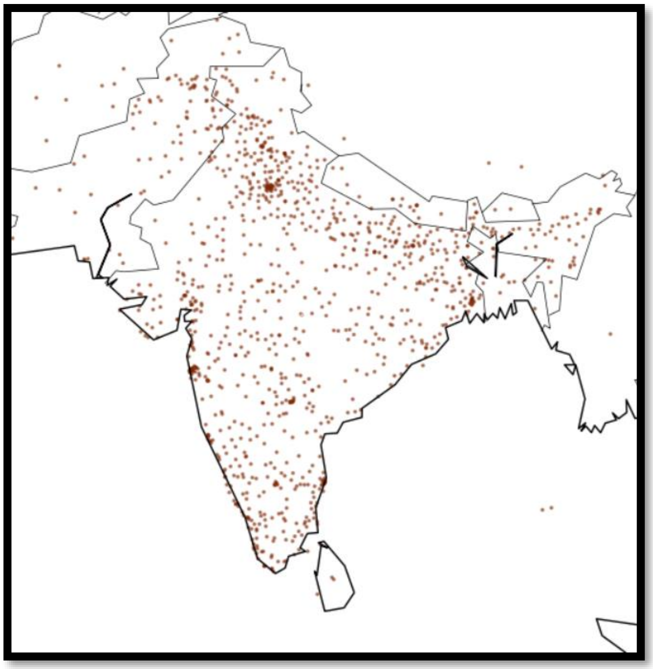

Fig 6: NATIONAL visualisation (neutral sentiment)

Sr. No.	City	No. of Tweets (neutral)
1	New Delhi	1,60,891
2	Mumbai	82,171
3	Bengaluru	33,222
4	Hyderabad	30,040
5	Chennai	24,042
6	Kolkata	17,254

Table 5: Major cities (neutral sentiment)

The cities New Delhi, Mumbai, Hyderabad, Bengaluru, Chennai and Kolkata were once again the locations that comprised the majority of the tweets (approximately 50%) from India.

Visualisation of Negative Sentiment

The figure given below (figure 7) depicts the global landscape for the negative sentiment i.e. tweets that oppose the CAA & NRC.

Tweets opposing the CAA & NRC were received from various nations like the USA, UK, Spain, Italy, France, Germany and Belgium in addition to India.

Similar to the previous landscapes, three nations namely the USA, UK and India comprised the majority of the negative sentiment.

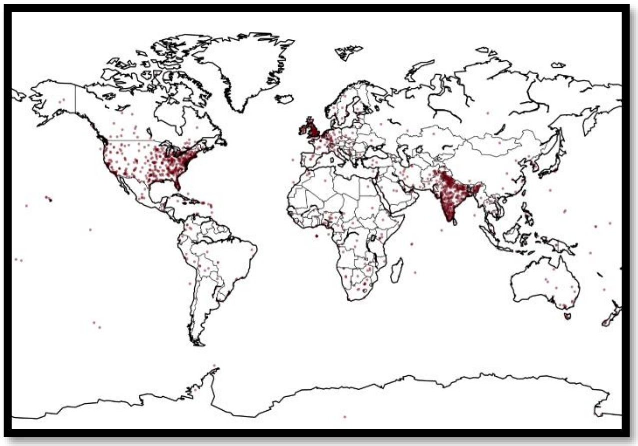

Fig 7: GLOBAL visualisation (negative sentiment)

Sr. No.	Nation	No. of Tweets (negative)
1	USA	40,191
2	UK	8,631
3	India	2,00,198

Table 6: GLOBAL visualisation (negative sentiment)

A national landscape (figure 8) was also generated. It depicts the major locations opposing the topic.

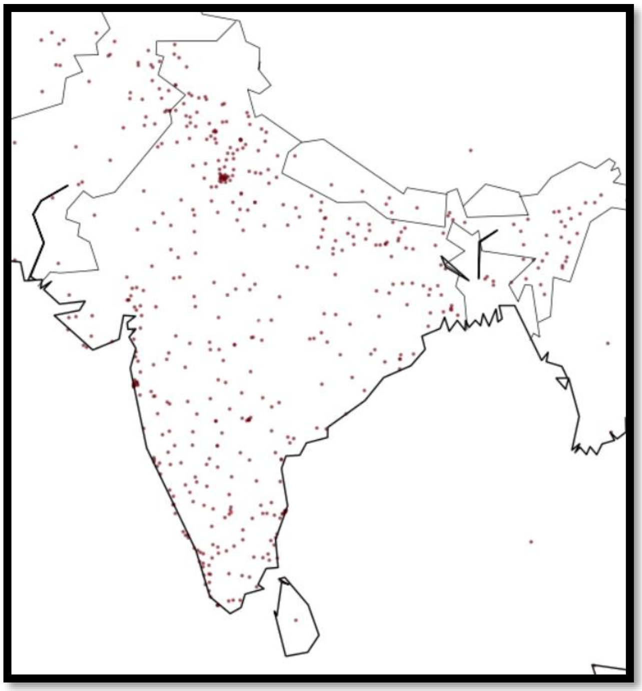

Fig 8: NATIONAL visualisation (negative sentiment)

Sr. No.	City	No. of Tweets (negative)
1	New Delhi	48,961
2	Mumbai	27,101
3	Hyderabad	9,802

Table 7: Major cities (negative sentiment)

The cities New Delhi, Mumbai and Hyderabad are the major locations with a large negative sentiment i.e. against CAA & NRC.

Bengaluru, Chennai and Kolkata are the cities that despite having large positive and neutral sentiments do not have any sizeable negative sentiment.

Overall, the negative sentiment towards CAA & NRC is much lower than the positive and neutral sentiments. This is also evident from the visual landscape where the location markers in the negative landscape are much smaller and lighter than the positive and neutral landscapes.

Combined Sentiment Based Scenario

This section presents a holistic picture of the Tweet sentiment distribution over all the major locations (globally & nationally) as follows:

Sr. No.	Nation	Sentiment (%)		
		Positive	Neutral	Negative
1	USA	35.77	46.47	17.74
2	UK	34.69	45.33	19.97
3	India	27.55	56.48	15.96

Table 8: GLOBAL combined view

Sr. No.	City	Sentiment (%)		
		Positive	Neutral	Negative
1	New Delhi	27.8	55.34	16.8
2	Mumbai	28.13	54.04	17.82
3	Bengaluru	33.19	58.08	8.68
4	Hyderabad	24.28	57.08	18.62
5	Chennai	32.33	57.98	9.67
6	Kolkata	30.61	60.74	8.63

Table 9: NATIONAL combined view

5 Conclusion and Future Work

In this work, we built a dataset of CAA & NRC related twitter data, developed a LSTM network for determining sentiments of tweets and performed geocoding for presenting Spatial Hotspot Identification and Sentiment Strength Visualisation (SI – SSV).

We used the GloVe word embedding weights for feature extraction. We geocoded the tweet locstionas and performed SA on our corpus. Then, the corpus was segmented based on the tweet sentiments and locations to get the sentiment distribution. The sentiment strength at each location was calculated.

Next, Sentiment Strength Visulisations were generated. A combined sentiment based scenario of the distribution at the national level as well as the global level was also presented. This analysis has revealed that the majority sentiments towards CAA & NRC are neutral or positive, with only a small negative sentiment (around 10 – 20%) at most locations.

In this work, the emphasis has been placed on spatial analysis. Also, only the English language tweets have been considered for analysis to prevent the added noise that would be generated due to language translation.

- Spatial Clustering of the locations can be performed to obtain clusters of locations having a similar sentiment distribution. Thus, clusters of locations with similar sentiments can be identified.
- Temporal Clustering (day wise and week wise) can be done on the dataset to obtain days / weeks with similar sentiment distributions over various locations. A temporal view can also be generated showing the day wise change in sentiments.

This can be the future scope of the current work.

6 References

Tae H. Kim, Thomas J. Cova and Andrea Brunelle. Exploratory Map Animation for Post-Event Analysis of Wildfire Protective Action Recommendations. In *Natural Hazards Review, Volume 7 Issue 1 (2006)* .

Andre Luiz Firmino Alves, Claudio de Souza Baptisa, Anderson Almeida Firmino, Maxwell Guimaraes de Oliviera and Anselmo Cardoso de Paiva. A Spatial and Temporal Sentiment Analysis Approach Applied to Twitter Microtexts. In *Journal of Information & Data Management, Volume 6 Issue 2 (2016)*.

Wei Zhang and Judith Gelernter. Geocoding location Expressions in Twitter messages: A preference Learning Method. In *Journal of Spatial Information Science, No 9 (2014).*

Abhinav Chandrachud. Secularism and the Citizenship Amendment Act. In *Social Science Research Netowrk (2020)*.

Rahul Jayaram. CAA - NRC: Citizenship or Nativism?. In *Deccan Herald (2020)*.

D. Williams and G. Hinton. Learning Representations by Back-Propagating Errors. In *Nature, vol. 323, no. 6088, pp. 533–538, (1986)*.

P. J. Werbos. Backpropagation Through Time: What It Does and How to Do It. *Proceedings of the IEEE (1990).*

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural computation (1997).*