

DA_Project_Conference_Paper_ 175_009_188

by Abhay Warriar

Submission date: 16-Nov-2022 03:11PM (UTC+0530)

Submission ID: 1955678961

File name: DA_Project_Conference_Paper_175_009_188.pdf (1.79M)

Word count: 2089

Character count: 11007

Twitter Sentiment Analysis of the Citizenship Amendment Act

¹ Kuval Kush Garg
PES2UG20CS175
CSE Department, PES University
Bangalore, India
kuvalkarg@gmail.com

¹ Abhay R Warriar
PES2UG20CS009
CSE Department, PES University
Bangalore India
abhaywarrier28@gmail.com

¹ Mayank Agrawal
PES2UG20CS188
CSE Department, PES University
Bangalore, India
mayankagrawal008@gmail.com

Abstract—The Citizenship Amendment Act (CAA) of 2019 caused a lot of public controversy when it was proposed. It allowed Indian citizenship to various religious minorities from the neighboring Muslim majority communities. However, this caused widespread protests across India for discrimination on the basis of religion. This project aims to analyze the sentiments of the public before a Bill is passed as an Act, using multiple Machine Learning models. Our prediction of the percentage of people who were for CAA coincides almost exactly with the results of many studies and surveys conducted by new outlets and organizations. Twitter API was used for the collection of tweets to conduct Sentiment Analysis and hence categorize the tweets as pro or against the CAA. Random Forest Classifier obtained the highest accuracy of all models.

I. INTRODUCTION

Bills, before being passed as Acts, may be scrutinized by the public. The feelings of the general public must be considered before passing an Act. Sentiment Analysis refers to the analysis of how a group of people react/respond to a product, service or in this case, a government policy. The CAA was not a very popular Act and gathered a lot of negative responses to it. It was important for the Indian government to consider the sentiments of the public and the backlash it caused, before passing this law. Our solution makes use of the Twitter API to gather tweets using appropriate keywords. We had to create a Developer account on Twitter for access to tweets which were then filtered out with keywords pertaining to CAA and India. Endpoints were then updated based on our permissions on the Twitter Develop portal. A JSON file containing the responses was then dumped and converted into a CSV file. We needed to make sure that the API was not spammed with requests. The duplicates were then dropped from the dataset. Training and Testing data was then created using the 80/20 split rule. This was followed by pre-processing, tokenization and stemming. The usernames, special characters and retweet values were then dropped and text vectorization and classification using multiple models was employed. A submission csv file was then created with the labelled predictions of the sentiments of the test data. The performance of various models is then compared. This then concludes whether the law is positively or negatively received.

II. PREVIOUS WORK

In the previous research papers and solutions, the dataset cleaned the data which was retweeted. However, this is an integral part of the dataset as it considers how members of the general mass feel about other people's opinions and ideas. Also, using an ensemble model makes more sense as it would compositely average out the best performing model and give the best result, which has not been implemented in previous works. The dataset was created from scratch for this

project, and based on the limited Twitter Developer access we were given, our training and testing dataset was limited in volume. Also, many common people as well as experts in the domain were asked to label the training data for classification. Assumptions made about the data were that only Indians were tweeting about the CAA, although this could be fine tuned by filtering out the geotags in our GET request to the Twitter API and including a parameter which includes the mention of only Indians being considered for this dataset. The hyperlinks used by users on their tweets were also ignored, although these may indicate important feelings about the user's standpoint, so this needs to be rectified in the future. The accuracy of the previous solutions was well below 85% and this is something we have improved drastically by achieving accuracy scores of 91% on many runs.

III. PROPOSED SOLUTION

Our solution consists of many sub-solutions, which are described in detail, ahead.

A Twitter Developer account was created with Elevated Access and the unique token generated was used to connect to the API through Jupyter Notebook. It was then set as an environment variable on our system and bearer tokens for the authorization were parsed so that it may be accessed by the Twitter API.

A function containing various parameters such as query, start time, end time, max results, etc. were then passed. These parameters are self-explanatory and were defined by us based on the number of tweets we require, the start date and end date of our parsing request and the maximum number of requests we wish to make per request so that we avoid spamming the API with GET requests.

We then connect to the endpoint based on our Twitter Developer permissions and a response code of 200, which indicates success was obtained. We made sure to set our keyword parameters to CAA and India so that only tweets containing these keywords are returned to us for our dataset.

The returned JSON file containing all relevant data and meta data related to our query was then dumped into our output file as is, with no processing done in this stage. This JSON file was then processed and converted into a CSV file.

Another function was created to obtain more tweets from all search pages and flags were setup to make sure that the maximum number of requests was not exceeded. A connection request was established and the next token was sent as a parameter. The CSV file was appended for each request loop and the duplicate tuples were dropped.

Following this, we divided our dataset into training and test data based on the standard 80/20 data split since we

cannot perform training on all of our data. Many open-source programs, manual labelling through friends and domain experts, etc. were employed for the labelling of our training dataset. Retweeted tweets were not dropped since they hold valuable information about how users feel about other users' opinions.

A ZERO represents tweets which are for the CAA, while a ONE indicates Tweets which are against the CAA. The value counts function was used to visualize the number of ZERO's and ONE's in the training data.

Username, special characters and punctuations were removed from the tweets and a new column containing the processed tweets was created. The RT tag which indicates a retweet was also removed as it holds no value in our analysis.

A Counter was used to keep track of the most common commonly occurring words in the dataset. A collection was created where elements are stored as dictionary keys and their counts are stored as values.

Porter Stemmer was then imported so that the process of stemming could be efficiently carried out. Words like comfort, comforting, comfy, etc were stemmed into a single word for ease of analysis and clean data.

Stop words were initialised in a list so that the most commonly occurring words which do not tilt the sentiments to either side of the classification (neutral words) were eliminated. Manual addition of other stop words was also done to completely get rid of all redundant words. Keeping in mind the Twitter lingo that is popular nowadays.

Punctuation of the words was checked and if it existed, it was removed and all other stop words were finalised, which made it past the previous stop word filter. This entire process was carried out on multiple loop runs and appended into the new tweet column of our dataset.

The entire process was done separately for the training and testing data so as to keep them independent from each other and eliminating bias and correlation.

Wordclouds were created to visualize the term frequency of the document and apply the function process to each word parsed through the loop. The list structuring and commas were removed so normal sentences were created in the new tweet column. Wordclouds for positively and negatively labelled data were created separately and interpolation was set to bilinear. Positive and negative data points were hence visually classified based on the frequency and occurrence commonality.

A Bag of Words was created and the training and testing data split was created again within this smaller sample set.

Term Frequency Inverse Document Frequency method/function was called for feature extraction and the list was converted into a vector on the basis of the count of each word. The relevance of each word in the series/corpus was calculated.

The raw frequencies of a token was scaled down so that the impact of more frequently occurring words is reduced, so the less informative features that occur in a small fraction of the training corpus are dropped.

Fitting and transformation was done to scale the data points and generalize them so that the distance between them is reduced.

The mean and standard deviation of the data was calculated. Fitting as well as transforming was performed on the training data whereas only transforming was performed on the testing data.

Random Forest Classifier was then built to check for the accuracy. Since it is a meta estimator that uses multiple decision trees and aggregates the average output, the accuracy was expected to be the highest among all models we used, which later turned out to be true.

200 was the parameter passed to the Random Forest Classifier function, which indicates the number of decision trees to be used. A higher value of n estimators results in a higher accuracy for larger datasets.

The accuracy, F score and confusion matrix were generated and a submission CSV file was created with all the newly predicted/labelled test data.

An accuracy of 91% was achieved with an F score of 87%. When tested against news articles and survey data conducted across India, our data classified that 64% was pro CAA, whereas the new articles reported that 62% of the population was pro CAA. This was extremely insightful and indicative of how strong our model is.

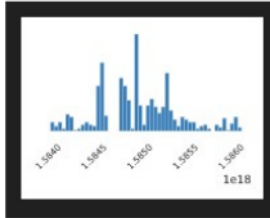
Various visualization plots were constructed and the same process was followed for the other two models.

Logistic Regression model had an accuracy of 87% while Gradient Boosting Regressor had an accuracy score of 90%. The comparison of the models was visualized using bar plots and it was concluded that Random Forest performed the best.

EDA

Dataset statistics		Variable types	
Number of variables	2	Numeric	1
Number of observations	627	Categorical	1
Missing cells	0		
Missing cells (%)	0.0%		
Duplicate rows	53		
Duplicate rows (%)	8.5%		
Total size in memory	9.9 KiB		
Average record size in memory	16.2 B		

1.58611E+18 Real number (R ₆₄)	Distinct	139	Minimum	1.58402 × 10 ¹⁸
	Distinct (%)	22.2%	Maximum	1.58611 × 10 ¹⁸
	Missing	0	Zeros	0
	Missing (%)	0.0%	Zeros (%)	0.0%
	Infinite	0	Negative	0
	Infinite (%)	0.0%	Negative (%)	0.0%
	Mean	1.584984354 × 10 ¹⁸	Memory size	5.0 KiB



RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's due. Inspite of not reading draft of NRC not reading about CAA...	Distinct	263	RT @MehboobaMufti: ...	152
Categorical	Distinct (%)	41.9%	RT @AparBharat: Hind...	69
HIGH CARDINALITY	Missing	0	RT @ThePollLady: @...	58
	Missing (%)	0.0%	RT @MaktoobMedia: It...	24
	Memory size	5.0 KIB	RT @pallavict: Every ...	13
			Other values (258)	311



First rows

1.58611E+18	RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's c
0	1.586110e+18 Almost enticing #ovlwar Amit Shah/BJP's #CAA triggered violence. 'With love (Islamophi
1	1.586080e+18 "There are organisations meant to work for mental patients or in leprosy colonies. But wh
2	1.586060e+18 @BJP4India as an Indian, I am eager to know what is population of INDIA. It will be very c
3	1.586060e+18 @azharlicks @RahulGandhi @INCIndia It's time you spoke against the dehumanization i
4	1.586060e+18 Damn he was hitting him hard. I want the same cops here to act against those anti-nation
5	1.586050e+18 RT @DocArifraHindu: @AskAnshul ASAP NRC, CAA implementation need in India .
6	1.586050e+18 RT @TimesNow: #TrumpHinduTribute/v/n/There are allegations against Joe Biden ∓
7	1.586050e+18 Devesh of @thewire_in also owned 4 Lakhs+ twitter bots to create Anti-CAA sentiments. I
8	1.586040e+18 @seemahegdearora @Poonam89625917 @NeelDas83815238 @minionair The article is
9	1.586040e+18 @TimesNow @Sanju_Verma_ @thenewshour @PadmajaJoshi Modi government should

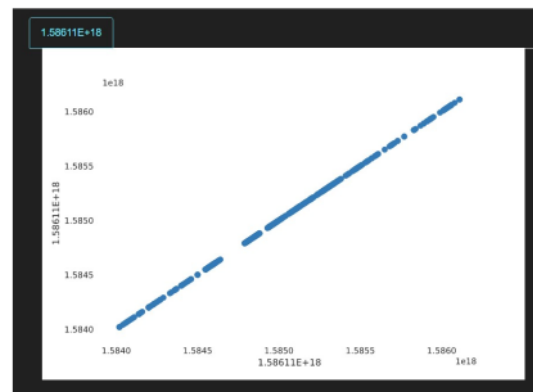
Last rows

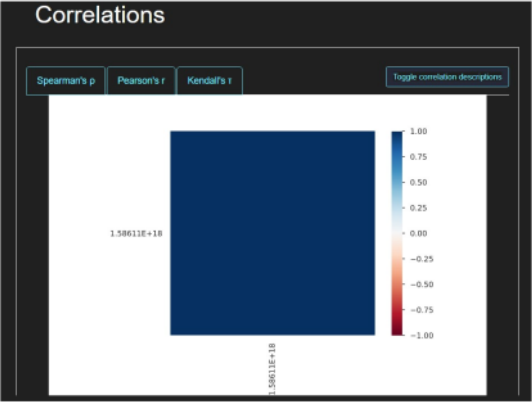
1.58611E+18	RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it'
617	1.585270e+18 @AskAnshul And radical mobs in India protests CAA, knowing full well the treatment H
618	1.585270e+18 @AskAnshul Why is @narendramodi @PMOIndia delaying the Implementation of the C
619	1.585270e+18 @BhoomjKolhi @grihs2 CAA law is for people like you to return to India unfortunately
620	1.585270e+18 @fiwaryshweta AK has done so much work in education and healthcare sectors, but th
621	1.585260e+18 @warispathan When you will say pakistan murdabad/n/Vande mataram/nMaking fatwa
622	1.585260e+18 RT @tarungupta970: @BhoomjKolhi Why don't you apply for Indian Citizenship under
623	1.585260e+18 #Democracy #Politics #Appeasement #Liberals #Hindus #leftists #congress #AAP #PV
624	1.585250e+18 @Dev_Uvacha @experienceluv @BhoomjKolhi CAA rules not notified. Plight of Pakis
625	1.585240e+18 Almost enticing civil war Amit Shah/BJP's CAA triggered violence. 'With love (Islamophi
626	1.586110e+18 RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's due



Duplicate rows

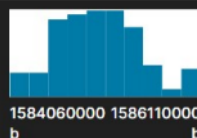
Most frequently occurring

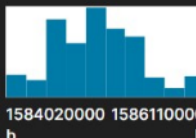
1.58611E+18	RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's
24	1.584940e+18 RT @AparBharat: Hindus living in Pakistan are also indians/r/nThat is the reason why k
7	1.584550e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
9	1.584570e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
10	1.584580e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
8	1.584560e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
12	1.584600e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
15	1.584800e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
17	1.584810e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a
1	1.584200e+18 RT @MaktoobMedia: It has been 1078 days since the movement against CAA began in
11	1.584580e+18 RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While a





# id	# label	tweet
 1584020000 1586110000 b	 0 1	RT @MehboobaM... 24% RT @AparBharat: ... 12% Other (322) 64%
1.58611E+18	0	RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's due.Inspite of not readi...
1.58531E+18	0	RT @ThePollady: @AskAnshul And radical mobs in India protests CAA, knowing full well the treatment ...
1.58458E+18	0	@MehboobaMufti we r happy for Sunak. CAA & NRC is not discriminatory. Muslims wanted separate Na...

# id	tweet
 1584060000 1586110000 b	RT @MehboobaM... 24% RT @ThePollady: ... 9% Other (85) 67%
1.58611E+18	RT @Informaticafan: @ReallySwara @RahulGandhi @bharatjodo Credit where it's due.Inspite of not readi...
1.58484E+18	RT @MehboobaMufti: Proud moment that UK will have its first Indian origin PM. While all of India rig...
1.5842E+18	RT @MaktoobMedia: It has been 1078 days since the movement against CAA began in India; it has been 1...

# 1.58611E+18	RT @Informaticaf...
 1584020000 1586110000 b	RT @MehboobaM... 24% RT @AparBharat: H... 11% Other (406) 65%
1.58611E+18	Almost enticing #civilwar Amit Shah/BJP's #CAA triggered violence. 'With love (Islamophilia) from We...
1.58608E+18	"There are organisations meant to work for mental patients or in leprosy colonies. But when their ac...
1.58606E+18	@BJP4India as an Indian, i am eager to know what is population of INDIA. it will be very commendable...

REFERENCES

- [1] <https://iopscience.iop.org/article/10.1088/1742-6596/1575/1/012083/meta>
- [2] <https://ieeexplore.ieee.org/document/8951367>
- [3] https://www.researchgate.net/publication/342871060_Sentiment_Analysis_on_Students'_Evaluation_of_Higher_Educational_Institutions
- [4] <http://www.ijadis.org/index.php/IJADIS/article/view/sentiment-analysis-approach-for-analyzing-iphone-release-using-s>
- [5] <https://www.sciencedirect.com/science/article/pii/S1568494620309959>
- [6] <https://kjar.spu.edu.iq/index.php/kjar/article/view/512/262>
- [7] <https://towardsdatascience.com/a-three-level-sentiment-classification-task-using-svm-with-an-imbalanced-twitter-dataset-ab88dcd1fb13>
- [8] <https://www.thehindia.com/news/national/62-per-cent-people-across-india-support-caa-survey-591624>

DA_Project_Conference_Paper_175_009_188

ORIGINALITY REPORT

2%

SIMILARITY INDEX

1%

INTERNET SOURCES

2%

PUBLICATIONS

0%

STUDENT PAPERS

PRIMARY SOURCES

1

Avnish Goel, Apoorv Kashyap, B. Devesha Reddy, Rochak Kaushik, S Nagasundari, Prasad B Honnavali. "Detection of VPN Network Traffic", 2022 IEEE Delhi Section Conference (DELCON), 2022

Publication

1%

2

www.siplab.fct.ualg.pt

Internet Source

1%

Exclude quotes On

Exclude matches < 5 words

Exclude bibliography On