Int. j. inf. tecnol. (August 2021) 13(4):1555–1567
https://doi.org/10.1007/s41870-020-00604-8

ORIGINAL RESEARCH

Sentimental study of CAA by location-based tweets

Geetika Vashisht[1] · **Yash Naveen Sinha**[1]

Received: 30 June 2020 / Accepted: 26 December 2020 / Published online: 22 March 2021

Abstract As people progressively resort to twitter to express their opinions or to disambiguate their sentiment, it's feasible to analyze the mass opinion to conclude the polarity of the subject at hand using sentiment analysis. Sentiment Analysis (SA) has revolutionized the way information is perceived today. Inspired by this, the work in this paper investigates the much-debated act- the Citizenship Amendment Act (CAA) by analyzing opinionated geo-tagged tweets, manually annotated and cross verified by six annotators. This is the first paper to the best of our knowledge to analyse CAA using SA and to provide a clear statistics of the mass opinion across the states of the nation. In this paper, machine learning approach is used for sentiment analysis of tweets. Support vector machine classifier is used to classify the tweets into three classes viz. positive, negative and neutral.

Keywords Citizen amendment act (CAA) · Support vector machine (SVM) · National register of citizen (NRC) · Sentiment analysis (SA) · CAB · AntiCAA · AntiCAA protests

✉ Geetika Vashisht
geetika.vashisht@gmail.com

Yash Naveen Sinha
yashnsinha@gmail.com

[1] Department of Computer Science, University of Delhi, New Delhi, India

1 Introduction

The Citizenship (Amendment) Act, 2019 was passed by Indian Parliament on 11 December 2019. As per the CAA, Hindu, Christian, Buddhist, Jain, Sikh and Parsi migrants who have entered India without a visa-on or before December 31, 2014 from the Muslim-majority countries like Pakistan, Bangladesh and Afghanistan and have stayed in the country for five years, are eligible to apply for Indian citizenship.

According to the Union government people of these faiths (mentioned earlier) have faced persecution in three Islamic countries, Pakistan, Bangladesh and Afghanistan, except the Muslims there. Therefore, it is the moral obligation of Indian's to provide them shelter. Since the CAA does not specifically mention "persecution" anywhere as a criterion to fall under the ACT, it is assumed to be discriminating against illegal Muslim immigrants from these countries.

This is for the first time that religion has been overtly used to provide citizenship to people in India. Although the outrage for CAA and NRC was seen at the same time, but there is certainly no connection between the two. National Register of Citizen (NRC) is an amendment of the Citizenship Act, 1995.The purpose of NRC is to document all the legal citizens of India to identify the illegal immigrants and deport them. This was initially state specific exercise unlike CAA, to keep its uniqueness still. But after its implementation in Assam it has been gaining popularity and many top BJP leaders wants it to be implemented nationwide.

The proposal of NRC Assam was followed by CAA and therefore it created a havoc in the minds of the people. They were unable to differentiate between the two and stared making false assumptions about the proof of

citizenship. As proposed by the BJP government, CAA does not look for ancestral proofs of citizenship but can only be claimed by manifestation of a Voter Id or Aadhaar Card.

The clash between pro and anti-CAA led to violence and destruction of public property. The protests soon converted into riots and led to devastation. The crowd was difficult to control so the police had to take extreme measures like lathi charge in places like Jaffrabad, Maujpur, Chandbagh. The silent protest at Shaheen Bagh, Delhi continued until the country was locked down due to Covid-19 pandemic. Many doubt these to be paid protests and no statistics is clearly available to categorize the mass opinion. This is indeed intriguing to know the actual sentiment of the majority. To get a clear statistics and to facilitate the research, twitter is targeted in this paper to mine the tweets opinionated on CAA. As per statistics provided by Twitter, it has 350 million monthly active users and 145 million daily active users making it a gold mine of opinionated data. This popular micro-blogging site has the potential to throw a light on the level of acceptance of the aforementioned act by the mass. To achieve the goal, tweets opinionated for the act under scrutiny were extracted and classified using the machine learning approach. Accuracy, precision, recall and F1-score are used as the evaluation metric to evaluate the performance of the machine learning model.

The paper commences with literature review in Sect. 2 consisting of a brief overview of the related work in the field. Further, Sect. 3 throws a light on the dataset, pre-processing techniques and the flow of work. It includes the exploratory data analysis with appropriate graphs followed by the sentiment classification task. Section 4 discusses results followed by conclusion in the last section.

2 Related work

Sentiment analysis aims at computationally identifying and categorizing the opinionated data according to the polarity mainly positive, negative and neutral. The analysis can be done at three levels viz. sentence-level, document-level and the phrase-level [1]. The sentiment analysis techniques are broadly classified into five categories, i.e., lexicon-based techniques, machine learning based techniques, hybrid techniques, rule-based techniques and ontology-based approaches [2–5].

There is another lesser explored technique known as Emoticons (Emojis) based sentiment analysis technique that exploits the graphical cues in the text to determine the sentiments. The lexicon-based approach relies on an annotated lexicon, machine learning approaches work by applying different algorithms and the hybrid approaches are based on the combination of these two approaches [6–8].

A number of studies have used these lexicon based approaches [9–13], machine learning based supervised [6, 14, 15, 16] or unsupervised [6, 17, 18] approaches, and hybrid approaches [6, 19] to classify sentiments into positive or negative categories. Sankar and, Subramaniyaswamy [20] analyses the machine learning classifiers at length and concludes that in the last few years, Naive Bayes and support vector machine classifiers surfaced as the most preferred choice of machine learning classifiers for analysing sentiments.

Kawade and Oza [21] uses machine learning classifiers, Naive Bayes and Genetic Algorithm to accomplish 4-way classification of people's sentiments on URI attack. [22] exploits the supervised machine learning classification technique using Support vector machine classifier for polarity detection. Gopalakrishnan and Ramaswamy [23] compares the performance of neural networks and support vector machine classifiers to mine opinions on social web in the health care domain. Moreno-ortiz and Fernández-cruz [12] uses lexicon based approach for identifying polarity of texts in specialised domains like the one considered by the author for financial texts. Bhoir [24] uses the SentiWordNet lexicon to categorize the movie reviews into two categories viz. positive and negative. Vu [25] uses a lexicon based approach with a heuristic approach for pre-processing the data improving the performance of the traditional lexicon-based methods. Das and Behera [26] uses RNNs LSTM (recurrent neural network—long short term memory) to do real-time sentiment analysis of tweets for stock prediction.

Further, the methods to measure the sentiment strength can be categorized based on the rating levels—one for identifying different aspects of a product and the other attempts to rate a review on a global level that considers only the polarity of the review (positive/negative) [27].

3 Data and methodology

This section discusses the work flow of the sentiment analysis process. The section is further sub-divided into three subsections that explores the dataset creation followed by a detailed analysis of the pre-processing steps to make the data ready to be fed to the classifiers. Exploratory data analysis is performed to get acquainted with the data followed by the sentiment classification.

3.1 Data collection and pre-processing

The micro-blogging site-Twitter is the most popular choice for sentiment analysis as the micro-blogs in Twitter have a

limited length and it is all about what's happening in the world. Twitter is widely used by people across the globe and gives us a great insight of what people are thinking and talking about [28]. Twitter can be easily accessed via smart phones or laptop or similar devices. Twitter also provides APIs (application programming interfaces) to provide broad access to tweets with of course certain amount of protections Twitter has in place for their use. Additionally, Twitter is easy to work with because of myriads of apps built over years to aid developers in interfacing with it. Around ten thousand tweets voicing opinions about Citizenship Amendment Act (CAA) were extracted from 1st to 30th March, 2020 for ten different regions of India viz. Assam, Andhra Pradesh, Uttar Pradesh, Maharashtra, West Bengal, Tamil Nadu, Kerala, Karnataka, Punjab and Delhi using Twitter's streaming API. Retweets (repetitions of previously posted messages) were removed by excluding messages which contain the "retweeted status" metadata or the "RT" token. A high variation is observed in the opinion of the mass towards CAA because of several factors like the campaigns or media influence. The data was extracted to capture public sentiments about the much debated act-CAA. A wide-ranging list of keywords and hashtags were used to collect the relevant tweets region-wise. For a better result, tweets were manually annotated for 3-way classification viz. positive, negative and neutral by two annotators and the work was cross-verified. Figure 1 presents the sequence of activities followed in this work.

The process of sentiment analysis initiates with the extraction of 'emotional data'. Here, the opinions of the people in the form of tweets are extracted:

1. A Twitter application is created to do the handshake to get the data from Twitter with R. The link apps.twitter.com/ can be followed to create the new application here. Twitter account is a pre-requisite.
2. Package twitteR is used to extract information from Twitter using Twitter APIs.
3. Package ROAuth is used to interface for OAuth. The OAuth settings will be needed for Twitter app.
4. Twitter authenticated credential object is created using consumer key, consumer secret, access token, access secret. Using these keys, Twitter account can be easily accessed.

Tweets are generally not structured and contain noise because of URLs, symbols, slangs etcetera [29]. Therefore, pre-processing is needed. It involves converting the text into same case usually lowercase; removal of URLs, white spaces, punctuations, symbols, numbers and stop words [30], lemmatisation.

Tweets were pre-processed in several stages:

1. *Conversion to lower case* Since the case of a string does not alter the meaning of the string, for accurate string matching, the tweets are converted to lower case.
2. *Removing URLs* All of the URLs are eliminated via regular expression matching.
3. *Removing@username* Remove "@username" via regular expression matching.
4. *Removing numbers and punctuations* Punctuations at the start and ending of the tweets are removed, e.g.: 'the day is beautiful!' replaced with 'the day is beautiful'.
5. *Striping additional white spaces* Multiple whitespaces are replaced with a single whitespace. Unnecessary tabs, spaces, newline characters are removed.
6. *Removal of stop words* Stop words are removed using existing stop words list and a customized list.
7. *Lemmatisation* It's the process of reducing words to their meaningful root word keeping the meaning intact. For example, good, better and best are reduced to their lemma which is good as all contribute to the similar meaning.
8. *Duplicate Tweets* For an accurate analysis, duplicate tweets shall be removed but retweeted tweets were kept as they denote unique sentiments leaving the count of tweets as mentioned in the Table 1.

The dataset is mildly unbalanced, thus, Synthetic Minority Oversampling Technique (SMOTE) is used to balance the classes post annotation as shown in Fig. 1. SMOTE is a statistical technique for increasing the number of cases in the dataset by generating new instances from existing minority cases that are supplied as input.

3.2 Exploratory data analysis

This section gives maximum insight into the dataset by summarizing the main characteristics graphically. The data is primarily explored to find out what it can tell us beyond the formal modelling task done in the next section. It is a vital step to acquaint oneself with the data before moving on to modelling the data. Figure 2 explores the reaction of people to CAA in Delhi using syuzhet package in R. Syuzhet package's get_nrc_sentiment function is used to generate each region's data sets into different sentiments and the result is displayed with the help of graph using barplot function from the ggplot2 package.

Figure 3 explores the reaction of people to CAA in Assam with the majority of the tweets not in the favour of the act.

Figure 4 explores the reaction of people to CAA in Uttar Pradesh with the majority of the tweets not in the favour of the act.

Fig. 1 Work flow

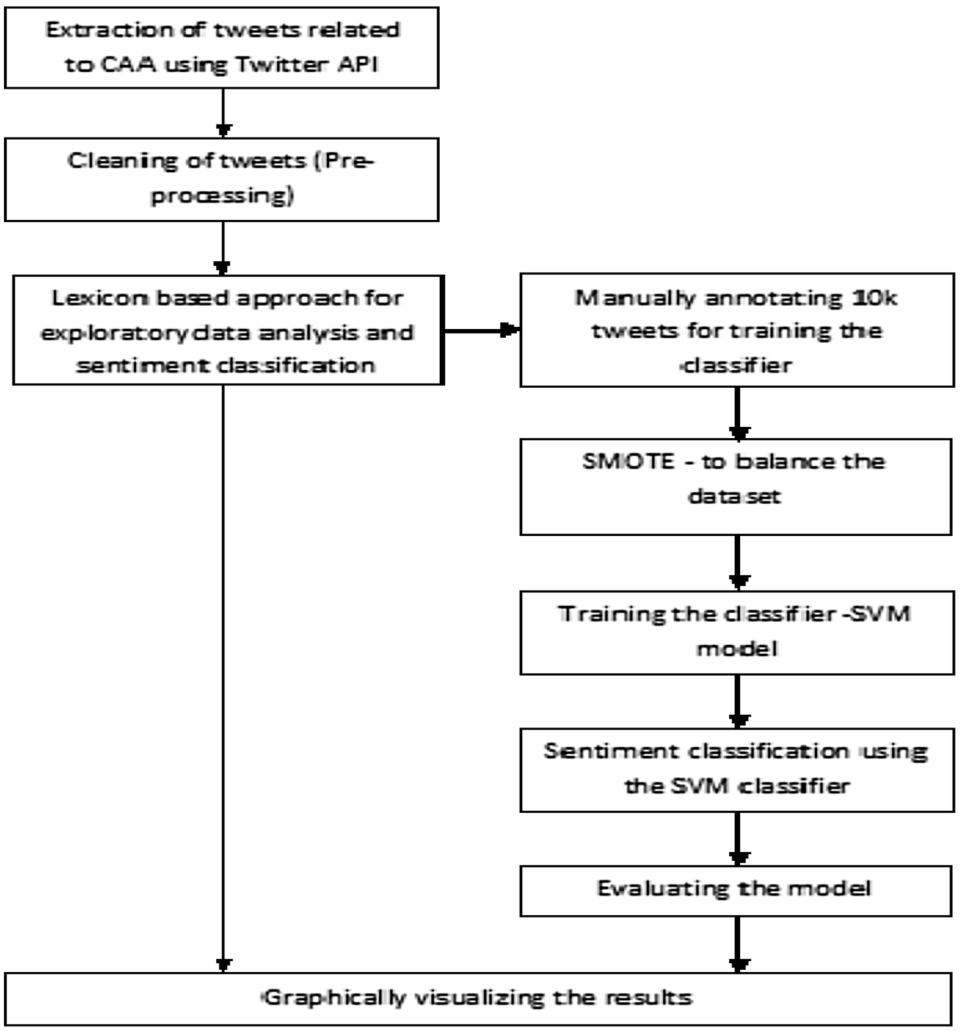

Table 1 Count of the total tweets region-wise

Region	#Tweets post pre-processing	Hash tags and usernames
Delhi	1568	#CAA
Assam	626	#CAB
Andhra Pradesh	3192	#AntiCAA
Uttar Pradesh	1480	#AntiCAAProtests
Maharashtra	3001	#CitizenshipAmmendementAct
West Bengal	1575	
Tamil Nadu	3211	
Kerala	1040	
Karnataka	2541	
Punjab	745	

Figure 5 explores the reaction of people to CAA in Punjab with the majority of the tweets supporting the act.

Figure 6 explores the reaction of people to CAA in Karnataka with the majority of the tweets supporting the act.

Figure 7 explores the reaction of people to CAA in Kerala. The negative tweets outnumber the positives by a marginal value.

Figure 8 explores the reaction of people to CAA in West Bengal where it is observed the count of the tweets not in

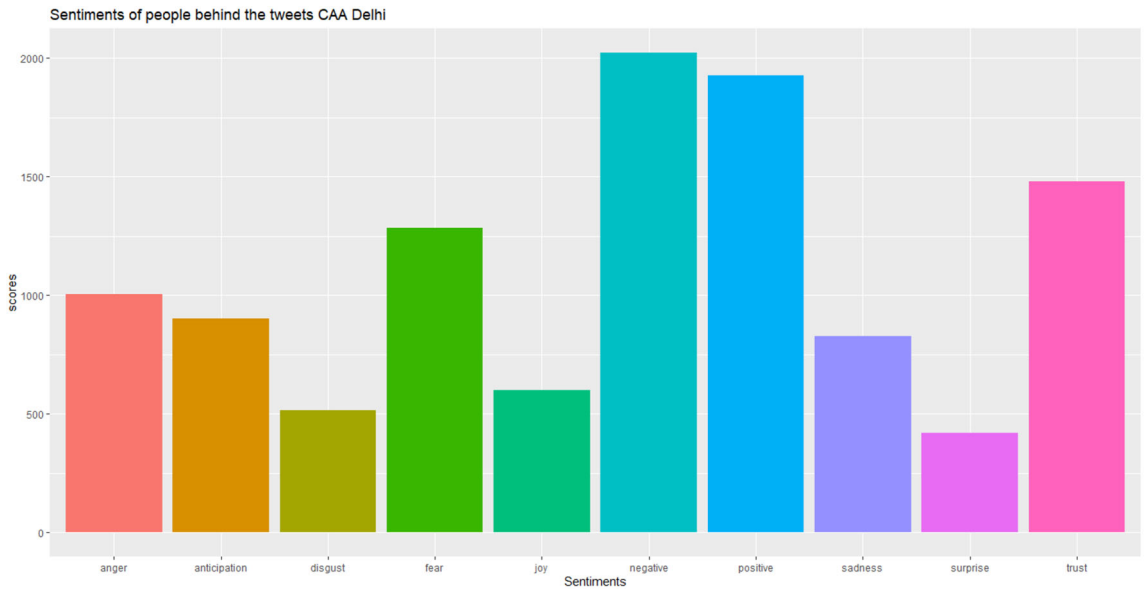

Fig. 2 Sentiment analysis graph of CAA tweets extracted from Delhi

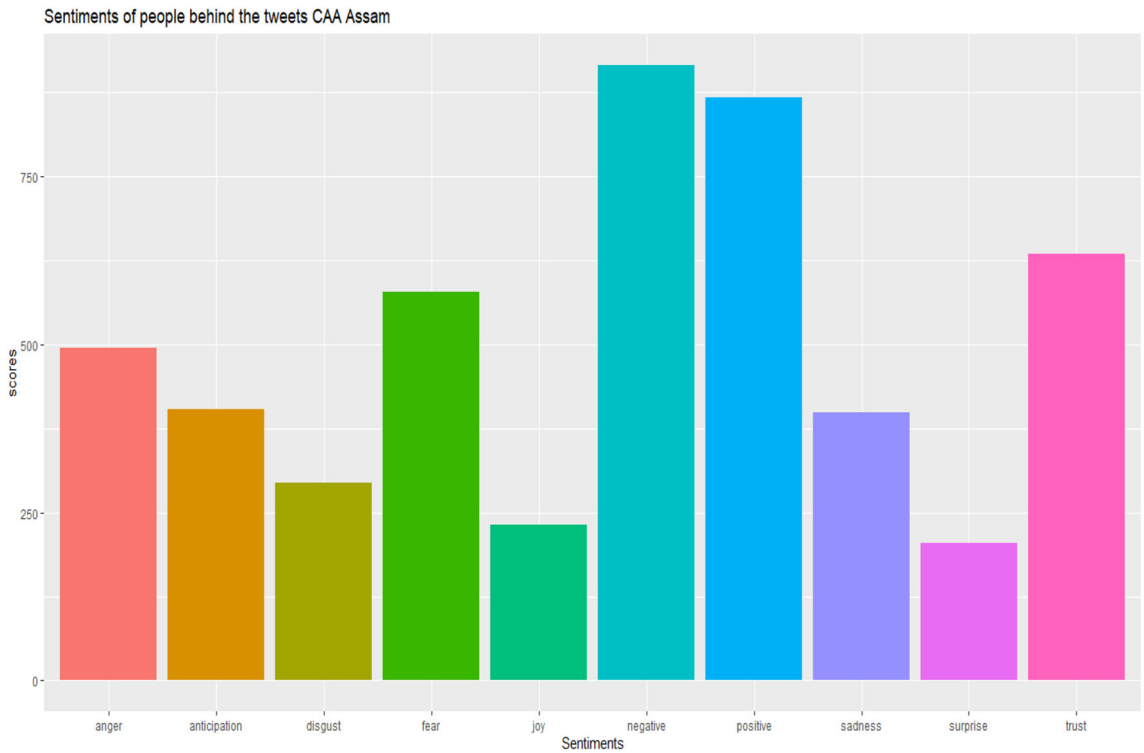

Fig. 3 Sentiment analysis graph of CAA tweets extracted from Assam

favour of the act are slightly higher than the tweets supporting the act.

Figure 9 explores the reaction of people to CAA in Andhra Pradesh and Telangana with the majority of the tweets in the favour of the act.

Figure 10 explores the reaction of people to CAA in Maharashtra with the majority of the tweets supporting the act.

Figure 11 explores the reaction of people to CAA in Tamil Nadu where it is observed the count of the tweets not in favour of the act are marginally higher than the tweets supporting the act.

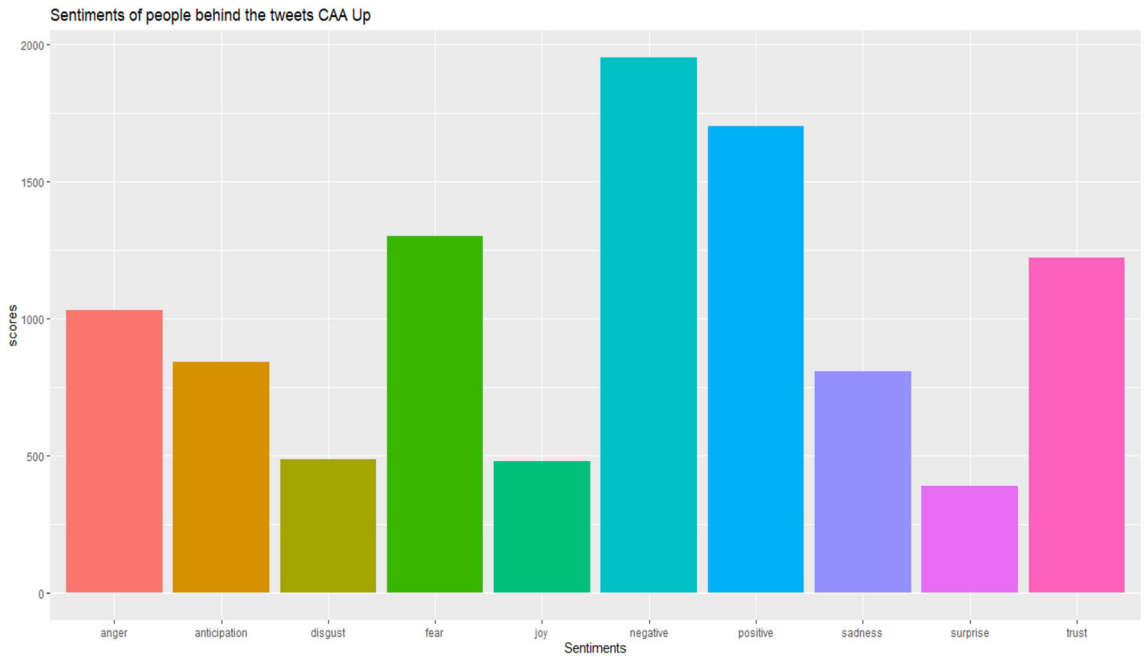

Fig. 4 Sentiment analysis graph of CAA tweets extracted from Uttar Pradesh

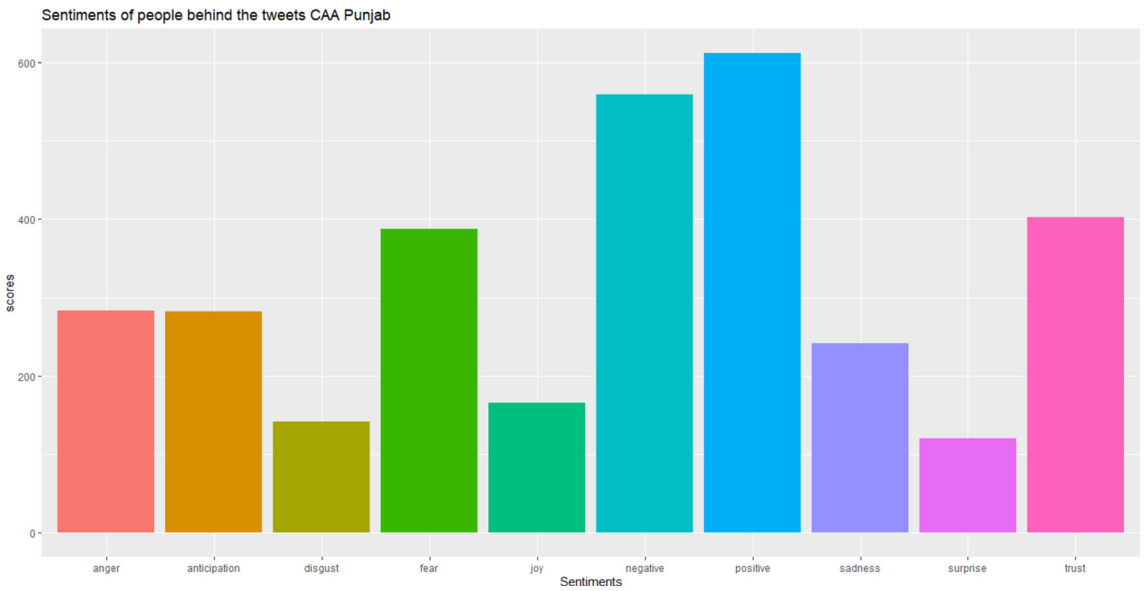

Fig. 5 Sentiment analysis graph of CAA tweets extracted from Punjab

Figure 12 does a comparative analysis of different emotions exhibited by tweets across the different regions of the nation indicating that the mass opinion is inclined towards the act and people have accepted and welcomed it at a great scale.

3.3 Sentiment classification

Multi-class support vector machine classifier is used for training and testing the data. Support vector machine or SVM is a supervised machine learning algorithm which can be used for both classification or regression problems. However, we mostly use it for classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n stands for the number of features we have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the classes. In this support vectors are actually the co-ordinates of all the individual observations. It is mainly used for two class

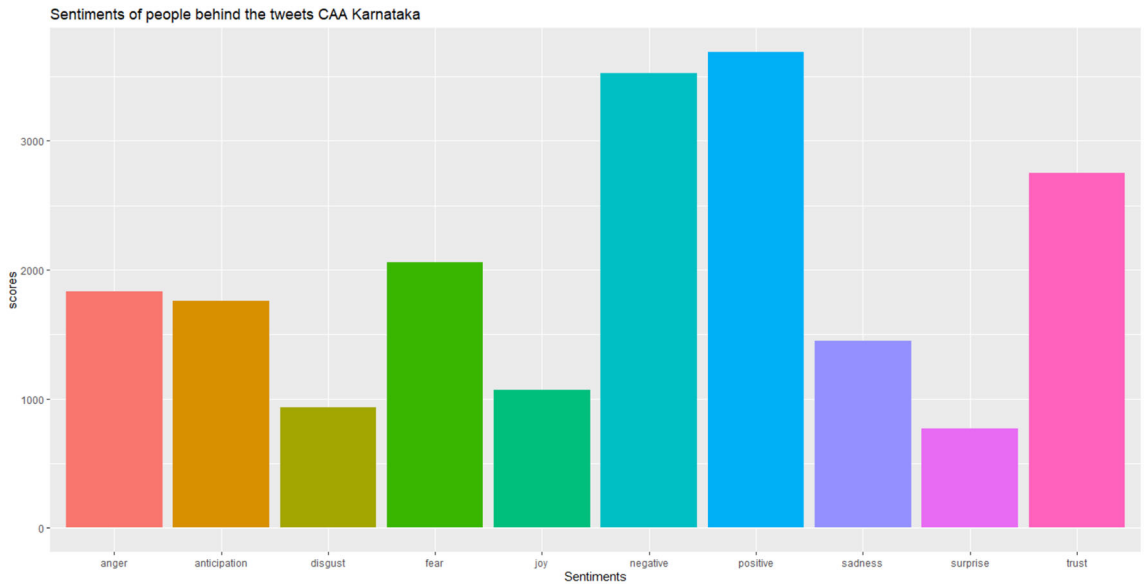

Fig. 6 Sentiment analysis graph of CAA tweets extracted from Karnataka

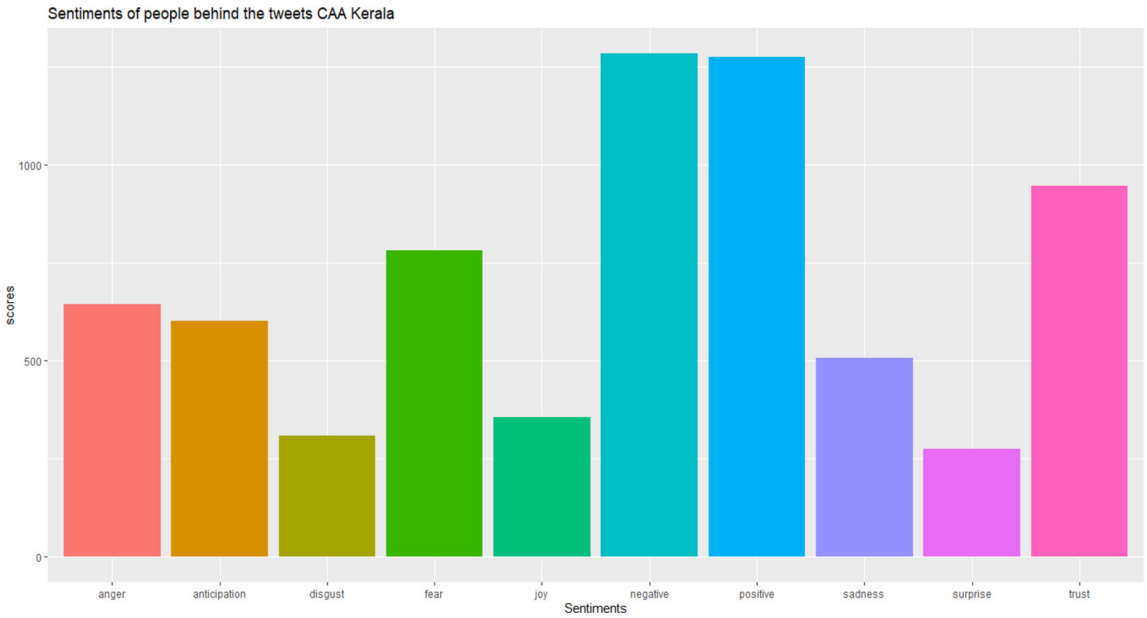

Fig. 7 Sentiment analysis graph of CAA tweets extracted from Kerala

problems (and is among one of the best algorithms to use for that task) but it can also be used for multi class classifications.

Why choose this model?

- Simple, fast and easy to implement.
- Highly scalable.
- Can be used for classification problems (both binary and multiclass).
- Insensitive to irrelevant features.
- Works with categorical data as well as continuous data.

The model is initially split into two sections viz. 'Training data' and 'Testing data' to train the model. First, the classifier is trained using 'training data set' and then tested the performance of the classifier on unseen 'test data set'. For this step, the dataset used in this work is split in 7:3 ratio, i.e., 70% training dataset and 30% testing dataset. Ten-fold cross validation is used to avoid model overfitting.

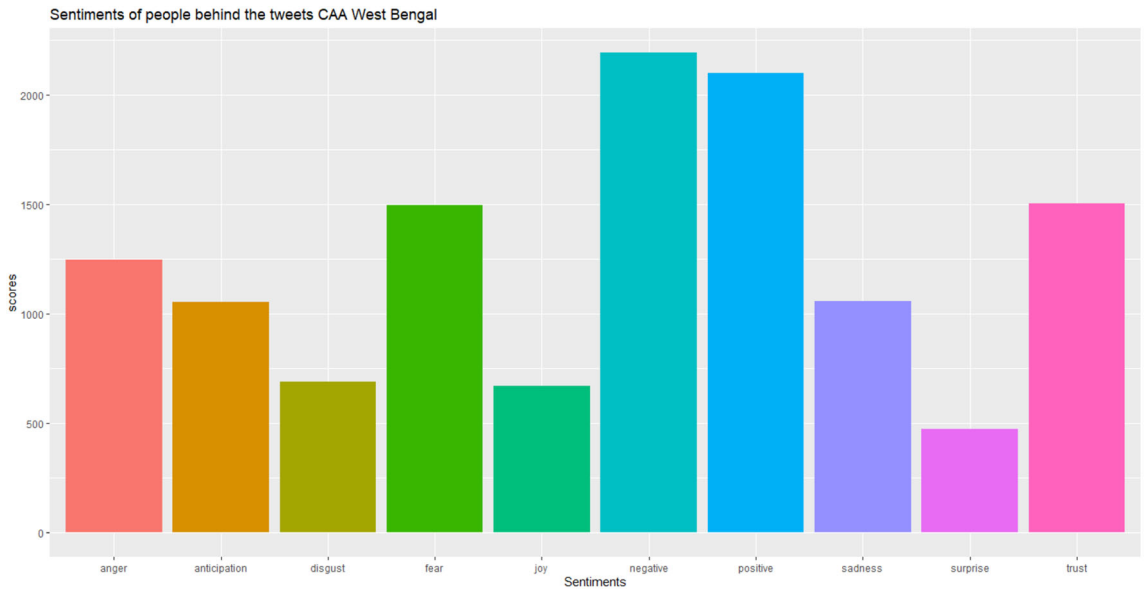

Fig. 8 Sentiment analysis graph of CAA tweets extracted from West Bengal

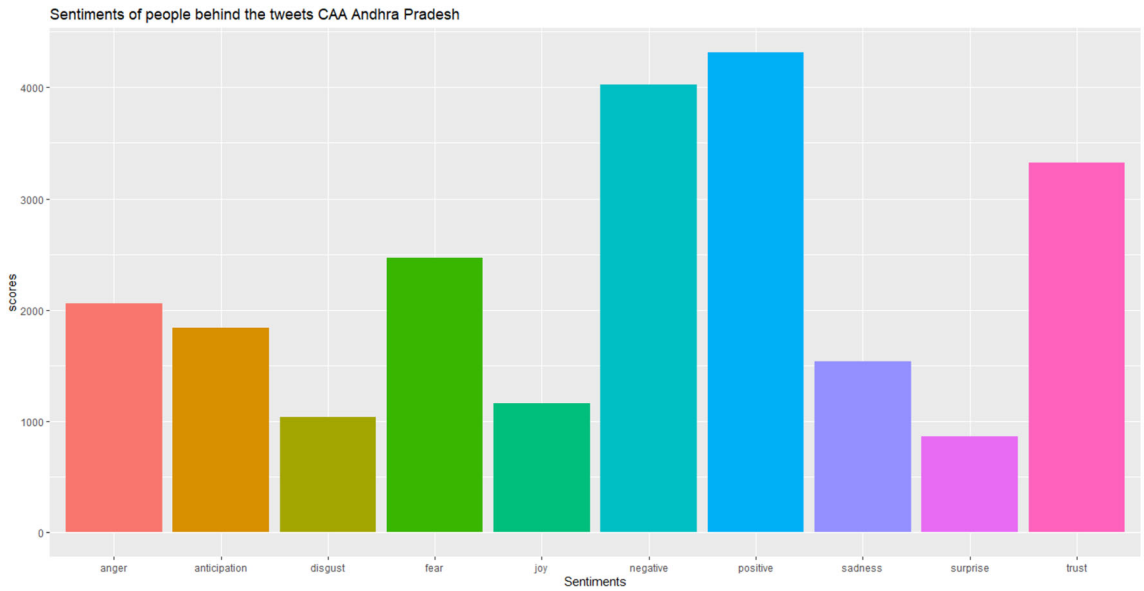

Fig. 9 Sentiment analysis graph of CAA tweets extracted from Andhra Pradesh and Telangana

4 Results and discussion

Confusion matrix (Fig. 13) is a tool to evaluate a classifier. Earlier, accuracy (Eq. 1) was the main measure to determine the performance of the classifier which determines how much positive class values are correctly classified as positive and vice-versa. In simple words, it tells that overall, how often the classifier is correct.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FN} + \text{FP}), \quad (1)$$

where, True Positive, TP is the Observation is positive and is predicted correctly as positive. False Positive, FP is the Observation is negative but is predicted positive. True Negative, TN is the Observation is negative and is predicted correctly as negative. False Negative, FN is the Observation is positive but is predicted as negative.

Though popular, accuracy cannot be relied upon when the data set is completely off-centre or imbalanced. The skewed data is biased towards the dominant class. Taking that into account, several other measures are introduced to accurately evaluate the classifier. The other most sought-after performance measures are sensitivity, specificity, kappa and prevalence.

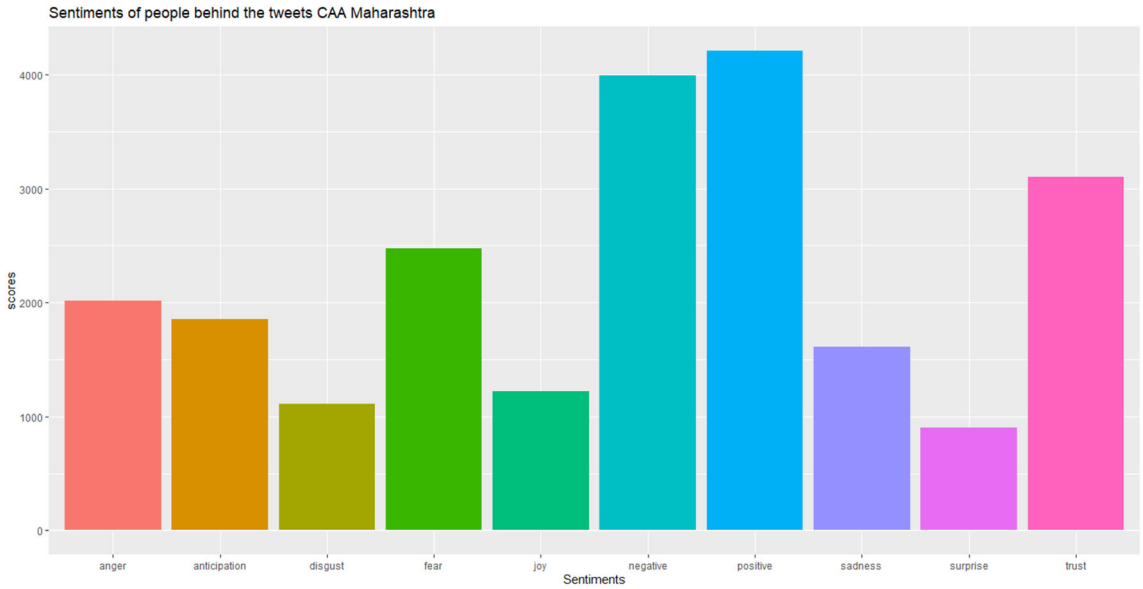

Fig. 10 Sentiment analysis graph of CAA tweets extracted from Maharashtra

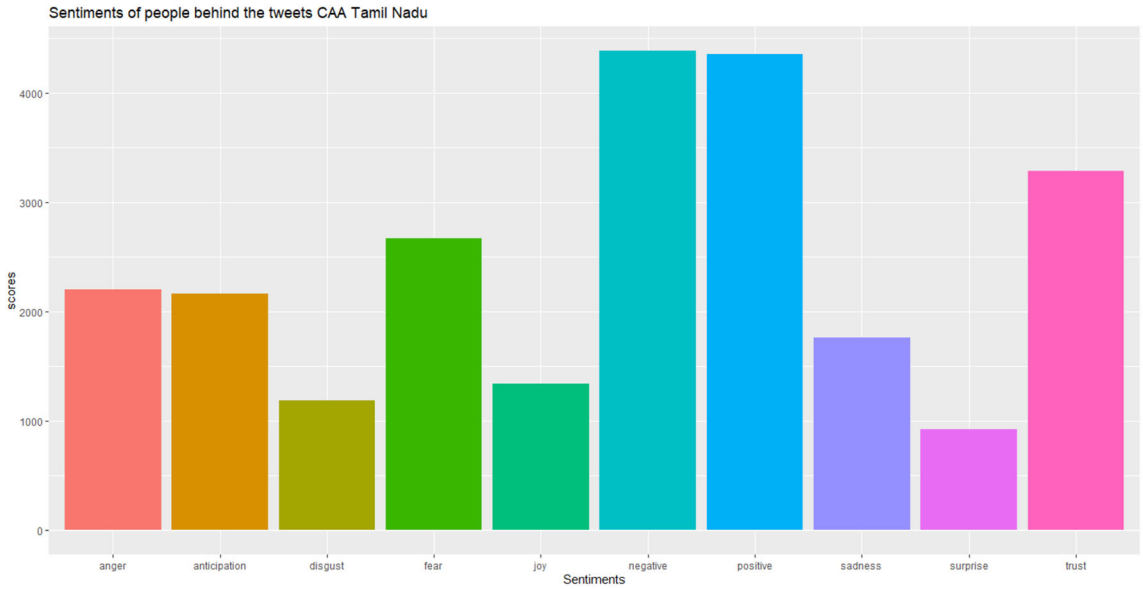

Fig. 11 Sentiment analysis graph of CAA tweets extracted from Tamil Nadu

Sensitivity (Eq. 2) also called as Recall is the capability of a classifier to correctly predict positive when it is actually positive i.e. the true positive rate,

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}). \quad (2)$$

Specificity (Eq. 3), also called True Negative Rate is the capability of the classifier to leave out the values of the positive class from negative class and vice-versa appropriately. It is the ability of a classifier to correctly identify the true negative rate or in other words, ability of the classifier to predict negative when it is actually negative. Specificity should be high. It's equivalent to 1 minus false positive rate.

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}). \quad (3)$$

People's reactions in tweets has been categorised in three different classes viz. positive, negative and neutral. Table 2 tabulates classification output showing the number of input test data.

Figure 14 shows the confusion matrix of the SVM model.

Kappa (Eq. 4) is used to measure the inter-rater reliability for categorical items. It compares the observed

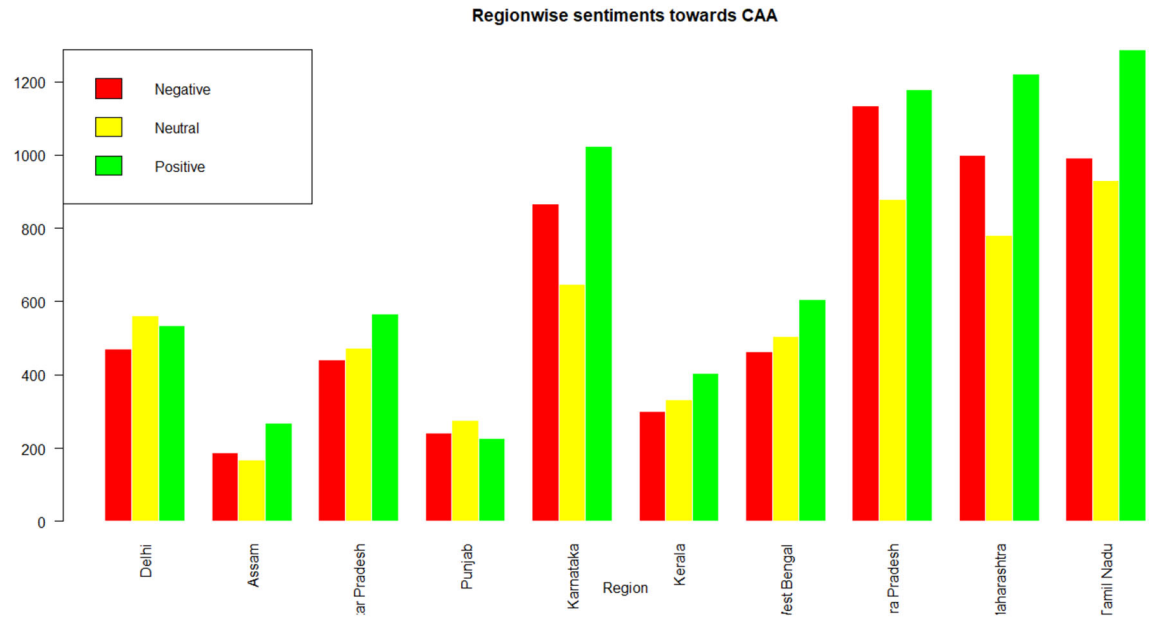

Fig. 12 This graph shows the comparison of sentiments of tweets in support of or against CAA from different regions in all over India

		Predicted				
	Confusion Matrix	Class 1: Positive	Class 2: Neutral	Class 3: Negative	False Negative (FN)	Recall
Actual	Class 1: Positive	A (TP)	B	C	B+C	A/(A+B+C)
	Class 2: Neutral	D	E (TNe)	F	D+F	E/(D+E+F)
	Class 3: Negative	G	H	I (TN)	G+H	I/(G+H+I)
	False Positive (FP)	D+G (FP)	B+H	C+F	Accuracy=A+E+I /(Sum of green and blue boxes)	
	Precision	A/(A+D+G)	E/(B+E+H)	I/(C+F+I)		

Fig. 13 Confusion matrix with advanced classification metrics for three classes

Table 2 Tweets classified into different classes

Predicted/expected	Positive	Neutral	Negative
Positive	963	131	109
Neutral	97	972	92
Negative	151	108	1010

accuracy with an expected accuracy. It is considered a promising measure in evaluating a classifier as it compares the classifier to that of an ideal one. It also finds if the classifier has in actual trained itself well to classify based on the relation between the attributes and the exact classification methodology rather than just replicating the class of the duplicated values,

$$\begin{aligned}\text{Kappa} &= (\text{accuracy} - \text{random})/(1 - \text{random})\\ \text{Random} &= ((\text{TN} + \text{FP}) \times (\text{TN} + \text{FN}) + (\text{FN} + \text{TP}) \times (\text{FP} + \text{TP}))/\\ &\quad (\text{TP} + \text{TN}) \times (\text{TP} + \text{TN})\end{aligned} \tag{4}$$

Prevalence (Eq. 5) is the ratio of the positive class to that of the total population,

$$\text{Prevalence} = (\text{TP} + \text{FN})/\text{N}. \tag{5}$$

Precision or positive prediction value (PPV) (Eq. 6) indicates how often the prediction is correct when it is positive,

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}). \tag{6}$$

Negative prediction value (NPV) (Eq. 7) indicates how often the prediction is correct when the prediction is classified as negative by the classifier.

$$\text{NPV} = \text{TN}/(\text{TN} + \text{FN}). \tag{7}$$

Misclassification rate (Eq. 8) tells how often the test is wrong,

$$\text{Misclassification rate} = (\text{FP} + \text{FN})/\text{N} = 1 - \text{accuracy}. \tag{8}$$

False positive rate (FPR) (Eq. 9) indicates the probability of false prediction as positive,

$$\text{FPR} = \text{FP}/(\text{TN} + \text{FP}) = 1 - \text{specificity}. \tag{9}$$

False negative rate (FNR) (Eq. 10) indicates the miss rate i.e. the probability of identifying the negative correctly.

$$\text{FNR} = \text{FN}/(\text{TP} + \text{FN}) = 1 - \text{sensitivity}. \tag{10}$$

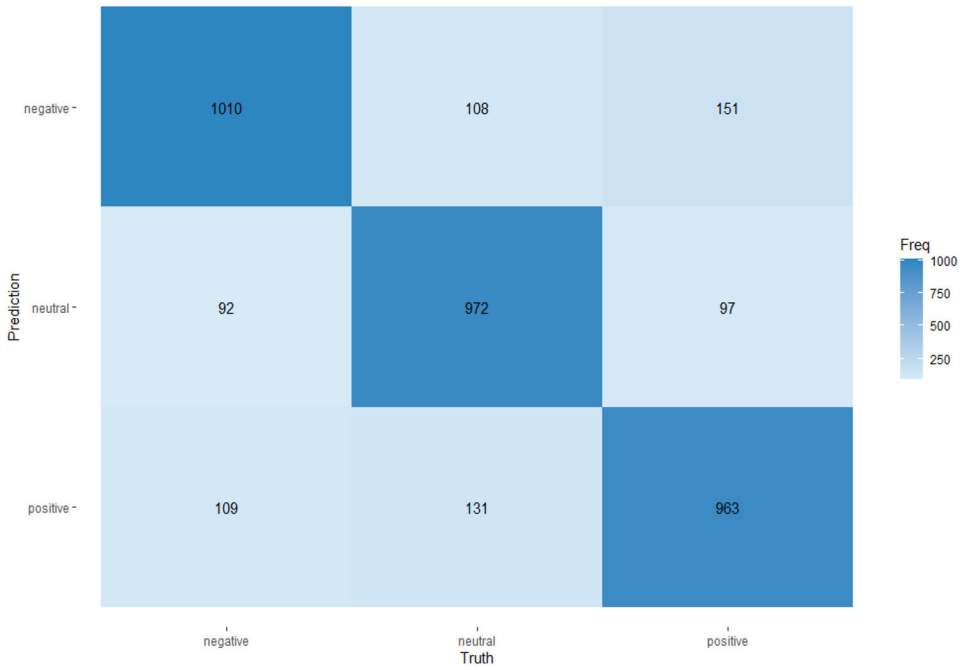

Fig.14 Graphical representation of the confusion matrix of the built SVM model

Table 3 Accuracy and kappa measures

Parameters	Value
Accuracy	963
Kappa	97

Table 4 Performance measures for the three classes

Predicted/expected	Positive	Neutral	Negative
Precision	0.8005	0.8372	0.7959
Recall	0.7952	0.8026	0.8340
F1	0.7978	0.8196	0.8145
Sensitivity	0.7952	0.8026	0.8340
Specificity	0.9009	0.9220	0.8931
Prevalence	0.3333	0.3333	0.3333
Balanced accuracy	0.8481	0.8623	0.8635

Positive likelihood ratio (Eq. 11) indicates the odds of a positive prediction given it is positive,

$$\text{Positive likelihood ratio (PLR)} = \text{TPR}/\text{FPR} = \text{sensitivity}/(1-\text{specificity}) \quad (11)$$

Negative likelihood ratio (Eq. 12) indicates the odds of a positive prediction when in actual it is negative,

$$\text{Negative likelihood ratio (NLR)} = \text{FNR}/\text{TNR} = (1-\text{sensitivity})/\text{specificity}. \quad (12)$$

Table 4 presents low recall and high precision for positive class indicating that FN is slightly higher missing few positive tweets but those predicted as positive are indeed positive (low FP). Neutral class also has low recall and high precision. While for the positive class recall is higher than precision which means that most of the negative tweets are correctly recognised but there are lot of false negatives.

It is clearly depicted in Fig. 15 that people have supported CAA. Although the positive tweets outnumber the negative tweets but the difference is not that much by which it can be concluded that the people might have mixed opinions about the act.

5 Conclusion

People are increasingly voicing their opinion on social media platforms, twitter being the most popular one is exploited in this work to classify mass opinion on the highly debated act like the Citizenship Amendment Act (CAA). Tweets opinionated on CAA were extracted from ten different regions of India viz. Assam, Andhra Pradesh, Uttar Pradesh, Maharashtra, West Bengal, Tamil Nadu, Kerala, Karnataka, Punjab and Delhi with an underlying motive of exploring the opinion of the mass on the newly introduced act by the government. A lexicon-based

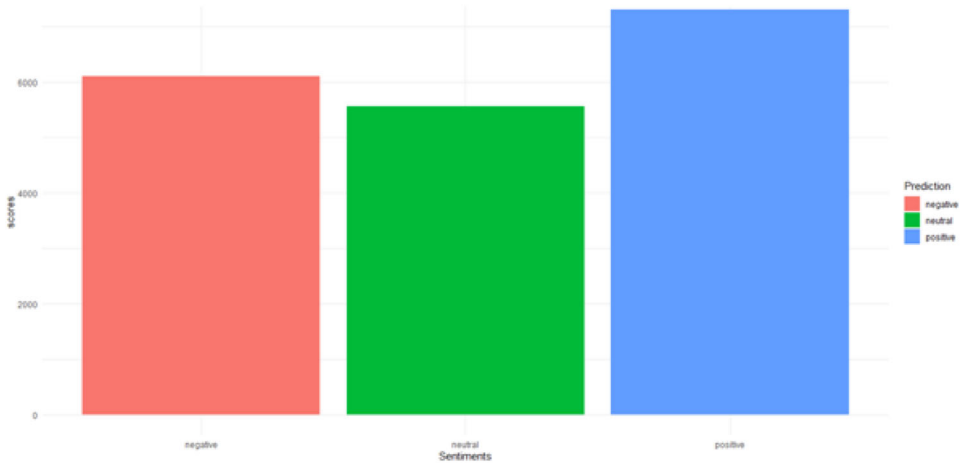

Fig.15 This graph shows the comparisons between all the sentiments of the extracted tweets

approach was used to do the exploratory data analysis while the support vector machine classifier was used to classify the tweets in three categories viz. positive, negative and neutral. The classifier came up with an accuracy of 81%. The works concludes that the people are more inclined towards the act than against it.

References

1. Wankhede Rohit RJ (2018) An approach to sentiment analysis. Int J Sci Res Sci Technol 4(2):1508–1513
2. Chetashri Bhadanea HD (2015) Sentiment analysis: measuring opinions. Elsevier Procedia Comput Sci 45:808–814. https://doi.org/10.1016/j.procs.2015.03.159
3. Devika MD, Sunitha C (2016) Sentiment analysis:a comparative study on different approaches. Procedia Comput Sci 87:44–49. https://doi.org/10.1016/j.procs.2016.05.124
4. Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM. https://doi.org/10.1145/2436256.2436274
5. Singh Y, Bhatia PK (2007) A review of studies on machine learning techniques. Int J Comput Sci Secur (IJCSS) 1(1):70–84
6. Deepali Arora KF (2015) Consumers' sentiment analysis of popular phonebrands and operating system preference using Twitter data: A feasibility study. In: IEEE 29th international conference on advanced information networking and applications, pp 680–686
7. Funk DM (2012) Automatic detection of political opinions in tweets. In: Castro RG, Fensel D, Antoniou G (eds) The Semantic Web: ESWC 2011 Workshops, ser. Lecture notes in computer science, vol 7117. Springer, Berlin, Heidelberg pp 88–89
8. Walaa Medhat AH (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113. https://doi.org/10.1016/j.asej.2014.04.011
9. Hu M, Liu B (2004) Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 168–177. New York, NY, USA, ACM. https://doi.org/10.1145/1014052.1014073
10. Maks I, Vossen P (2012) A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 53(4):680–688. https://doi.org/10.1016/j.dss.2012.05.025
11. Matteo Cristani CT (2018) Making sentiment analysis algorithms scalable. Research Gate
12. Moreno-Ortiza A, Fernández-Cruz J (2015) Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. Procedia Soc Behav Sci 198:330–338. https://doi.org/10.1016/j.sbspro.2015.07.451
13. Siau EA (2014) An approach to sentiment analysis the case of airline quality rating. In: Proceedings of the Pacific Asia conference on information systems (PACIS), pp 363–368. Corpus ID: 27201667
14. Li HXSJ (2011) Sentiment analysis model for hotel reviews based on supervised learning. In: International conference on machine learning and cybernetics (ICMLC), vol 3, pp 950–954. https://doi.org/10.1109/ICMLC.2011.6016866
15. Wang W (2010) Sentiment analysis of online product reviews with semisupervised topic sentiment mixture mode. In: Fuzzy systems and knowledge discovery (FSKD), 2010 seventh international conference vol 5, pp 2385–2389. https://doi.org/10.1109/FSKD.2010.5569528
16. Content W, Analysis S (2012) Applying supervised opinion mining techniques on online user reviews. Informatica Economică 16:81–91
17. Devi MU (2013) Analysis of sentiments using unsupervised learning techniques. In: Information communication and embedded systems (ICICES), pp 241–245
18. Liu GL (2010) A clustering-based approach on sentiment analysis. In: IEEE Intelligent systems and knowledge engineering (ISKE), pp 331–337. https://doi.org/10.1109/ISKE.2010.5680859
19. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B (2011) Combining lexicon-based and learning-based methods for twitter sentiment analysis (Online). Corpus ID: 16228540. http://www.hpl.hp.com/techreports/2011/
20. Sankar H, Subramaniyaswamy V (2017) Investigating sentiment analysis using machine learning approach. In: 2017 international conference on intelligent sustainable systems (ICISS). ICISS, pp 87–92. https://doi.org/10.1109/ISS1.2017.8389293
21. Kawade DR, Oza KS (2017) Sentiment analysis: machine learning approach. Int J Eng Technol 9(3):2183–2186. https://doi.org/10.21817/ijet/2017/v9i3/1709030151
22. Ahmad M, Aftab S, Bashir MS, Hameed N, Ali I, Nawaz Z (2018) SVM optimization for sentiment analysis. Int J Adv Comput Sci Appl 9(4). https://doi.org/10.14569/IJACSA.2018.090455
23. Gopalakrishnan V, Ramaswamy C (2017) Patient opinion mining to analyze drugs satisfaction using supervised learning. Rev c de

Trastornos Alimentarios 15(4):311–319. https://doi.org/10.1016/j.jart.2017.02.005
24. Bhoir P (2015) Sentiment analysis of movie reviews using Lexicon approach. IEEE international conference on computational intelligence and computing research (ICCIC). https://doi.org/10.1109/ICCIC.2015.7435796
25. Vu L (2017) A lexicon-based method for Sentiment Analysis using social network data. International conference on information and knowledge engineering (IKE'17). At: Las Vegas, Nevada, USA. ISBN: 1-60132-463-4
26. Das S, Behera RK (2018) ScienceDirect real-time sentiment of streaming for stock real-time sentiment analysis of Twitter streaming data for stock prediction real-time sentiment analysis of Twitter streaming data for stock prediction. Procedia Comput Sci 132(Iccids):956–964. https://doi.org/10.1016/j.procs.2018.05.111
27. Collomb A, Costea C, Joyeux D, Hasan O, Brunie L (2014) A study and comparison of sentiment analysis methods for reputation evaluation. Rapport de Recherche RR-LIRIS-2014
28. Arora D, Li KF, Neville SW (2015) Consumers sentiment analysis of popular phone brands and operating system preference using Twitter data: a feasibility study. https://doi.org/10.1109/AINA.2015.253
29. Sharef NM, Azmi Murad MA, Mustapha N, Zin HM (2017) The effects of pre-processing strategies in sentiment analysis of online movie reviews. In: AIP conference proceedings 1891, 020089
30. Duwairi R, El-Orfali M (2014) A study of the effects of pre-processing strategies on sentiment analysis for Arabic text. J Inform Sci 40(4):501–513. https://doi.org/10.1177/0165551514534143