.

PREDECTION OF SOCIO ECONOMIC STATUS USING ASPECT BASED LEARNING

Neha Gupta
CSE Department
Guru Nanak Dev Engineering College,
Ludhiana, India

Dr. Kiran Jyoti
IT Department
Guru Nanak Dev Engineering College,
Ludhiana, India

Abstract : Sentiment Analysis is the method of figuring out and categorizing critiques expressed in a bit of textual content, particularly as a way to decide whether the writer's mindset towards a specific subject matter or product is fantastic, bad, or impartial. With the increasing use of micro blogging websites which include twitter, facebook and different social media, each day a whole lot of critiques are being made available on-line. These opinions may be of a product, film or it is able to be an unbiased statement describing a state of affairs. Sentiment analysis is as a result used to classify those statements as a effective one or a poor one. There are various blessings of Sentiment Analysis. It makes the person privy to the numerous tremendous and terrible capabilities of any product. It enables the users in powerful choice making. Furthermore, SA enables agencies to seek feedback from these evaluations and alleviate their products/services anywhere essential. This research work design a framework for analysis the tweets related to multiple social –economic factor.

Index Term – Sentiment analysis, twitter, tweets, aspect based learning, machine learning

I. INTRODUCTION

Sentiments investigation is an interdisciplinary region which involves regular dialect preparing, content examination and computational semantics to distinguish the content notion. As the obstinate writings are frequently an excessive number of for individuals to swim through to settle on a choice, a programmed estimation order technique is important to characterize instant messages into various opinion introduction classes (e.g. positive/negative). The point of notion characterization is to proficiently recognize the feelings communicated as instant messages. Estimation examination can be connected to general information, in spite of the fact that it is more powerful when connected to particular since word implications and assessment may contrast crosswise over areas. A few words bear an unmistakable positive or negative assumption, for example, the words 'great' or 'terrible', while the assessment of some different words relies upon the area. Assumption investigation can be connected for various purposes, for example, political crusades and uproars, be that as it may, it is regularly connected on surveys. The audits can be from various areas including motion pictures, commercials, items, autos, advanced cells, tourism and e-learning. Conclusion investigation subsequently turns into a powerful method for understanding popular suppositions. Most research on slant examination concentrated on interpersonal organizations, item surveys and furthermore on money markets. Itemized investigation of the sentiments and musings accessible in the content is spoken to which permits the extra preparing of the information, for the collection of the conclusions and for acknowledgment of repudiating feelings. Magnificence of the result gave by conclusion mining is basic and critical for the accomplishments and achievement of all succeeding and progressive errands and because of this it turns into a vital and trying for the basic leadership process[1].

In aspect-based sentiment analysis the point is to distinguish the parts of elements and the feeling communicated for every angle. A definitive objective is to have the capacity to create synopses posting every one of the viewpoints and their general polarity[2]. For some application situations archive level audit arrangement is excessively coarse-grained and does not give the coveted data. The majority of the current methodologies depend on word-level investigation of writings and can distinguish just unequivocal articulations of conclusion. The ABSA errand Datasets comprising of client surveys with human-created comments distinguishing the specified parts of the objective substances and the estimation extremity of every viewpoint were given. The analyses were keep running in two areas: eatery and workstation audits.

Machine learning approach utilizes surely understood machine learning systems, for example, terms nearness, recurrence, parts of discourse, supposition words and refutations. While broadly expounding of the terms, following portrayals can be given.

- Terms presence and recurrence: These highlights are singular words or word n-grams and their recurrence checks. It either gives the words double weighting (zero if the word shows up or one assuming generally) or utilizations term recurrence weights to demonstrate the relative significance of highlights [3].
- Parts of Speech (POS): Finding descriptors, as they are vital markers of assessments.
- Opinion words and expressions: These are words usually used to express feelings including great or terrible, as or detest. Then again, a few expressions express suppositions without utilizing conclusion words. For instance: cost me dearly.
- Negations: The presence of negative words may change the feeling introduction like not great is proportional to terrible.

II. RELATED WORK

Bokányi et al. [4] defined that by displaying full scale efficient markers utilizing advanced hints of human exercises on versatile or informal communities, they can give vital bits of knowledge to forms beforehand got to through paper-based studies or surveys as it were. They gathered totaled workday action courses of events of US districts from the standardized number of messages sent in every hour on the online informal organization Twitter. They indicated how region business and joblessness insights are encoded in the day by day cadence of individuals by decaying the action courses of events into a direct blend of two predominant examples. The blending proportion of these examples characterizes a measure for every area, that connects essentially with business (0.46} 0.02) and joblessness rates (– 0.34}0.02). In

.

this manner, the two predominant action examples can be connected to rhythms flagging nearness or absence of customary working hours of people. The examination could give approach creators a superior knowledge into the procedures overseeing business, where issues couldn't just be recognized in view of the quantity of authoritatively enrolled jobless, yet additionally based on the advanced impressions individuals leave on various stages.

Simionescu et al. [5] described that an internet or "big" information are progressively estimating the significant exercises of people, family units, firms and open operators timelily. The data set includes vast quantities of perceptions and grasps adaptable theoretical structures and trial settings. In this manner, web information are to a great degree valuable to think about a wide assortment of human asset issues including guaging, now casting, recognizing medical problems and prosperity, catching the coordinating procedure in different parts of individual life, and estimating complex procedures where customary information have known shortfalls.

Pham et al. [6] discussed the rising interest of angle based assessment investigation, with a specific end goal to decide feeling evaluations and significance degrees of item perspectives. A novel multi-layer engineering for speaking to client audits proposed in this paper. The observations has shown that the general assumption for an item is made from slants out of its angles, and thus every viewpoint has its opinions communicated in related sentences which are likewise the structures from their words. Machine learning procedures including word inserting and compositional vector models, and connected a back-spread calculation in view of slope drop to take in the model. This model thus creates the perspective evaluations and additionally viewpoint weights (i.e. viewpoint significance degrees). The tests directed on an informational index of audits from lodging space & results shown that the proposed technique performs better than the previous.

Pannal et al. [7] explained aspect base sentiment analysis concept in machine learning era. This paper mainly explored sentiment analysis in light of the prepared informational collection to give the positive, negative and unbiased surveys for various items in the advertising scene. In angle based estimation examination (ABSA) the point is to recognize the parts of substances and the notion communicated for every perspective. A definitive objective is to have the capacity to produce rundowns posting every one of the angles and their general extremity. To prepare the application for the given informational collections SVM (bolster vector machine) and ME (Maximum Entropy) arrangement calculations have been utilized. Performance of algorithms is analysed based on precision, recall and Fmeasure.

Cagatay et al. [8] exploit a sentiment classification model on basis of Vote ensemble classifier uses from three individual classifiers: Bagging, Naïve Bayes & Support Vector Machines. Moreover, in bagging they utilized SVM as base classifier. The main focus of this research is to enhance the execution of machine learning classifiers for feeling grouping of Turkish audits and documents. Their experimental results show that multiple classifier system based approaches are much better for sentiment classification of Turkish documents. They performed experiments on three different domains such as book review, movie reviews and shopping reviews. The authors concluded that this approach is not restricted to just one domain and can be extended to several other domains as well.

Mishra et al. [9] defined that the understudies' employability is a noteworthy worry for the establishments offering advanced education and a strategy for early forecast of employability of the understudies is constantly attractive to make auspicious move. It utilizes different grouping strategies of information mining, as Bayesian techniques, Multilayer Perceptrons and Sequential Minimal Optimization (SMO), Ensemble Methods and Decision Trees, to anticipate the employability of Master of Computer Applications (MCA) understudies and discover the calculation which is most appropriate for this issue. For this reason, an informational index is created with the customary parameters like financial conditions, scholarly execution and some extra enthusiastic aptitude parameters. A relative examination infers that J48(a pruned C4. 5 choice tree) most appropriate for employability expectation with 70. 19% precision, simple elucidation and model building time (0. 02Sec) not as much as Random Forest, which has somewhat better expectation precision (71. 30%), higher building time(0. 11) and troublesome elucidation. Further, Empathy, Drive and Stress Management capacities are observed to be the major enthusiastic parameters that influence employability.

Azam et al. [10] aimed of the examination was to analyze the connection between military uses and joblessness rate in the board of chosen SAARC nations to be specific India, Nepal, Pakistan and Sri Lanka over the period extending from 1990 to 2013. Specifically, the examination utilized multivariate system to look at the long-run connection between military uses and joblessness rate by considering the impact of macroeconomic factors to be specific vitality utilization, GDP per capita, Foreign Direct Investment (FDI) net inflows and populace development rate. The experimental outcomes demonstrate that every one of the factors display non-stationary conduct and have long-run connections between them. The consequences of board DOLS demonstrate that military uses support the business rate in the SAARC area, as the evaluated coefficient of military use has a negative and more flexible association with the joblessness rate. Different components i.e., vitality utilization, GDP per capita and approaching FDI essentially diminishes the joblessness rate in the SAARC area. Nonetheless, populace development rate does not display huge relationship with the joblessness rate amid the period under the examination. In a moment relapse mechanical assembly, vitality utilization, and GDP per capita fundamentally diminish the military spending, be that as it may, populace development rate essentially builds the military consumptions in the area.

Sundaram et al. [11] defined that the uniting the after effects of the NSS 66th Round Employment-Unemployment Survey and the Provisional Population Totals of the 2011 Population Census, they analyzed the between play of statistic change, choices on tutoring and cooperation in the work constrain, and the adjustments in the economy in molding the size and structure of business and the resultant effects on work profitability, genuine wages and neediness among those inside and outside the work compel in India over the period 2005– 2010. They additionally offer a concise exchange on a few issues in the estimation of destitution in India. A lazy development in the aggregate number of specialists on UPSS close by an outright lessening in the span of female workforce, and in the quantity of laborers in farming and chaotic assembling are among the key after effects of our examination. On the positive side we locate a solid development in work in the sorted out assembling part and in the number and offer of standard wage pay specialists; and, a solid development in labor efficiency and in genuine wages. They likewise discover a no matter how you look at it diminishment in the extent and check of the working poor in the vicinity of 2005 and 2010. These changes in nature of business must temper our failure with the little development in the span of the workforce.

.

III. PROPOSED WORK

1. Search tweets: Tweets are extracted using twitter api on particular keyword or hash tag.
2. Collect tweets: The extracted tweets are then stored in the array list or database to further processing.
3. Data Pre-processing: The tweets are then processed in order to remove the stop words, punctuation marks, urls, repeated characters etc are removed and filtered tweet is then given to the sentiword net dictionary to get the score.
4. Sentiment Analysis: The pre-processed tweet is then tested by referring the dictionary. The overall score of each tweet is calculated. Then the scale is applied to give the category to each tweet like positive, negative and neutral.
5. Aspect based Learning : The particular aspect based on the text in the tweets is extracted using aspect based learning techniques like TF-IDF etc.
6. Data Mining Techniques: The dataset created using the above framework can be tested using data mining techniques and then comparative analysis can be done between various classification algorithms.

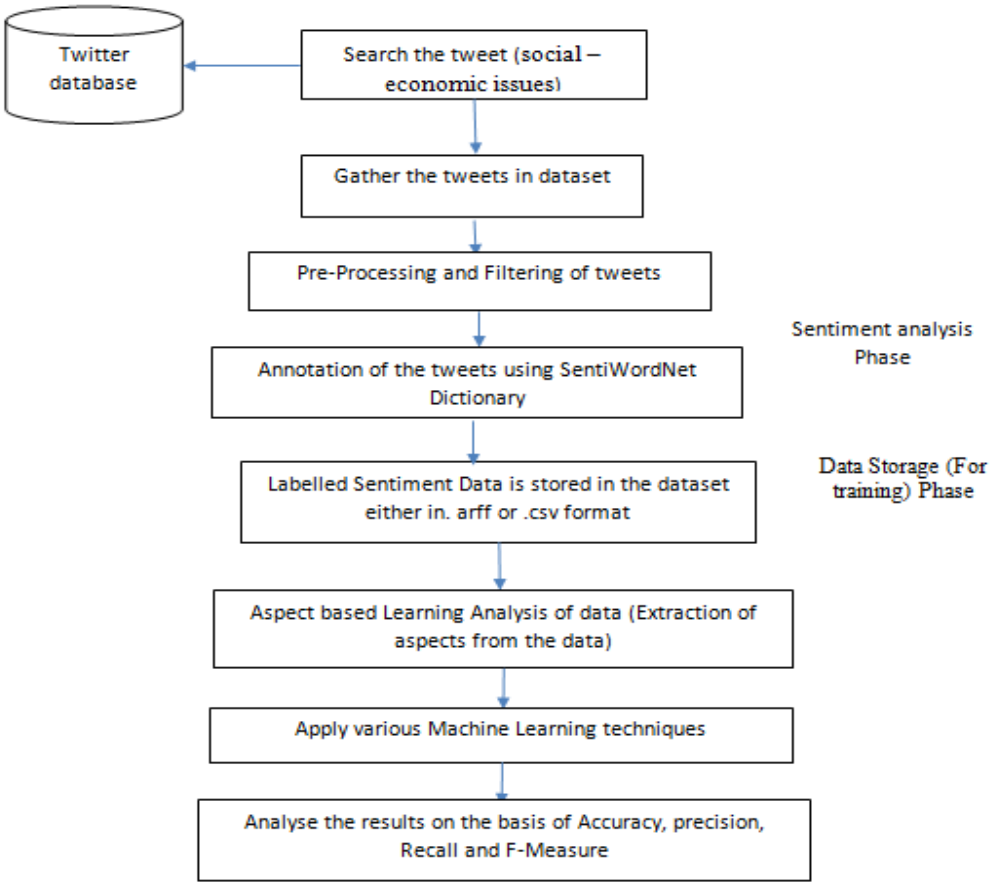

Figure 1: Flowchart of Proposed Technique

IV. EXPERIMENTALL RESULTS

Following are the parameters for evaluation of performance

i. Precision and recall

Precision and recall are the two metrics that are widely for evaluating performance in text mining, and in text analysis field like information retrieval. These parameters are used for measuring exactness and completeness respectively.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

ii. F-measure

F-Measure is the harmonic mean of precision and recall. The value calculated using F-measure is a balance between precision and recall.

$$\text{F measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision+recall}}$$

iii. Accuracy

Accuracy is the common measure for classification performance. Accuracy can be measured as correctly classified **instances** to the total number of **instances**, while error rate uses incorrectly classified instances instead of correctly classified instances.

.

$$\text{Accuracy}=\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{False Positive}+\text{True Negative}+\text{False Negative}}$$

Table 1: Comparison of accuracy of various classifiers for different keyword

Keywords Param-eters	Deficit (import and export)	Educa-tion	GDP	Inco-me	Stock Exchan-ge	Unempl-oyment
Naïve Bayes	65.2174	56.4706	52.8571	61.1111	61.5385	46.8354
KNN	65.2174	54.1176	57.1429	56.9444	69.2308	45.5696
Random Forest	60.8696	56.4706	58.5714	58.3333	76.9231	49.3671
Decision Tree	47.8261	44.7059	45.7143	55.5556	76.9231	40.5063

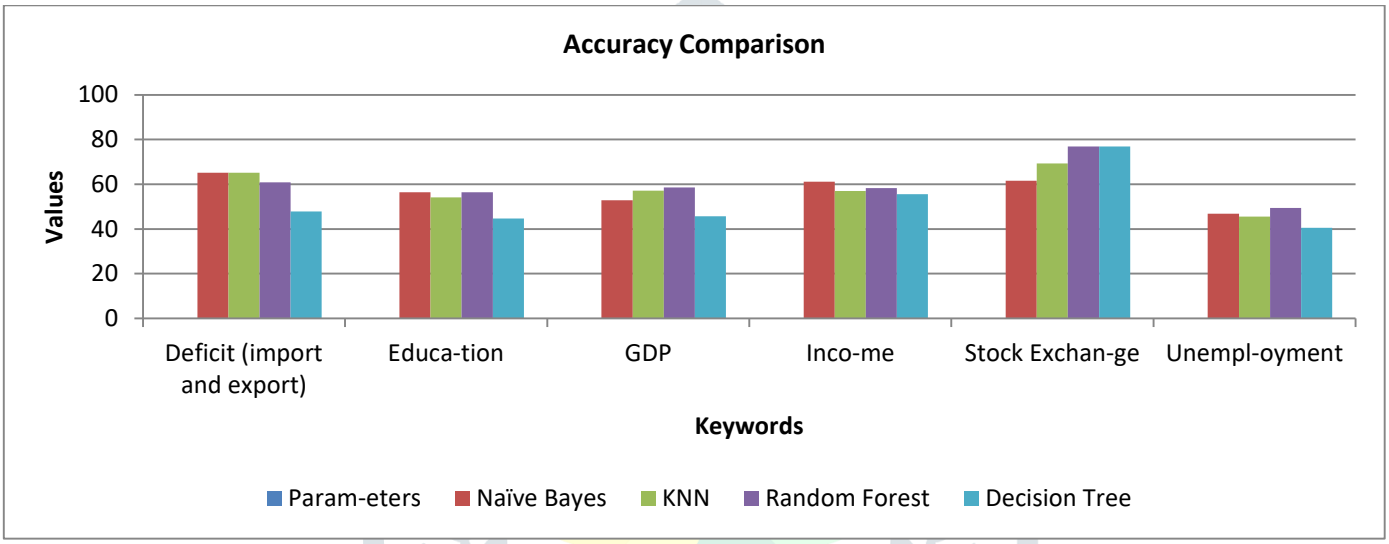

Figure 2: Accuracy comparison of various classifiers for different keywords

Table 2: Comparison of precision of various classifiers for different keywords

Keywords Parameters	Deficit (import and export)	Educa-tion	GDP	Inco-me	Stock Excha-nge	Unempl-oyment
Naïve Bayes	0.618	0.568	0.507	0.612	0.559	0.437
KNN	0.717	0.599	0.607	0.672	0.577	0.392
Random Forest	0.644	0.62	0.602	0.645	0.592	0.456
Decision Tree	0.447	0.285	0.428	0.454	0.592	0.348

.

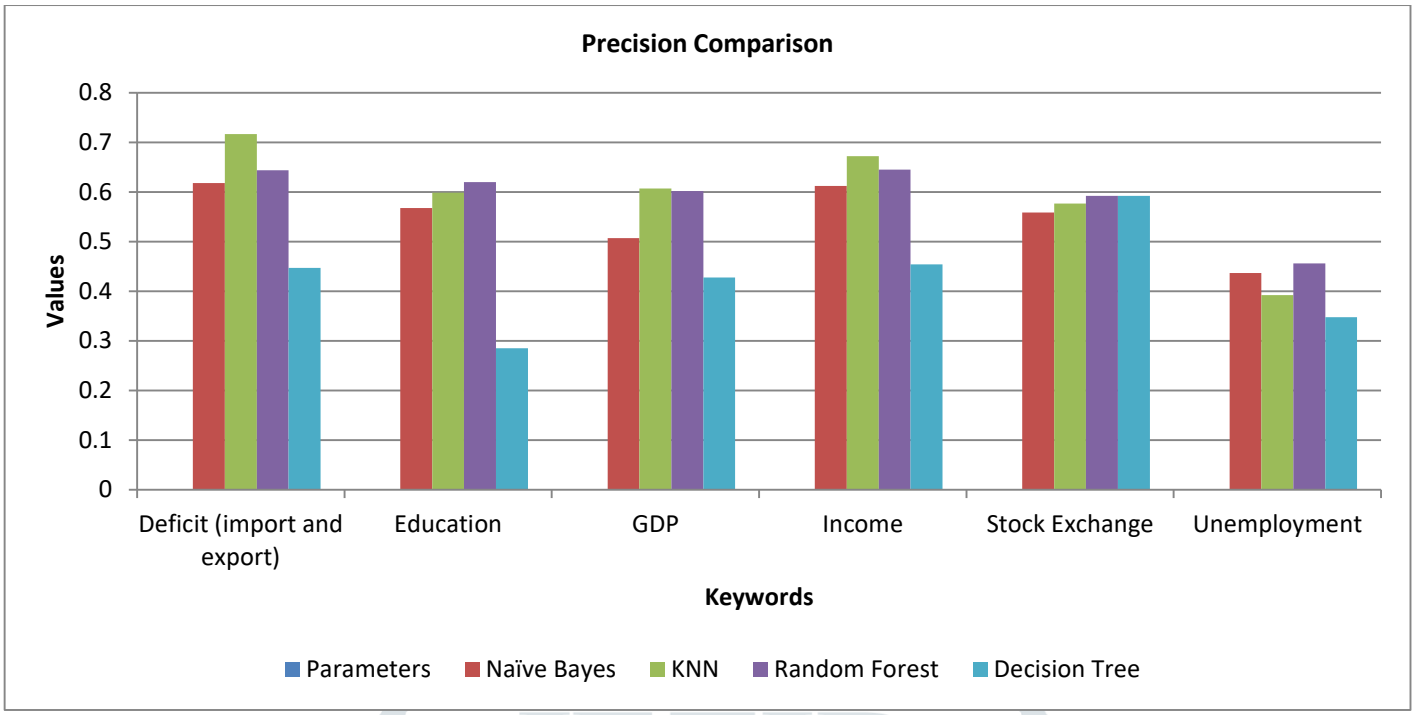

Figure 3: Precision comparison of various classifiers for different keywords

Table 3: Comparison of recall of various classifiers for different keywords

Keywords Parameters	Deficit (import and export)	Educ-ation	GDP	Inco-me	Stock Exchange	Unemp-loyment
Naïve Bayes	0.652	0.565	0.529	0.611	0.615	0.468
KNN	0.652	0.541	0.571	0.569	0.692	0.456
Random Forest	0.609	0.565	0.586	0.583	0.769	0.494
Decision Tree	0.478	0.447	0.457	0.556	0.769	0.405

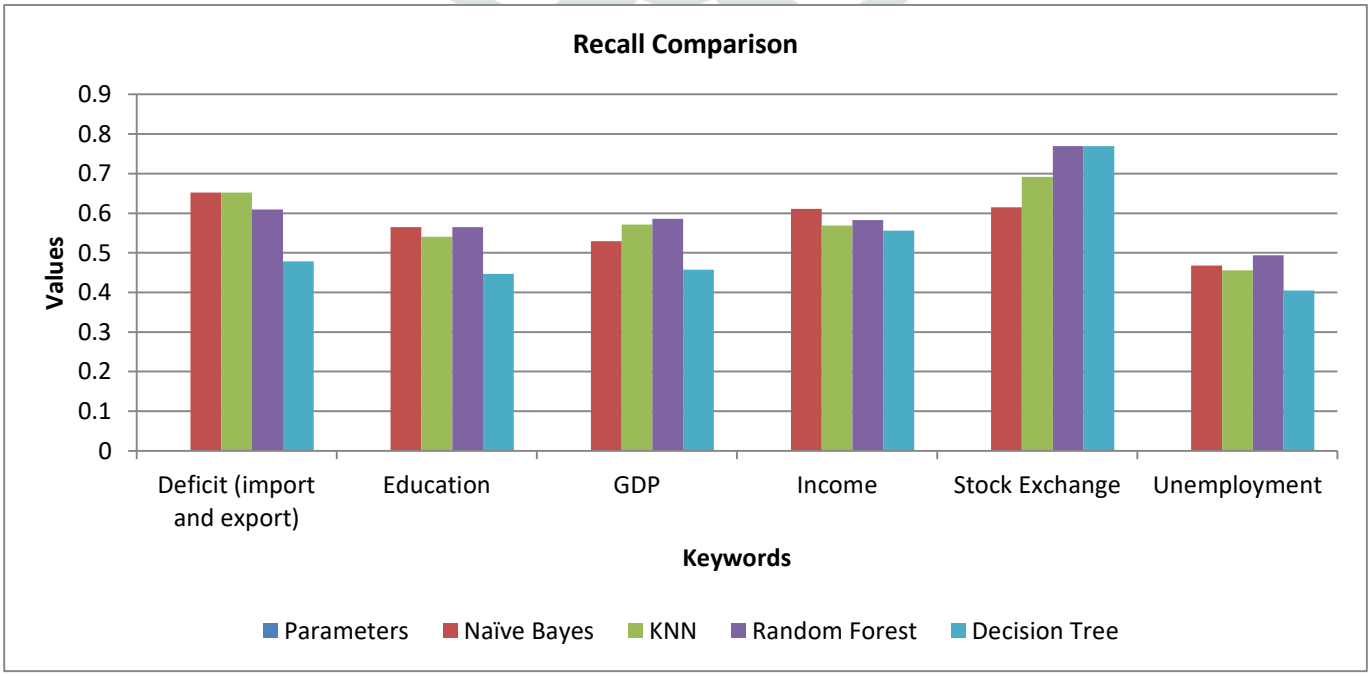

Figure 4: Recall comparison of various classifiers for different keywords

.

Table 4: Comparison of f-measure of various classifiers for different keywords

Keywords Parameters	Deficit (import and export)	Educa-tion	GDP	Inco-me	Stock Exchange	Unemplo-yment
Naïve Bayes	0.631	0.493	0.495	0.567	0.586	0.433
KNN	0.632	0.444	0.526	0.573	0.629	0.389
Random Forest	0.593	0.47	0.541	0.524	0.669	0.439
Decision Tree	0.443	0.327	0.397	0.466	0.669	0.373

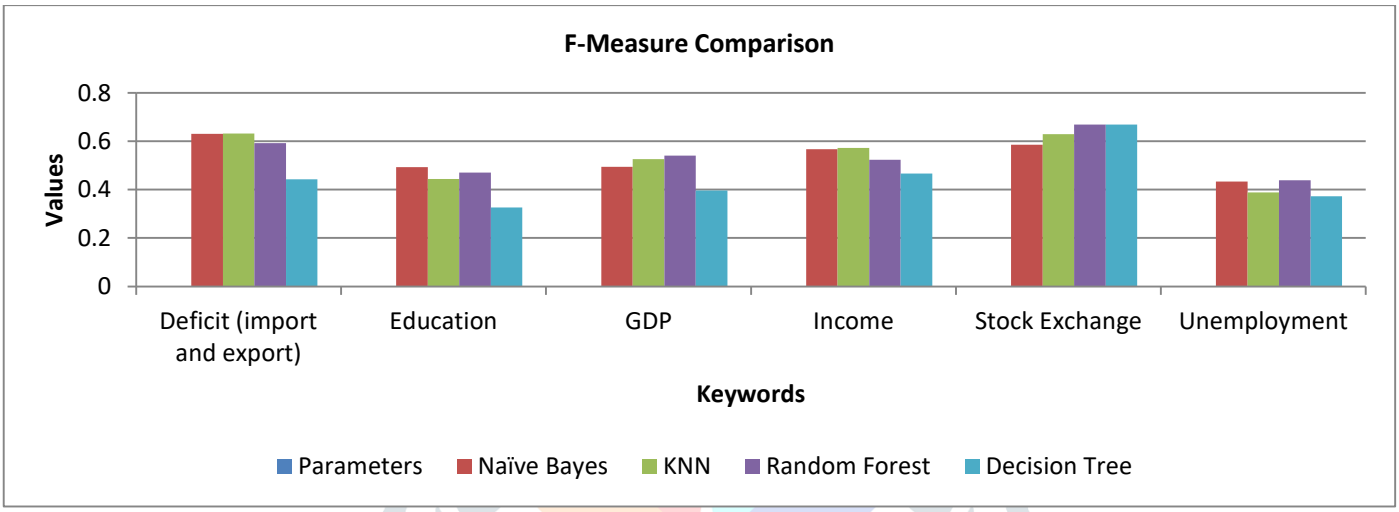

Figure 5: F-Measure comparison of various classifiers for different keywords

V. CONCLUSION

This research work design a framework for analysis the tweets related to multiple social –economic factor. We utilize the information stream unreservedly gave by Twitter through their Application Program Interface, which adds up to every sent message. In this investigation, we center around the piece of the information stream with geolocation data. Aspect based learning technique are used to access the twitter data. Tweets are extracted using twitter api on particular keyword or hash tag. Proposed technique is implemented using NetBeans software. Various parameters like precision, recall, accuracy and f-measure are used to evaluate the performance of this work.

REFERENCES

[1] Dzogang, F., Lesot, M. J., Rifqi, M., andBouchon-Meunier, B. (2010),” Expressions of graduality for sentiments analysis -a survey”, In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on* (pp. 1-7). IEEE.

[2] Saias, Jos, and Ramalho. R. R,(2015), “Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12.”, *International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 767–771

[3] Ahmed Hassan "Sentiment analysis algorithms and applications: A survey" [On- line]. Available: http://ac.elscdn.com/S2090447914000550/1-s2.0 S2090447914000550main.pdf [May. 01, 2016].

[4] Bokányi, E., Lábszki, Z., and Vattay, G. (2017),”Prediction of employment and unemployment rates from Twitter daily rhythms in the US”, *arXiv preprint arXiv:1703.07708*.

[5] Simionescu, M., and Zimmermann, K. F. (2017),” *Big Data and Unemployment Analysis*”, GLO Discussion Paper,vol 81.

[6] Duc-Hong Phamac and Anh-CuongLeb, “Learning multiple layers of knowledge representation for aspect based sentiment analysis”, Data & Knowledge Engineering, Elsevier, 2017

[7] Nipuna Upeka Pannala, Chamira Priya manthi Nawarathna and J.T.K. Jayakody, “Supervised Learning Based Approach to Aspect Based Sentiment Analysis”, IEEE, pp.662-666, 2016

[8] Catal, Cagatay, and Mehmet Nangir. "A sentiment classification model based on multiple classifiers." Applied Soft Computing 50 (2017): 135-141.

[9] Mishra, T., Kumar, D., & Gupta, S. (2016),” Students' Employability Prediction Model through Data Mining”, *International Journal of Applied Engineering Research*, vol11, no. 4,pp: 2275-2282.

[10] Azam, M., Khan, F., Zaman, K., andRasli, A. M. (2016),” Military expenditures and unemployment nexus for selected South Asian countries”, *Social Indicators Research*, vol127, no. 3,pp: 1103-1117.

[11] Sundaram, K. (2017),”Some recent trends in population, employment and poverty in India: An analysis”, In *Perspectives on Economic Development and Policy in India* (pp. 129-167). Springer Singapore.