International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Sentiment Analysis using Optimized Feature Sets in Different Twitter Dataset Domains

Anup Raut, Rahul K. Pandey

Abstract in the last years, the relevance of sentiment analysis is broad and dominant. The capability to take out insights from social data is a tradition that is being extensively accepted by all over globe. Sentiment Analysis has turn out to be a hot-trend issue of technical and marketplace research in the area of Natural Language Processing (NLP) and Machine Learning. Sentiment analysis is enormously useful in social media supervising as it permits us to expand an impression of the wider open estimation behind definite topics. Investigation of social media streams is typically limited to just essential sentiment analysis and count based metrics. This is of the same kind to just scratching the outside and missing out on those elevated value insight that is ahead of you to be discovered. There's a lot of effort to be done, but perfections are being prepared every day. It is a way to appraise on paper or verbal language to settle on if the expression is favorable, unfavorable, or unbiased, and to what level. Today's algorithm-based sentiment analysis tools can touch vast amount of client response constantly and precisely. Balancing with text analytics, sentiment analysis exposes the customer's estimation concerning topics ranging from your goods and services to your position, your advertisements, or even your challengers. These efforts scrutinize the crisis of studying texts, like posts and reviews, uploaded by user on Twitter. The Support Vector Machine (SVM), k-nearest neighbors algorithm (KNN) and proposed optimized feature sets model is offered to progression the tweet features and to recognize the out of sight sentiments from these tweets. These essential concepts when used in combinations become a very significant tool for analyzing millions of variety conversations with human echelon accurateness. The projected optimized feature sets model Sentiment Analysis exercise the assessment metrics of Precision, Recall, F-score, and Accuracy. Also, average measures weighted F1-scores are constructive for categorization of Positive, Negative and Neutral multi-class problems. The running time of the technique is evaluates by accomplishing diverse methods in the same investigational setup consisting a cluster of 8 nodes. Planned optimized feature sets model Sentiment Analysis reachs 82 % accuracy as compare with SVM 78.6 % and KNN 75 %. Further, while analyzing sentiments of tweets we have measured only tweets in English acknowledged by Twitter streaming API.

Keywords **: NLP, SVM, KNN,F1-score.**

I. INTRODUCTION

A progression that done routine mining of attitudes, views, opinions and emotions from speech, text, tweets and database sources through Natural Language Processing (NLP) is called Sentiment analysis. Sentiment analysis engages classifying opinions in text into classes like ""negative" or "positive" or "neutral". It's also referred as subjectivity analysis, appraisal extraction and Sentiment Analysis. The words opinion, view, sentiment and belief are utilized interchangeably but there are differences between them.

- Opinion: A conclusion opens to disagreement (because contradictory experts have diverse opinions)
- View: subjective estimation
- Belief: purposeful receiving and intellectual concur
- Sentiment: view representing ones feelings

Sentiment Analysis is a perception that possibly will contains numerous responsibilities such as sentiment extraction, subjectivity classification and sentiment classification, summarization of opinions or opinion spam detection, among others. It intends to analyze people's attitudes, opinions emotions, sentiments etc. in the direction of fundamentals such as products, individuals, organizations, topics, and services.

Sentiment Analysis removes opinionated text datasets summarizing them in an understandable form for end users. It takes out "positive", "negative" or "neutral opinions" from amorphous data. It involves computational running of view and text subjectivity. Natural Language Processing (NLP) grips text part processing which is distorted to machine format by NLP. Artificial Intelligence (AI) uses NLP provided information applying math's to settle on whether an opinion is positive or negative. A variety of methods exist to conclude a user's view on topics from natural language textual information. Sentiment Analysis tracks the frame of mind of the people relating to a precise product or topic. This provides routine mining of opinions, emotions and sentiments in text and tracks outlook and feelings on the web. Tracking products or brands and formative whether they are positive or negative can be done using the web. Sentiment Analysis is referred by numerous names like "opinion extraction", "sentiment analysis", "subjectivity analysis", "sentiment mining", "affect analysis", "emotion analysis" and "review mining". But, all come under Sentiment Analysis.

A lot of approaches are old in Sentiment Analysis, the most ordinary being "lexicon based and machine learning". In lexicon, straightforward text illustration is a "bag-of-words" approach where documents are considered as an anthology of all words with no help of relations between single words. Sentiment compass reading and words are resources associating to Opinion lexicons. The method's disadvantage is that a word well thought-out positive and negative depends on a state of affairs.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Anup Raut *, PhD Research Student, MUIT, Lucknow.
Dr. Rahul K. Pandey, Research Guide, MUIT, Lucknow

II. LITERATURE SURVEY

In topical years, computer users have begun using social media for communication. Consequently, user interested in issues can be acknowledged by carrying out sentiment analysis. Countless people make use of social networks for business, and the increase of social media has in turn culminated in a rise in sentiment analysis. For instance, businesspeople recognize new opportunities and reinforce reputations on scrutinizing reviews/ ratings available on the web. This not only equivalent the relevant search but also point out relationships among words, so that both words that adjust the sentiment and what the sentiment is about can be precisely acknowledged. Certain scaling configurations are also used in such systems to differentiate positive, negative and neutral sentiments. Sentiment Analysis (SA) errand is to tag individuals' opinions as a variety of classifications, for instance, positive and negative from a known bit of content. Another errand is to decide whether a given content is subjective, communicating the author's opinions, or objective, communicating. These assignments were carried out at a variety of stages of analysis extending from the report level, to the sentence and expression level. An additional assignment is angle removal which began from point of view based sentiment analysis in express level. Every one of these errands is under the sunshade of SA. As of late innumerable, techniques and developments have been projected for the matter of SA in a range of assignments at a mixture of levels. This impression means to categorize SA techniques as a rule, with no absorbed on meticulous level or errand. And moreover to survey the primary research issues in late articles displayed in this field. We found that machine learning-based techniques including governed learning, unsupervised learning and semi-regulated learning techniques, Vocabulary based techniques and half and half techniques are the most succeeding techniques utilized. The unlock issues are that present techniques are as yet out of shape to function admirably in a variety of space; sentiment sort in view of inadequate labeled data is as up till now a testing issue; there is nonattendance of SA look into in dialects other than English; and obtainable techniques are as yet unfit to administer multifaceted sentences that necessitates more than sentiment words and uncomplicated parsing.

Sentiment analysis is to a great degree expensive in social media scrutinizing as it facilitates us to select up an outline of the more widespread public view at the back precise subjects. Social media observing devices like Brandwatch Analytics create that procedure quicker and fewer demanding than at any other instance, because of real-time checking capabilities. The uses of sentiment analysis are wide and successful. The capability to extract bits of knowledge from social data is a training that is as a regulation generally squeezed by organizations in excess of the globe. Moves in sentiment on social media have been come into view to attach with shifts in money markets. The Obama association utilized sentiment analysis to gauge public opinion to understanding declarations and battle messages in front of 2012 presidential election. The aptitude to rapidly understand client states of mind and counter as needs be is something that Expedia Canada browbeaten when they saw that there was a relentless increase in negative feedback to the music utilized as a part of one of their TV adverts.

Sentiment analysis led by the product exposed that the music played on the business had twisted out to be extremely bothering following a variety of airings, and buyers were running to social media to find expression for their dissatisfactions. Two or three weeks following the advert originally broadcast, in excess of portion of online discussion about the battle was negative. As different to chalking up the advert as dissatisfaction, Expedia could address the negative sentiment in an energetic and self-knowing pathway via ventilation another description of the advert which featured the culpable violin being damaged.

Document level sentiment analysis (Danushka Bollegala et al. 2013) techniques described whether the whole document describes a positive or negative opinion. The main principle of document-level sentiment analysis is to categorize a document to identify its on the whole subjectivity and/or sentiments articulated in it. In such a casing, it is possible to apply both supervised and unsupervised learning methods to perform document-level psychoanalysis. From the literature review carry out in this work, it is seen that the chief benefit of performing opinion mining using sentiment analysis is that it settle on the overall opinion of a meticulous creation from a document and the disadvantage is that determining the estimation about the characteristic of a particular product is an unfeasible task in document stage analysis.

Sentence-level sentiment analysis categorizes each sentence based on its sentiment polarization (positive, negative, neutral) about a topic (Danushka Bollegala et al. 2013). This level of analysis depends completely on bias categorization, which classifies sentences as objective and subjective. The objective sentences portray the suitable information comfortable from the sentences. On the other hand, the subjective sentences are used to explain the subjective views and the equivalent opinions. Such subjectivity classification is measured a chief benefit of sentiment-level analysis. Classification methods appropriate to document-level analysis are functional to sentence-level analysis.

III. METHODOLOGY

Figure 1 demonstrates the working of Sentiment Analysis in the extraction phase where social media acts as a data source. Data from social media like twitter is updated frequently. So, it gives the feeling of real time representation of sentiments. To obtain data on run-time an internet both known as web crawler is used. This browses through the World Wide Web in an organized manner to index web pages. In pre-processing, the extracted data is cleaned as it has large amounts of noise before sending the text for analysis. Extracted text has grammatical errors as text is of limited length. In analysis, sentiments in data, number of repetitions observed in tweets and their location is analyzed. In Knowledge Discovery phase, to find opinions of people regarding any particular occurrence, it is essential to store data related to the event. Once sentiments polarity is known it generates statistical graphs and charts.

There are four groups used in sentiment analysis:

1) Syntactic feature: It employs word or Part Of Speech (POS) ticket, Ngrams, punctuation, or phrase patterns one among all. The authors identify that "phrase patterns like "n+aj" (positive adjective) symbolize positive sentiment direction, at the same time as "n+dj" (negative adjective) articulate negative sentiment".

2) Semantic feature: Semantic feature concentrate on relation between signifiers like phrases, words, and score-base method classify these stands on title figure of encompass positive or negative semantic feature.

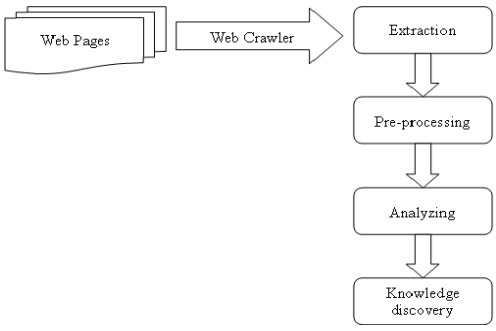

Figure 1Working of Sentiment Analysis

3) Link base feature: Using links present among relation and relations, link base samples are classified.
4) Stylistic feature: Artist's use this to go by a message to us. Use of symbolism: This is where a writer or an artist employs a symbol to illustrate, symbolize or differentiate a person, thing or place.
The planned system encloses various phases of growth. A dataset is fashioned using twitter posts of movie reviews. As we know that tweets hold jargon words and misspelling. So, we carry out a sentence level sentiment investigation on tweets. This is completed in three stages. In a first stage preprocessing is prepared. Then feature vector is formed by means of applicable features. Lastly, using dissimilar classifiers, tweets are classified into positive, negative and neutral classes.

Algorithm

Input: Training and Testing Datasets
Step 1: Browse Training dataset
Step 2: Delete stopwords from training dataset.
Step 3: Tokenization of training dataset.
Step 4: Stemming of training dataset.
Step 5: Browse Testing dataset
Step 6: Delete stopwords from testing dataset.
Step 7: Tokenization of testing dataset.
Step 8: Stemming of testing dataset.
Step 9: Extract special keywords.
Step 10: Extract positive and negative keywords.
Step 11: Extract positive and negative tags.
Step 12: Form 8-features based vector
Step 13: Perform Classification
Step 14: Form BoW vector.
Step 15: Form Hybrid Vector by fusion of vector 1 and 2.
Step 16: Measure Precision, Recall and Accuracy Performance for existing and proposed method.

IV. RESULTS

Efficiency of sentiment analysis submission is calculated through experiments in the form of test data. For binary classifiers there are different metrics to measure the performance as following.

Precision:

In binary classification precision (also called positive predictive value) is the portion of the numeral of true forecasted documents to total number of predicted documents. In the other words precision is quotient of TP and total of TP and FP as following

$$Precision = \frac{TP}{TP + FP}$$

Generally, far above the ground precision means that the classifiers execute well to properly classify or predict the true polarity of test data set.

Recall:

Recall or sensitivity is defined as the number of true predicted document to all existing documents including true and false predicted as follow

$$Recall = \frac{TP}{TP + FN}$$

High recall means that the classifier can predict correctly most of the documents.

Accuracy:

Another statistical metric of how well a binary classifier performs correctly on test data is accuracy. To calculate the accuracy equally true positive and true negative values among the whole number of scrutinize cases are measured. The accuracy formula is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The classification algorithm joining KNN and SVM is given below. Confusion matrix for KNN and SVM is in use from weka by opening the Train Features Set and Test Features Set there directly and is provided below

Table 1: Confusion matrix for KNN

		Predicted		
	Class	Negative	Neutral	Positive
Actual	Negative	77	19	17
	Neutral	9	46	16
	Positive	21	14	79

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 2: Confusion Matrix for SVM

		Predicted		
	Class	Negative	Neutral	Positive
Actual	Negative	77	14	22
	Neutral	11	47	43
	Positive	16	20	78

Confusion matrix for projected approach is considered by the analysis results and with the help of confusion matrix precision, recall and f-measure is calculated

Analysis results show True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) for each sentiment class. Thus confusion matrix is derived below:

Table 3: Confusion Matrix for proposed work

		Predicted		
	Class	Negative	Neutral	Positive
	Negative	77	12	24
Actual	Neutral	0	48	23
	Positive	0	12	102

With the help of confusion matrices Accuracy, Precision, Recall and F-measure for positive, negative and neutral classes are calculated and are also compared for the 3 approaches used above

Table 4 Accuracy of each technique

Method	Accuracy
KNN	75%
SVM	78.60%
Proposed	82.00%

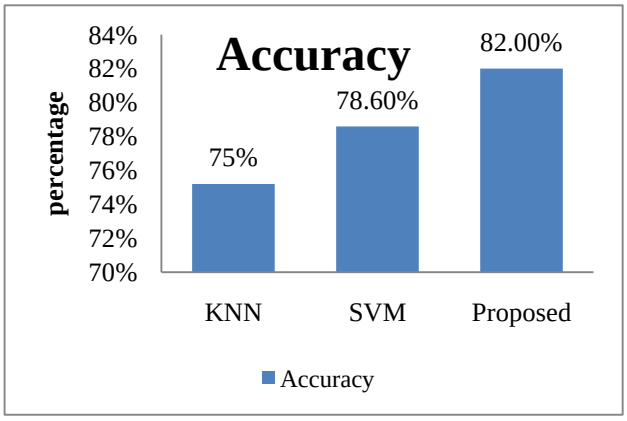

Figure 2: Showing Overall Accuracy for 3 Approaches

Now here we discuss the precision analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In Figure 3 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

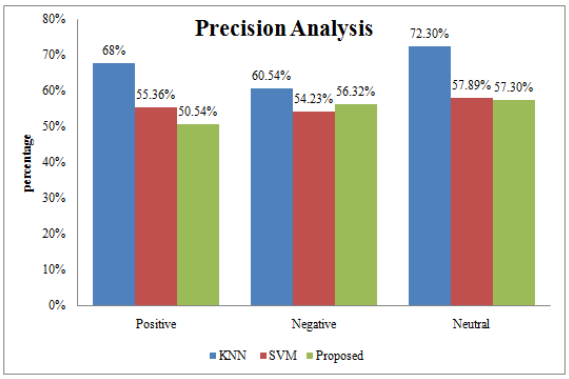

Figure 3 Precision analysis

Now here we discuss the recall analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In Figure 4 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

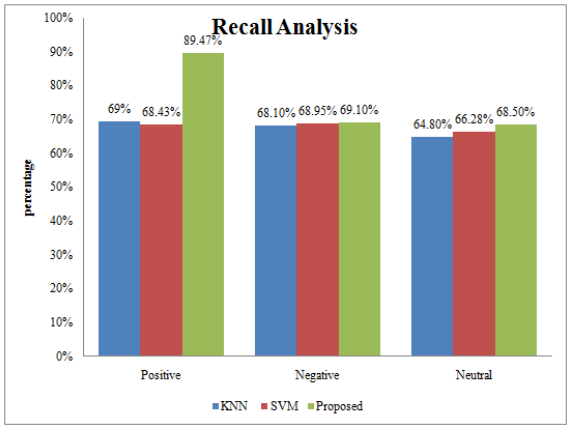

Figure 4 Recall analysis

Now here we discuss the F-score analysis show blow table show comparative value of each categories of tweet like positive, negative and neutral with respect to the each technique. In Figure 5 x axis show tweet category with classification technique and y axis shows the percentage value of precision.

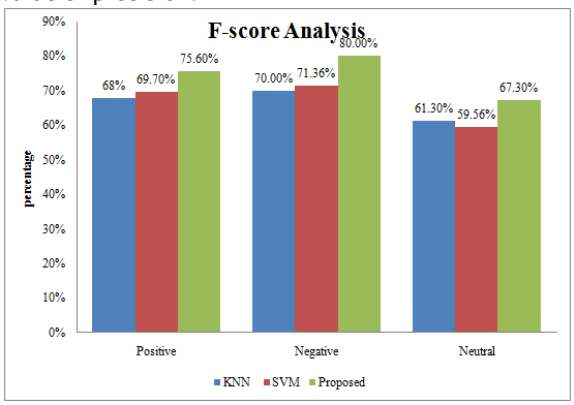

Figure 5 F-score analysis

In this section we first examine the performance of our sentiment analyzer by comparing with other methods. We also compare the computation time taken by the algorithms in the parallel python environment. Finally sentiment analysis on real time tweets are done and sentiment based cliques were generated for further analysis.

V. CONCLUSION AND FUTURE WORK

We presented the important of sentiment analysis, and discussed the different techniques of sentiment analysis proposed since in last decade. The detailed comparative analysis of such techniques is presented in this project report. After the study of existing methods, their problems have been openly discussed. Finally, the proposed methodology is presented with wide set of contributions. For future work, we will work on contributions, designing approaches for solving the current research problems. We create a sentiment classifier for twitter using features sets. We also investigate the relevance of using a double step classifier and negation detection for the purpose of sentiment analysis. Our baseline classifier that uses just the unigrams achieves an accuracy of around 80.00%. Accuracy of the classifier increases if we use negation detection or introduce bigrams and trigrams. Thus we can conclude that both Negation Detection and higher order n-grams are useful for the purpose of text classification. However, if we use both n-grams and negation detection, the accuracy falls marginally. We also note that Single step classifiers outperform double step classifiers. In general, Naive Bayes Classifier performs better than Maximum Entropy Classifier. We have also introduced a new feature reduction technique that only reduces the number of features drastically but also enhances F-score when used with threshold.

Future work is doing same job with for Hindi tweets There are many users on Twitter that use primarily Hindi language. The approach discussed here can be used to create a Hindi language sentiment classifier.

REFERENCES

1. Go, Alec, Richa Bhayani, and Lei Huang "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.
2. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010
3. Spencer, James, and Gulden Uchyigit. "Sentimentor: Sentiment analysis of twitter data." Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2012.
4. Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!" Icwsm 11 (2011): 538-541.
5. Narr, Sascha, Michael Hulfenhaus and Sahin Albayrak "Language-independent twitter sentiment analysis" Knowledge Discovery and Machine Learning (KDML), LWA (2012): 12-14.

A. Celikyilmaz, D. Hakkani-Tur and J. Feng “Probabilistic model-based sentiment analysis of twitter messages,” in Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 79–84, IEEE, 2010.

6. Y. Wu and F. Ren, “Learning sentimental influence in twitter,” in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

A. Pak and P. Paroubek “Twitter as a corpus for sentiment analysis and opinion mining,” in Proceedings of LREC, vol. 2010

7. R. Xia, C. Zong, and S. Li “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.
8. Neethu M.S., Rajasree R "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India, IEEE.
9. Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, MANUSCRIPT, 2014.
10. Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi and Tao Li" Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE Transactions on Knowledge and Data Engineering, 2015.
11. R. Suresh Ramanujam, R. Nancyamala, Nivedha, "SENTIMENT ANALYSIS USING BIG DATA", 2015 International Conference On Computation Of Power, Energy, Information And Communication
12. Saif, Hassan, Yulan He, and Harith Alani "Semantic sentiment analysis of twitter" International Semantic Web Conference. Springer Berlin Heidelberg, 2012.
13. Mukwazvure, Addlight, and K. P. Supreethi "A hybrid approach to sentiment analysis of news comments." Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on. IEEE, 2015
14. [17] Khan, Jawad and Byeong Soo Jeong "Summarizing customer review based on product feature and opinion." Machine Learning and Cybernetics (ICMLC), 2016 International Conference on. IEEE, 2016