iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34313

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Twitter Sentimental Analysis using Machine Learning

Prof. Shikha Malik[1], Sakshi Palaw[2], Saloni Palav[3], Parth Sonpal[4], Anuj Poojary[5]
[1, 2, 3, 4, 5]*Atharva College of Engineering, Department of Electronics and Telecommunication, Mumbai University, Mumbai, India*

Abstract: Nowadays, social networking sites are at the boom, therefore great deal of knowledge is generated. Millions of people are sharing their views daily on twitter. This paper contributes to the sentiment analysis for customers' review classification which is useful to analyze the knowledge within the sort of the amount of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of those two. For this we first pre-processed the data set, because of the unstructured tweets; the technique of pre-processing the raw data is Removal of punctuation, Removal of common words (Stop words), Normalization of Words and lastly vectorisation. thereafter applied machine learning based classification algorithms namely: 1. Naïve Bayes Algorithm 2. Logistic Regression Algorithm 3. Decision tree Algorithm 4. Random Forest Algorithm. The Naive Bayes algorithm was for simple classification and while Logistic Regression, Decision Tree and Random Forest Algorithm was used for Standardization. We also conclude that through this , that Logistic regression algorithm provides highest accuracy of 95% whereas Decision Tree and Logistic Regression give accuracy of 93% and 94% respectively.

I. INTRODUCTION

Nowadays, the age of Technology has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media ,etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, etc. to express their emotions, opinion and share each and every views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data within the sort of tweets, status updates, blog posts, comments, reviews, etc. Moreover, social media provides a chance for businesses by giving a platform to connect with their customers for advertising. People mostly depend upon user generated content over online to an excellent extent for decision making. For e.g. if someone wants to buy a product or wants to use any service, then they firstly search its reviews online, discuss about the same social media before taking a decision. the quantity of content generated by users is just too vast for a normal user to analyze. So there is a need to automate this, various sentiment analysis techniques are widely used. Sentiment analysis tells user whether the information about the product is satisfactory or not before they pip out . Marketers use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements. In this paper we have implemented the Naïve bayes algorithm. For example, recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative opinions about those items by making use of Sentimental Analysis.

II. LITERATURE SURVEY

Sentiment analysis is the process of using natural language processing, analytic thinking, and applied mathematics to analyze customer sentiment. The best businesses understand the sentiment of their customers; what people are saying, how they're saying it, and what they mean. Customer sentiment are often found in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the realm of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace. Sentimental analysis is predominantly used for brand monitoring, customer service and market research and analysis.

A. Existing System

The Naive Bayes algorithm is one of the most widely used algorithms for performing sentimental analysis. It is undoubtedly useful to a certain extend but in our case with handling the twitter data set, it turned out to give us slightly low accuracy than we were expecting. Also this existing system works on the data set which is constrained to particular topic hence has a limited scope. Further, accuracy in this case plays an important role as it is a major factor of the output of our system and cannot be ignored.

In sentimental analysis if the system is not accurate enough it may lead to discrepancies which may further affect the model .Hence it becomes extremely important for the model to be up to the mark with the standard accuracy rates.

Tan and Zhang (2008) compared different feature selection (Mutual Information, Information Gain, Chi Squared and Document Frequency) and learning methods (centroid classifier, K-nearest neighbor, window classifier, Naïve Bayes and SVM) in extracting opinion from Chinese documents. Information Gain performed the best for sentimental terms extraction and SVM exhibited the best performance for sentiment classification. In a study by Dasgupta and Ng (2009), a weakly-supervised sentiment classification algorithm was proposed.[13]. This study showed opinions on one topic with different graph structures. Chen et al. (2006) presented term clusters with polarity information, words coordination, and decision tree based review representation. Boiy et al. (2007) performed experiments using SVM, naive Bayes multinomial and maximum entropy on movie and car brands review. The best accuracy of the study was up to 90.25% . [14] Boiy et al. (2007) performed experiments using SVM, naive Bayes multinomial and maximum entropy on movie and car brands review. The best accuracy of the study was up to 90.25% . The similar movie reviews dataset was used for experiment by Annett and Kondrak (2008). Different approaches like SVM, NB, alternating decision tree and lexical (WordNet) based approach were utilized for sentiment analysis and greater than 75% accuracy was achieved.[15].In 2015, Ms. Umaa Ramakrishnan and Ms. Rashmi Shankar, and Mr. Ganesha K have recommended an approach, which gives the users the full privilege to make use of any of the classifiers and generate the overall result. By utilizing the OAuth Tool, they have obtained the private and public key from Twitter to access the Tweets and obtain the data sets. The methods such as Bayesian Network, Maximum Entropy Classifier, Naïve-Bayes Classifier, SVM (Support Vector Machine), Decision Tree Classifiers, were used for classifying the sentiment analysis.

These methodologies are useful in deriving a resultant which shows the number of positive, negative and neutral tweets generated for a particular topic that is searched for. The algorithm produced accurate results but takes too long to execute them.[16].This study mainly uses the context of German federal elections to investigate whether Twitter is used for the political deliberation and whether the online message actually correspond to the offline election results. Here, the LIWC(Linguistic Inquiry and Word Count) text analysis software is used for analysis of over 100,000 messages which contain the reference of either a political party or a politician. The use of micro blogging message content is used as a valid indicator of political sentiment .The data set used contained 104,003 political tweets which were published in Twitter's pubic message board. Then the tweets that contained the names of the major 6 parties and prominent politicians were collected. Hereby they had 70,000 tweets mentioning 6 major parties and 35,000 tweets referring to their politicians.

Then the LIWC text analysis technique was used to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. This mainly focused on 12 dimensions in order to profile political sentiment: Future orientation, past orientation, positive emotions, negative emotions, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money. In the final results it was found that Twitter is indeed used as a platform for political deliberation.

The mere number of tweets reflects voter preferences and comes close to traditional election polls, while the sentiment of Twitter messages closely corresponds to political programs, candidate profiles, and evidence from the media coverage of the campaign trail.[6]

B. Proposed System

In this paper we have implemented three more machine learning based classification algorithms along with the Naive Bayes algorithm, which are:

1) Logistic Regression Algorithm
2) Decision Tree Algorithm
3) Random Forest Algorithm

We use these algorithms after the pre-processing is done. The Naive Bayes algorithm is used for simple classification while the remaining three algorithms are used for the process of standardization.

Logistic regression models the possibilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. The goal of using a Decision Tree is to produce a training model which will be used predict the category or value of the target variable by learning simple decision rules inferred from prior data(training data).Random forest is a supervised learning algorithm.

The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. Therefore, combining the advantages of all these algorithms we designed, trained and tested a model which was giving us an output and accuracy much better than the existing system. The individual algorithms used, give us and accuracy as follows(rounded up):

a) Logistic Regression Algorithm: 95%
b) Decision Tree Algorithm: 93%
c) Random Forest Algorithm: 94%

Thus our aim of obtaining the standardized accuracy is obtained through these changes hence the proposed system proves to better than the existing one. Also these three algorithms have better accuracy rate than the Naive Bayes algorithm.

III. METHODOLOGY

A. Data Collection

For our model input, we have chosen the twitter tweets of an user which he tweeted in the past couple of years. We start by first importing important libraries(ex. :pandas) and many such libraries available online. The data set chosen is a data set available online in one of the many online libraries. This particular data set contains about 1400+ tweets which we have used as to train the data set at first.

id	label	tweet
1	0	@user when a father is dysfunctional and is so self
2	0	@user @user thanks for #lyft credit i can't use cau
3	0	bihday your majesty
4	0	#model i love u take with u all the time in urð□□±
5	0	factsguide: society now #motivation
6	0	[2/2] huge fan fare and big talking before they leav
7	0	@user camping tomorrow @user @user @user @
8	0	the next school year is the year for exams.ð□□¯ ca
9	0	we won!!! love the land!!! #allin #cavs #champions
10	0	@user @user welcome here ! i'm it's so #gr8 !

As you can see in the above data set, we have three features in here which are: ID , LABEL and TWEET .Here we have a training data set which contains tweets labelled as “1” or “0” respectively.

LABEL “0”: Positive sentiment
LABEL “1”: Negative sentiment

Then we read the data using Pandas.

Here we are done with the data collection part and now we will move forward pre-process the obtained data through collection so that we can further work on that data with ease.

B. Data Pre-Processing

After acquiring the data which we will actual work on, still some perforation is to be done to separate out only the most important parts needed for actually analyzing the complete data set. This makes it easier for us to develop a Natural Language Processing (NLP) model .This whole process is divided into three parts:

1) *Removal Of Punctuations:* Punctuations are going to be always a disturbance in NLP specially hashtags and “@” play a serious role in tweets. The overlooked punctuations and other unusual notations are going to be removed within the upcoming preprocessing techniques.
2) *Removal Of Commonly Used Words (Stop Words):* In an NLP task the stop words (most common words e.g: is, are, have) do not make sense in learning because they don’t have connections with sentiments. So removing them saves the computational power also as increases the accuracy of the model.
3) *Normalization Of Words:* All the weird symbols and therefore the numerical values were removed and returned a pure list with words as shown above. But still we may encounter multiple representations of an equivalent word. (e.g: play, plays, played, playing) albeit the words are different they carry us an equivalent meaning because the normal word “play”. So we need to do Lexicon Normalization approach to solve this issue. Lexical normalization is the task of translating/transforming a non standard text to a standard register.

This completes the task of pre-processing the raw data that is ; our test part is now completed. Here filtering out the data was required so that we get an highly accurate output for our input.

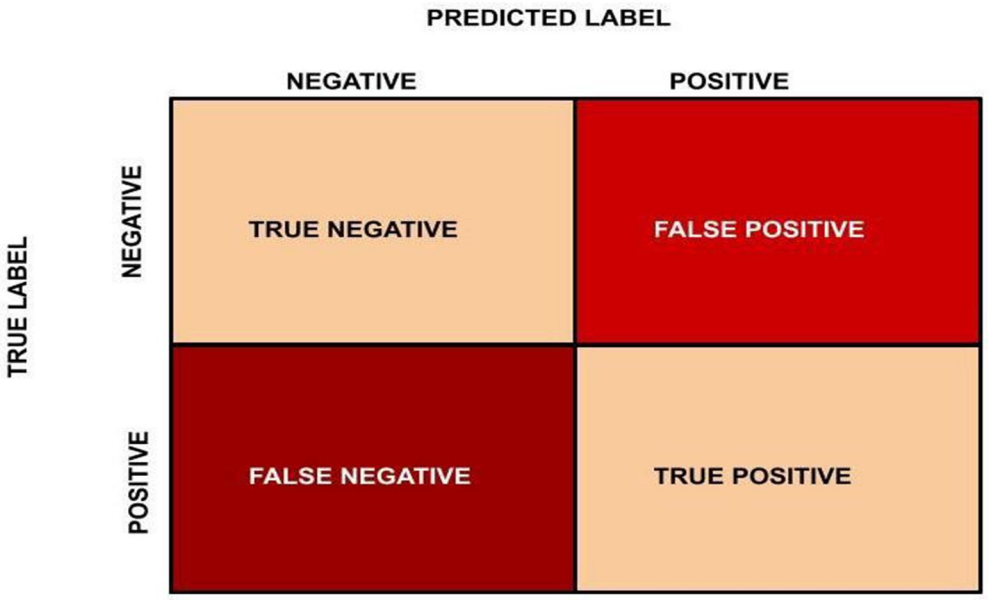
C. Machine Learning Algorithms

Naive Bayes algorithm

1) Naive Bayes may be a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. Here we first look towards the classification report , then we find the precision, then recall and lastly the f1 score.

2) A Classification report is employed to live the standard of predictions from a classification algorithm. How many predictions are True and the way many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are wont to predict the metrics of a classification report .The metrics are calculated by using true and false positives, true and false negatives. Positive and negative during this case are generic names for the anticipated classes.

3) Confusion matrix C is a square matrix where Cij represents the number of data instances which are known to be in group I(true label) and predicted to be in group j (predicted label).



The same procedure is repeated with the other three algorithms .

Here the accuracy we get in the remaining three algorithms is :

a) Logistic Regression Algorithm: 95%

b) Decision Tree Algorithm: 93%

c) Random Forest Algorithm: 94%

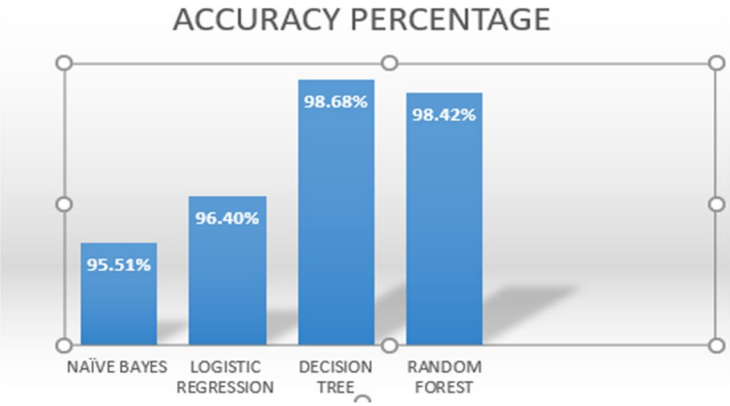
Thus we have now obtained our desired output. All we needed was accuracy as high as possible and we have now achieved it in our model. hence using this methodology we can very easily analyze the sentiments of datasets with high accuracy and precision.

IV. RESULT AND DISCUSSION

```
*NAIVE BAYES CLASSIFIER*
-------Classification Report---------
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      6200
           1       0.40      0.92      0.55       193

    accuracy                           0.96      6393
   macro avg       0.70      0.94      0.77      6393
weighted avg       0.98      0.96      0.96      6393

--------Confusion Matrix-------------
[[5928  272]
 [  15  178]]
-----------Accuracy of the Model------------
0.9551071484436102
```

```
Models 0
Classification Report
              precision    recall  f1-score   support

           0       0.97      1.00      0.98      5933
           1       0.90      0.56      0.69       460

    accuracy                           0.96      6393
   macro avg       0.93      0.78      0.84      6393
weighted avg       0.96      0.96      0.96      6393

Accuracy
0.9640231503206632

Models 1
Classification Report
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5933
           1       0.92      0.90      0.91       460

    accuracy                           0.99      6393
   macro avg       0.95      0.95      0.95      6393
weighted avg       0.99      0.99      0.99      6393

Accuracy
0.9868606288127639

Models 2
Classification Report
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      5933
           1       0.95      0.83      0.88       460

    accuracy                           0.98      6393
   macro avg       0.97      0.91      0.94      6393
weighted avg       0.98      0.98      0.98      6393

Accuracy
0.9842014703582043
```

1) *Model 0 :* Logistic Regression
2) *Model 1 :* Decision Tree
3) *Model 2:* Random forest classifier.



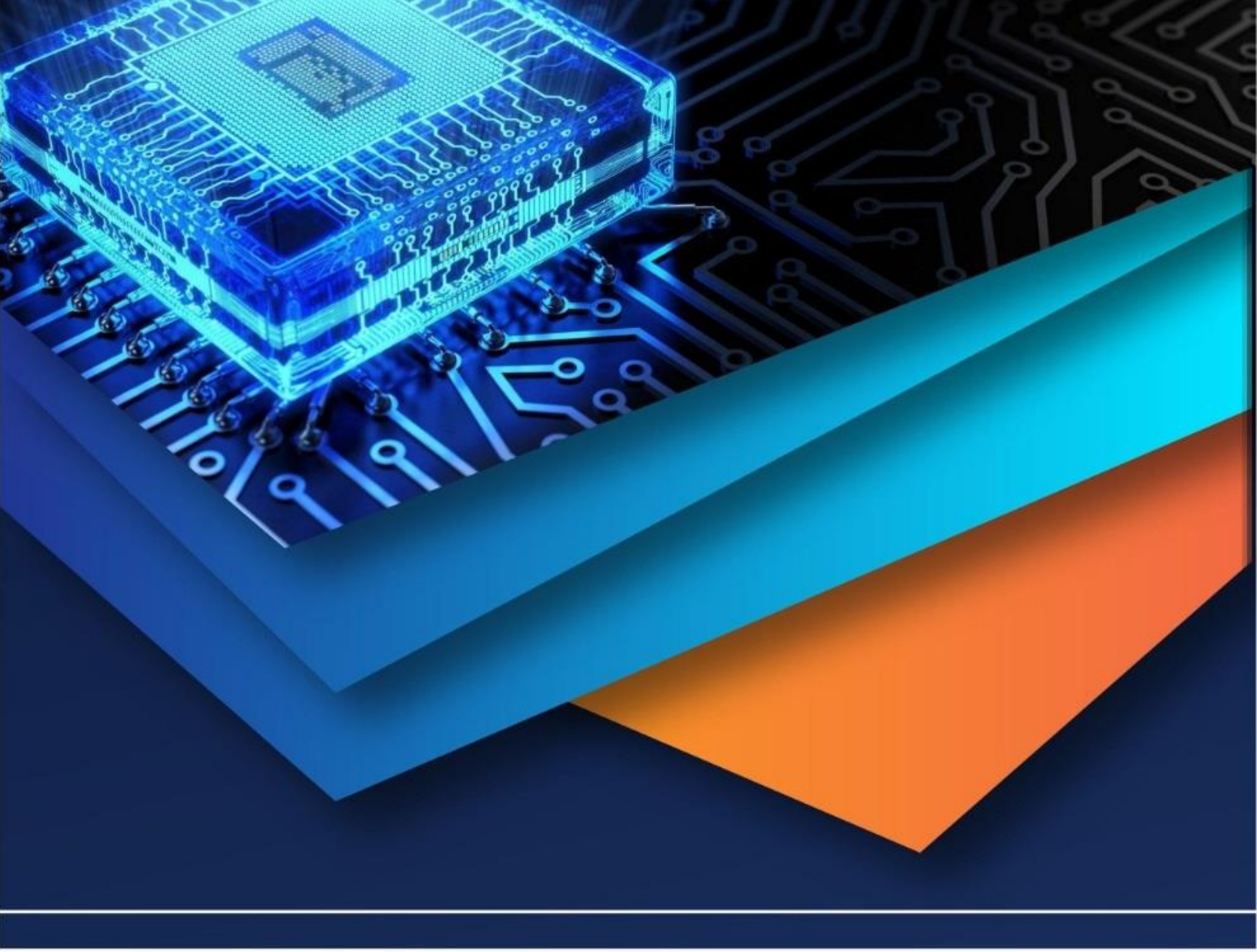
Thus we have now obtained our desired output.All we needed was accuracy as high as possible and we have now achieved it in our model.hence using this methodology we can very easily analyze the sentiments of datasets with high accuracy and precision.

V. CONCLUSION

Sentimental analysis is developed to analyze customer sentiment which is an important factor as it helps businesses quickly understand the overall opinions of their customers. This paper has discussed techniques for preprocessing the dataset and retrieval of tweets through Twitter. This program has used the machine learning approach which is very important to classify the text or reviews from the dataset, together with the natural language processing technique. We have studied the Naive Bayes algorithm for the text classification which is used to find out the polarity of the tweet. Also, we have combined the Logistic regressions algorithm, Decision tree Algorithm, and Random forest Algorithm to obtain standardized accuracy from the system. The program has categorized the sentiments into positive and negative, which is represented through a pie chart. Thus it will assist the companies and the organizations the mindset of the users and customers and can accordingly tailor their products and needs of their consumers. In future work, we will explore even richer linguistic, delve deeper beyond the concept of the number of likes, comments, and shares in a post, and having a better understanding of the emotions and sentiments of the consumers.

REFERENCES

[1] Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and theOMG!", (Vol.5). International AAAI, 2011.
[2] A.Sharma, and S. Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis," Association for the advancement of Artificial Intelligence, 2012.
[3] J. Spencer and G. Uchyigit, "Sentiment or: Sentiment Analysis of Twitter Data," Second Joint Conference on Lexicon and Computational Semantics. Brighton:University of Brighton, 2008.
[4] A. Blom and S. Thorsen, "Automatic Twitter replies with Python," International conference "Dialog 2012".
[5] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1), 2008.
[6] Tumasjan A., Sprenger T.O., Sandner P.G., Welpe I. M., Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. AAAI ,2010
[7] Terveen L., Hill W., Amento B., McDonald D., and Creter J., PHOAKS: A system for sharing recommendations. In Communications of the Association for Computing Machinery (CACM),2007, 40(3):59–62
[8] Taboada M., Gillies M. A., and McFetridge P., Sentiment classification techniques for tracking literary reputation.In LREC Workshop: Towards Computational Models of Literary Analysis, 2006: 36–43.
[9] Piao S., Ananiadou S., Tsuruoka Y., Sasaki Y., and McNaught J., .Mining opinion polarity relations of citations. In International Workshop on Computational Semantics 84 (IWCS), 2007:366–371.
[10] Kumar, A. & Ahmad, N. ComEx Miner: Expert Mining in Virtual Communities, International Journal of Advanced Computer Science and Applications (IJACSA), Vol.3, No. 6, June 2012, The Science and Information Organization Inc, USA.
[11] Seki Y., Eguchi K., Kando N., and Aono M., Multi-document summarization with subjectivity analysis at DUC 2005. In Proceedings of the Document Understanding Conference (DUC)
[12] Spertus E., Smokey: Automatic recognition of hostile message. In Proceedings of Innovative Applications of Artificial Intelligence (IAAI),1997: 1058–1065.
[13] Annett, M. and Kondrak, G. (2008). A comparison of sentiment analysis techniques: Polarizing movie blogs. Advances in Artificial Intelligence, 5032:25–35.
[14] Chen C, Ibekwe-SanJuan F, SanJuan E, Weaver C (2006) Visual analysis of conflicting opinions. In: IEEE Symposium on Visual Analytics Science and Technology, pp 59–66
[15] Boiy, Erik, Hens, Pieter, Deschacht, K, and Moens, Marie-Francine. 2007. Automatic sentiment analysis of on-line text. In Proceedings of the 11th International Conference on Electronic Publishing. Vienna, Austria.
[16] International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 16291-16301 © Research India Publications http://www.ripublication.com Sentiment Analysis of Twitter Data: Based on User-Behavior Ms.Umaa Ramakrishnan and Ms.Rashmi Shankar



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY