ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Effective Sentiment Analysis of a Launched Product using Clustering and Decision Trees

Rishabh Soni[1], K. James Mathai[2]

PG Scholar, Dept. of Computer Engineering and Applications, NITTTR Bhopal, Madhya Pradesh, India[1]

Associate Professor, Dept. of Computer Engineering and Applications, NITTTR Bhopal, Madhya Pradesh, India[2]

ABSTRACT: Sentiment analysis is one of the many areas of computational studies that analyzes people's sentiments, opinions, evaluations, appraisals, attitudes, and emotions towards entities such as products, services and events. Many previous researches in the area of sentiment analysis used twitter data to classify the tweets into positive or negative, depending on the sentiment. A number of researchers have tried to improve the accuracy of such a classification. In this research, an approach of 'Cluster-then-Predict' is used to first cluster the tweets using k-means algorithm and then perform classification using Classification Trees. This clustering operation makes the data domain-specific, which results in creation of better predictive models which has led to more accurate classification of sentiments of a recently launched product. The rationale of this research is to effectively perform sentiment analysis of a recently product 'iPhone 6s' developed by the company Apple using the 'Cluster-then-Predict' approach.

KEYWORDS: Sentiment Analysis; Text Mining; Classification Algorithms; Clustering; Social Networks

I. INTRODUCTION

Opinion Mining and Sentiment Analysis alludes to the field of Natural Language Processing (NLP), computational linguistics and text mining involving the study of opinions, sentiments and emotions expressed in texts. An opinion may be viewed as a statement in which the opinion holder makes a specific claim about a product, news article, post, etc. using a certain sentiment. Knowledge acquired from social networks such as Twitter and Facebook are extremely valuable because thousands of opinions expressed about a certain topic are highly unlikely to be unfair or biased. The intuitive nature of such opinions makes them an effective tool by the majority of researchers, which make them the basis for making decisions regarding product review, marketing research, stock market prediction, etc.

The rationale of this research is to effectively perform Sentiment Analysis of a particular product from a company. User opinion on the product 'iPhone 6s' were mined from the popular microblogging website Twitter, using its Application Programming Interface (API). The main goal of such a Sentiment Analysis is to discover the sentiment the users have towards the product 'iPhone 6s'. This type of analysis helps the companies to make better decisions, based on the sentiment of the users.

Sentiment classification is done to classify the textual data, or 'tweets' in case of Twitter, as showing positive or negative sentiment. Until now, the unsupervised learning in the form of clustering for the purpose of sentiment classification had not been investigated sufficiently. In this research, the twitter data that is collected from Twitter was classified into two categories; positive and negative. An analysis was then performed on the classified data to investigate what percentage of the consumer sample falls into each category. A hybrid approach named 'Cluster-then-Predict' is used in which first clustering on our data is performed and then the sentiment is predicted. This improves the accuracy of the classifier, as well as interpretability of the solution.

Particular emphasis is placed on evaluating various machine learning algorithms and proposed 'Cluster-then-Predict' method for the task of twitter sentiment analysis. The evaluation is performed on the basis of accuracy of prediction, F-Score, interpretability of the solution, etc.

II. RELATED WORK

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

The emotions of fans during a live ongoing match can also be measured using Twitter. Authors Yang Yu and Xiao Wang in their research [1] track the emotions of U.S. soccer fans during five 2014 FIFA World Cup games, including three games featuring their own teams. They use Twitter search API to extract tweets during the matches. The research conceded that the sports fans use Twitter for emotional purposes and that big data approach to analyze sports fans' sentiment showed results consistent with the predictions of the disposition theory when the fanship was clear and showed good predictive accuracy. They use big data to analyze a large number of tweets to extract sentiment and found that the results are consistent with their expectations.

Election results is also other area in which Twitter is used for prediction. Authors Vadim Kagan et al. [2] use a sophisticated mix of social of social network analysis and methods to learn diffusion models from Twitter data. Using this, authors could accurately project the winner of the Pakistani 2013 and Indian 2014 election – and, because it was real-time, they were able to predict the most influential individuals on specific topics on social issues on social media, allowing potential candidates in the future to use this knowledge to better shape their campaign strategy.

Twitter messages are also used to determine sentiment towards a brand. Authors M. Ghiassi et al. [3] use twitter messages to determine sentiment towards a Justin Bieber brand. They introduced an approach to supervise feature reduction using statistical analysis and n-grams to develop a lexicon which is Twitter-specific, for sentiment analysis. This reduced Twitter-specific lexicon with brand-specific terms for brand-related tweet. They show that the reduced lexicon set, while well smaller, lowers modeling capacity and maintains a high degree of coverage over Twitter corpus and yields improved sentiment classification accuracy. To demonstrate the efficacy of the devised Twitter-specific lexicon compared to conventional sentiment lexicon, they develop comparable sentiment classification models using SVM. They then develop sentiment classification models using the DAN2 machine learning and Twitter-specific lexicon approach. The author Mohamed M. Mostafa in his research [4] used a random sample of 3516 tweets to evaluate consumer's sentiment towards well-known brands such as IBM, Nokia, DHL, T-Mobile, KLM. He used 6800 adjectives with expert-defined lexicon with known orientation to perform the analysis. Their results indicated a mostly positive consumer sentiment towards several famous brands. By using both a quantitative and qualitative methodology to analyze tweets related to brands, study adds breadth and depth to the discussion over attitudes towards cosmopolitan brands.

Ensemble classifier approaches such as boosting and bagging are also used in sentiment classification. In the paper [5], M. Govindarajan used bagging classification approach for movie sentiment classification and proved that this ensemble learning approach is superior to individual approaches in terms of accuracy. Sentiment analysis is also carried out in sporting events like FIFA world cup. In a research, Peiman Barnaghi et al. [6] extract the sentiment form Twitter to look for a correlation between these sentiments and FIFA World Cuo 2014 events of interest. In this process, they also classify the sentiment as positive or negative using Logistic Regression Classification (LRC).

Previous researches in twitter sentiment analysis used twitter data to classify the tweets into positive or negative, depending on the sentiment. A number of researchers have tried to improve the accuracy of prediction of such a classification. Till now, classification algorithms such as Naïve Bayes, Support Vector Machines, Classification and Regression Trees, Random Forest, were used for classification. Different attempts have been made in order to increase the prediction accuracy by using different methods.

Author Anders Westling in his paper [7] realize that clustering can be done to divide the texts into groups such that the text in the same group are more similar to each other than to text in other groups. When combined with other methods, clustering can be useful when no classified data is available. Author John Dodd in his research [8] recognizes that using k-means clustering may provide more valuable insights when combined with sentiment analysis. A clustering algorithm may discover groups of tweets about a particular feature of the product, or other information related to its release. Authors AK Jose et al. in their project [9] also propose the idea of using data clustering for making the classification domain-specific and thus to improve accuracy. Author Akhavan Rahnama in a research [10] recognizes clustering of social streams in real-time as an improvement in sentiment analysis. Author Yanchang Zhao [11] extracts text from Twitter to build a document-term matrix. He also performs clustering to find groups of words and also groups of tweets. But he does not partition his tweet data according to the cluster to perform prediction on it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

From the studies and analysis of papers published before, it was observed that classification algorithms like SVM and RandomForest gives better prediction accuracy, especially when compared with CART algorithm. But CART has advantage over others in that trees can be visualized to observe which keywords are significant. The 'Cluster-then-Predict' approach can improves the accuracy of CART to a level comparable with other algorithms, while significantly enhancing the interpretability of the result.

III. RESEARCH METHODOLOGY

This section contains the steps that were taken to implement 'Cluster-then-Predict' model. The figure 3.3 contains the flowchart of implementation, showing the steps needed to be performed from collecting raw data to prediction, using the proposed approach.

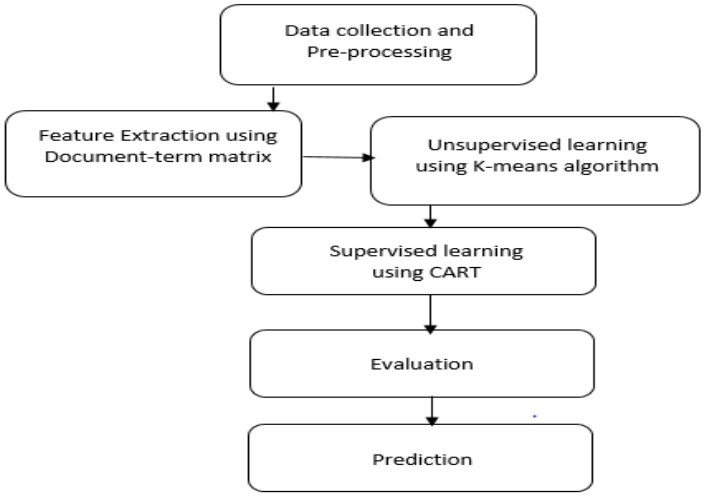

Fig.1. Flowchart of 'Cluster-then-Predict' approach for sentiment analysis

A. *Obtaining Raw Data:*

The raw was obtained using Twitter Streaming API. This was implemented in order to get live tweets showing some sentiment towards the product 'iPhone 6s'. A database of 2785 tweets were collected for analysis. The text of the tweet was saved in a CSV file simultaneously while it was streaming. After that, sentiment values for each tweet was assigned. The sentiment value was taken as an average of ratings by five persons who were familiar with smartphones and apple products. This was done for supervised training purpose.

B. *Cleaning and Pre-processing the data:*

It is hard for a machine to understand raw twitter data because of the heavy use of homonyms and metaphors. Also, sarcasm is widely used in tweets which are hard for computers to interpret. Also, tweet data is loosely structured, textual, has spelling mistakes, and could be multi-lingual. The data cleaning process was done to remove any unwanted content from the training data and the input tweets. Data cleaning process not only simplified the classification task for the machine learning model but it also served to greatly decrease processing cost in the training phase.Some pre-processing stepswere performed to ensure that our algorithm works well with data.

1. Change the text so that all the words are either in lower-case or upper-case.
2. Punctuations can also cause problems- basic approach is to remove everything that is not alphabet or numeric. All punctuations were removed so that '@iPhone 6s '#iPhone 6s', 'iPhone 6s!' will all count as just 'iPhone 6s'.
3. Removing unhelpful terms. Many word s are frequently used but are only meaningful in a sentence, these are called "stopwords". Examples: is, the, at, which.These are unlikely to improve machine learning prediction quality, and also reduce the size of the data.
4. Stemming: This step is motivated by the goal to represent words with different endings as the same word. Example: "argument", "argued", "arguing" are all changed to simply "argue".

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

To pre-process the data, 'tm' text mining package and 'SnowballC' packages in 'R' had been used.A database of 2785 tweets was collected for analysis. It contained 1891 tweets which showed positive sentiment and 894 tweets which showed negative sentiment towards 'iPhone 6s'.

C. *Feature Engineering:*

For feature extraction from the twitter corpus, 'Bag-of-Words' model was used. This model is usually used in document classification, where the frequency of each word in the tweets is used as a feature for training a classifier.

The data obtained from the document-term matrix is called sparse, means that there are many zeroes in our matrix.Thisis an issue for two main reasons; first one is computational- more terms means more independent variables, which typically means that it takes longer to build our models. The other is, in building models the ratio of independent variables to observations will affect how good will model generalize. So, the terms that don't appear very often were removed, using sparsity threshold.A new matrix is produced which contains the terms which appear in at least 1% of the tweets.

D. *Performing unsupervised learning using K-means clustering:*

A number of K clusters which partition our data in such a way that features in one cluster, rate utterances with particular linguistic properties are most similar to each other. In other words, tweets must be distributed in clusters such that similar tweets, based on words they contain, gets partitioned in one cluster. Thus, tweets were clustered based on the words. The K-means clustering technique was used for this application.

The contents of the each cluster shows the following popular words in each of the four clusters:

cluster 1: win case enter
cluster 2: itune ipod iphone
cluster 3: new phone plus

From the above top words, generate from R code, it can be observed that the clusters are of different topics. The cluster 1 focusses on contests to win cover cases of the iPhone. Cluster 2 focusses on iTunes on Apple iPod and iPhone. Cluster analysis is very helpful for the companies to analyze consumer sentiment towards their products. Relation between various keywords within a cluster can give useful insights.

E. *Supervised learning through CART algorithm:*

Decision trees on one of the most widely used machine learning algorithms much of their popularity is due to the fact that they can be adapted to almost any type of data, in particular textual data. Classification and Regression Trees (CART) algorithm was used in this research to classify the tweets as positive or negative. The CART algorithm was applied to each of the test clusters to classify the sentiment. By plotting the model, the prominent keywords which sway public perception had been depicted.

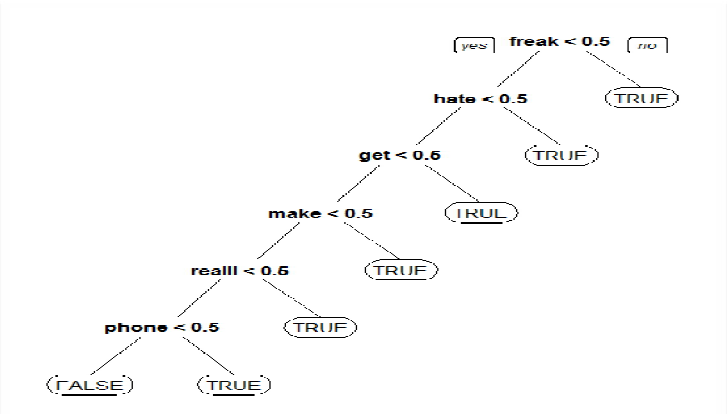

Fig.2. Plotting a Classification Tree

Using decision trees such as CART makes the solution much more interpretable. For example, in the case in figure 2, if the word freak is likely to appear in the tweet, then it is likely to be classified into TRUE, or showing negative emotion towards that product.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

At last, the predict values of sentiment are combined for the entire test set.

IV. RESULT ANALYSIS

The results and the performance of the proposed 'Cluster-then-Predict' approach against other machine learning classification techniques such as Support Vector Machines (SVM), Random Forests, Naïve Bayes, etc. are examined and explained in this section; with respect to various parameters such as accuracy, F-Score, precision, etc.

To obtain the results using the above performance parameters, confusion matrix was used. The sentiment value of the tweets is classified into '0' – showing positive sentiment and '1' – showing negative sentiment.

A. *Accuracy of the Classification:*

The accuracy is used as a statistical measure of how well a binary classification test correctly identifies the sentiment of the tweet as positive or negative. In other words, accuracy is the proportion of true results (true positives and true negatives) among the total number of cases.

The accuracy of the proposed 'Cluster-then-Predict' classifier is found to be the highest, as shown in the figure 3 below:

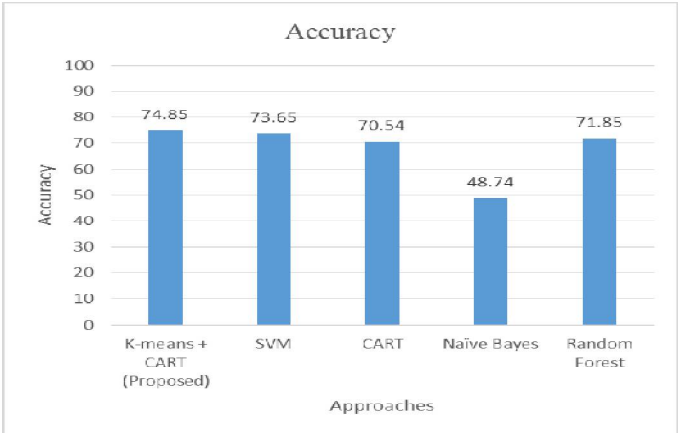

Figure 3: Comparative accuracy of the Proposed, SVM, CART, Naïve Bayes and Random Forest approaches

As shown in the figure 4, the accuracy of classification of the sentiment for proposed approach is highest. It means that using the 'cluster-then-predict' approach, 74.85% of the tweets in test data are correctly predicted as showing the actual sentiment of the users towards the product 'iPhone 6s. This accuracy value is significantly higher than CART, on which the 'cluster-then-predict' approach is based.

B. *Sensitivity and Specificity:*

Sensitivity or recall or True Positive Rate (TPR) measures the proportion of 1's, i.e. negative sentiment, which are correctly calculated as having negative sentiment.Specificity measure the proportion of 0's, i.e. positive sentiment, which are correctly identified as having positive sentiment.

The following figure shows the sensitivity and specificity values of various approaches:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

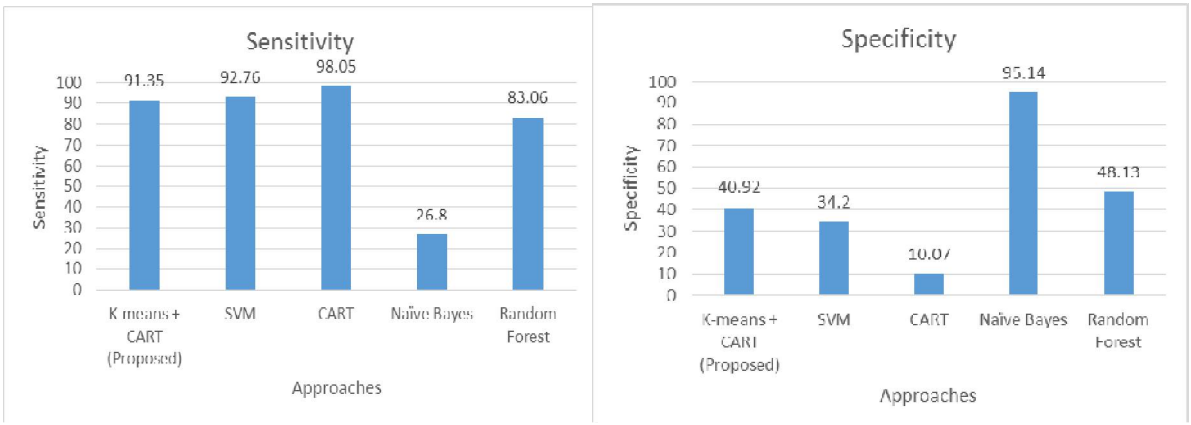

Figure 4: Comparative sensitivity and specificity values of the Proposed, SVM, CART, Naïve Bayes and Random Forest approaches

From the figure 4, it shows that using the proposed 'Cluster-then-Predict' approach, 91.35% of the tweets are predicted correctly to show negative sentiment towards the product 'iPhone6s'. It also shows that using the proposed approach, 40.92% of the tweets are predicted correctly to show positive sentiment towards the product 'iPhone 6s'. Using Naïve Bayes, high value for specificity is obtained because in predicts almost all of the tweets as showing positive sentiment.

From the above sensitivity and specificity statistics, it could be observed that the proposed approach gives a better balance between the two. This fact is further established when we examine the harmonic mean of Precision and Recall (F Score).

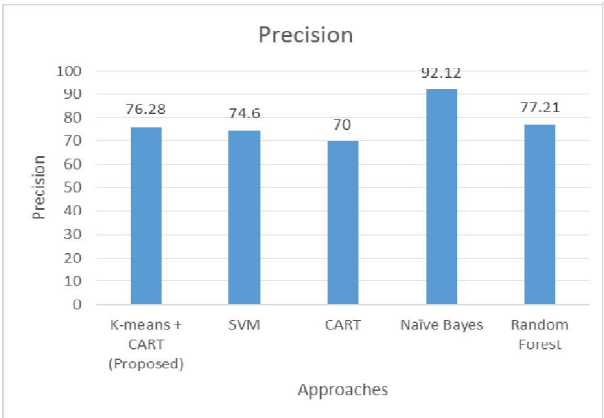

Figure 5: Comparative precision values of the Proposed, SVM, CART, Naïve Bayes and Random Forest approaches

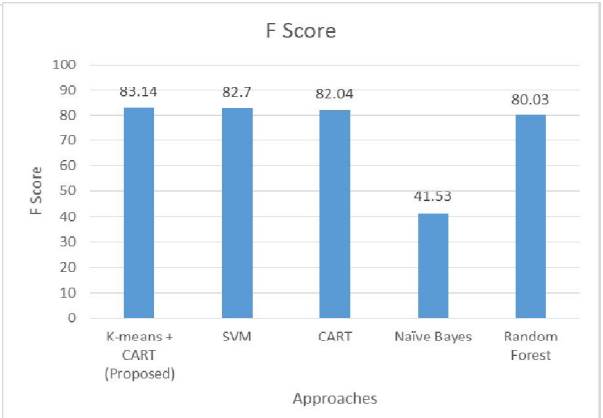

Figure 6: Comparative F-Scores of the Proposed, SVM, CART, Naïve Bayes and Random Forest approaches

C. *Precision of the classification:*

Precision is analogous to positive predictive value. It means that if '1' (or negative sentiment) is predicted, how often is it correct. The following figure 5 shows the precision values of various approaches.The figure 5 shows that the precision score of Naïve Bayes is highest, this means that Naïve Bayes most accurately predicts the negative sentiment of a tweet of a user towards 'iPhone 6s'. The proposed approach gave the precision measure 0.7628. This means that among all those tweets that showed negative emotion, the probability of the tweet actually showing negative emotion is 76.28%. The proposed approach has higher values of precision than SVM and CART.

D. *F-Score of the classification:*

The F Score measures the accuracy using the precision and recall. The F1 metric weights recall and precision simultaneously. Therefore, reasonably good performance on both will be favored over extremely good performance on

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

one and poor performance on the other. The following figure 6 compares the F Score of various approaches using a bar plot.

From the figure 6, it is concluded that the proposed approach gives the best F Score. This means that the sentiment of the user towards 'iPhone 6s' is most correctly predicted as negative or positive using the proposed approach. Since the F Score is harmonic mean of Precision and Recall, it is implied that proposed approach has the best performance and the model would generalize well on the test data.

A review table (Table 1) is shown below which specifies various evaluation parameters and results of evaluation.

Technique	Parameters					
	Accuracy (in %)	*TPR/ Sensitivity (in %)*	*FPR/ Specificity (in %)*	*Precision (in %)*	*F Score*	*AUC*
K-means + CART (Proposed)	74.85	91.35	40.92	76.28	83.14	76.24
SVM	73.65	92.76	34.20	74.60	82.70	72.35
CART	70.54	98.05	10.07	70	82.04	59.64
Naïve Bayes	48.74	26.80	95.14	92.12	41.53	52.92
Random Forest	71.85	83.06	48.13	77.21	80.03	70.47

Table 1: Performance of the proposed 'Cluster-then-Predict' approach

Thus, it could be observed that the parameters which are most important, which are accuracy, AUC and F Score are higher as compared to other algorithms. Also, the interpretability quotient of the innovative 'Cluster-then-Predict' approach is higher as well. This is due to the fact that cluster analysis can be performed after clustering. Also, classification trees gives important keywords which affect the public view towards the product.

Using the proposed 'Cluster-then-Predict' approach, it was found that out of 736 tweets in the test data, 704 showed positive sentiment, and only 32 showed negative sentiment towards iPhone 6s.

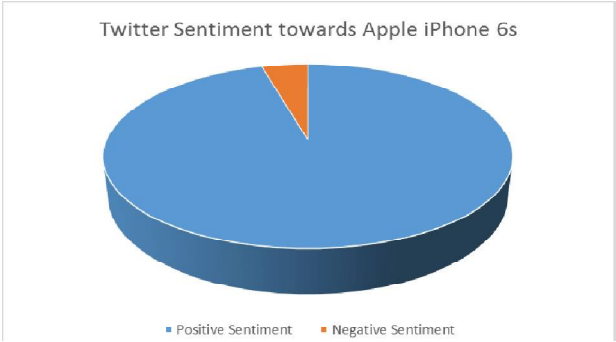

Figure 7: Pie chart showing predicted Twitter user's sentiment towards Apple iPhone 6s

So, using the 'cluster-then-predict' approach of sentiment analysis, it was known that public largely accepted and liked the product and also gave good opinion about it on Twitter.

V. CONCLUSION

To conclude, this research has illustrated that an effective sentiment analysis can be performed on a product, iPhone 6s in this case, by collecting a sample audience opinions from Twitter. Throughout the duration of this research many different data analysis tools were employed to collect, clean, mine and evaluate sentiment from the dataset. Such an analysis could provide valuable feedback to the companies and help them to spot a negative turn in viewer's perception

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

of their brand image. Discovering negative trends early on can allow them to make educated decisions on how to target specific aspects of their services and products in order to increase its customer's satisfaction.

It is shown in this research that hybrid approach of 'Cluster-then-Predict' has a major effect on the overall accuracy of the analysis. This approach has an accuracy of about 75% for classification. Comparison between different algorithms and proposed approach shows that proposed approach is superior in critical evaluation parameters of accuracy, AUC and F Score. The solution obtained from the proposed approach is more interpretable as well.

REFERENCES

1. Yu, Yang, and Xiao Wang., "World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets." Computers in Human Behavior Vol. 48, pp. 392-400, 2015.
2. Kagan, Vadim, Andrew Stevens, and V. S. Subrahmanian, "Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election." IEEE Intelligent Systems Vol. 30, Issue 1, pp. 2-5, 2015
3. Ghiassi, M., J. Skinner, and D. Zimbra. "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", Expert Systems with applications Vol. 40, Issue 16,pp. 6266-6282,2013.
4. Mostafa, Mohamed M., "More than words: Social networks' text mining for consumer brand sentiments." Expert Systems with Applications Vol. 40, Issue 10, pp. 4241-4251, 2013.
5. Govindarajan, M.,"Bagged Ensemble Classifiers for Sentiment Classification of Movie Reviews". International Journal of Engineering and Computer Science Vol. 3 Issue 2, 2014.
6. Barnaghi, Peiman, Parsa Ghaffari, and John G. Breslin, "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets." ResearchGate (2015).
7. Westling, Anders,"Sentiment Analysis of Microblog Posts from a Crisis Event using Machine Learning". Master of Science Thesis, KTH CSC, Stockholm, Sweden, 2013
8. Dodd, J.,"Twitter Sentiment Analysis. Final Project Report", National College of Ireland, 2015.
9. Jose, A.K., Bhatia, N., Krishna, S."Twitter Sentiment Analysis. Major Project Report", NIT Calicut, 2010.
10. Rahnama, Amir Hossein Akhavan, "Distributed real-time sentiment analysis for big data social streams."IEEE,International Conference onControl, Decision and Information Technologies (CoDIT), pp. 789-794., 2014.
11. Zhao Y.,"R and data mining: Examples and case studies". Academic Press. pp. 97-109, 2012.

BIOGRAPHY

Rishabh Soniis a PG Scholar in the Department of Computer Engineering and Applications, National Institute of Technical Teachers' Training and Research, Bhopal, India. He received Bachelor of Engineering (BE) degree in 2012 from IISTIndore, M.P., India. His research interests are Machine Learning, Sentiment Analysis, Big Data Analytics, etc.

K. James Mathaiisan Associate Professor in the Department of Computer Engineering and Applications, National Institute of Technical Teachers' Training and Research, Bhopal, India.