International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Sentiment Analysis on Social Media Big Data With Multiple Tweet Words

S. Uma Maheswari, S. S. Dhenakaran

Abstract-The main objective of this paper is Analyze the reviews of Social Media Big Data of E-Commerce product's. And provides helpful result to online shopping customers about the product quality and also provides helpful decision making idea to the business about the customer's mostly liking and buying products. This covers all features or opinion words, like capitalized words, sequence of repeated letters, emoji, slang words, exclamatory words, intensifiers, modifiers, conjunction words and negation words etc available in tweets. The existing work has considered only two or three features to perform Sentiment Analysis with the machine learning technique Natural Language Processing (NLP). In this proposed work familiar Machine Learning classification models namely Multinomial Naïve Bayes, Support Vector Machine, Decision Tree Classifier, and, Random Forest Classifier are used for sentiment classification. The sentiment classification is used as a decision support system for the customers and also for the business.

Keywords-Opinion Mining, Social Media, Big Data, Support Vector Machine, NLP.

I. INTRODUCTION

Attaining Sentiment Analysis/Opinion Mining on real time Big Data is a challenging task. In twitter, for every second very huge volume of data, like customer's opinion/feelings posted, and shared in very high speed. These twitter posts can be different types such as text, pdf, image, audio and video. In terms of Volume, Velocity, and Variety these types of data called as Big Data. So collecting this real time data and processing this type of data for Sentiment Analysis or Opinion Mining is a highly challenging process. Because of Big Data is neither structured nor un-structured; it is semi-structured and huge volume of data. To Simplify the Big Data Analysis complications Hadoop frame work used as a platform in this proposed work.due to the popularity of opinion mining on social media big data has become a decision support system for decision makers to take decision on retail/e-commerce business. Social media provides feelings of customers on product and services. Customers in social media express their opinion by sharing 'reviews, comments, likes, and emotion symbols. Generally, in E-commerce business, social media review provides the best approach. Sentiment analysis is also referred to as opinion mining. This proposed work is done with twitter data of Amazon product.

Manuscript published on 30 August 2019.
*Correspondence Author(s)
S. Uma Maheswari Department of Computer Science, Alagappa University. Tamil Nadu
Dr. S. S. Dhenakaran Department of Computer Science, Alagappa University. Tamil Nadu

The work incorporates all the types of words and expression of words to give best and more impact on sentiment analysis. This proposed work helps the customers to know the product, brand, popular and quality of the products. Also helps retailers or organization to take the decision about the product, brand and quality based on the customer's likes, feelings/opinion. The work of paper is arranged as follows. Section 2 provides the related work on sentiment analysis. Section 3 proposes the methods and materials for mining and analysis used in the work. Section 4 represents the result and discussion of the proposed work. Finally, the conclusion is presented in section 6.

II. RELATAED WORK

A. Sentiment Analysis on TF-IDF

Rini Wongso et al. [5] have proposed Sentiment Analysis using TF-IDF in Indonesian articles. 5000 News articles are collected using web crawling with each article labeled by Economy, Health, Technology, Sports and Politics. Text preprocessed by performing lemmatizing, stop words removal. Feature words are selected by TF-IDF (Term Frequency-Inverse Document Frequency) to calculate SVD (Singular Value Decomposition) for Naïve Bayes algorithm to classify News articles.

B. Opinion Mining on Intensifier Detection

Dibakar Ray et al. [6] have put forward the Sentiment Analysis on twitter data using lexicon based sentiment. The analysis has used negation words, degree modifier words, capitalized words, slang words, repeated letters in a word and multiple punctuations in a word. It has provided good impact on Sentiment Analysis.

C. Sentiment Analysis on Twitter data using Likes - ReTweet

Rizal Setya Perdana et al. [7] have suggested the combination of ReTweet and likes with the Sentiment Analysis. It has used Naïve Bayes classifier, ReTweet count and likes count tweets for analysis. Here the tweets are classified as positive or negative to find F1 score which is used for better analysis.

D. Opinion Mining on Emoji

Valmeekam Karthik et al. [8] have advocated the Sentiment Analysis on emojis(facial expression) by Machine Learning Classifiers, Convolutional Neural Network, and Artificial Neural Network.

Tweets without emoji are discarded. All emojis are converted to Unicode. When emoji is not present then all slash symbols appearing in tweet (\) and Xs are removed and other characters are converted to ASCII code then to as Unicode.Existing research work has avoided special types of data by implementing stop words removal, and preprocessing. Several works have already done in Sentiment Analysis by only taking the positive and negative sentiment words (good or bad). A few research works has done with emoji symbols and slang words (OMG, LOL) [1], [2] and some other work are done with negation words (not good, not bad) [3], [4]. Existing works done using limited features, such as negation words, degree modifier words, capitalized words, slang words, repeated letters, and, emoji symbols used as separately or combined one or two features together. But there is no existing work using all these features combined together. If use these all features it will give a good impact on Sentiment Analysis accuracy. So this proposed work done using all these features together.

III. PROPOSED WORK

Methods and Material of Proposed Work

In this proposed work to extract the real time Big Data on Social Media about the E-Commerce products, Twitter API used. For this proposed work Amazon product reviews collected from the tweets in real time manner. To analyze this type of Big Data and visualize the results, Hadoop, Twitter API, Python and Nature Language Processing Toolkit used. In this proposed work for Sentiment Analysis and classification purpose popular Machine learning Techniques used.

Implementation of the Proposed Work

Opinion Mining is searching the opinion/emotional words or sentences and extracts the opinion words or sentences. In this proposed work opinionated words identified and extracted from the tweets and analyzed. The Following methods used for extract and analyze the sentiment/opinion words.

A. Lexicon Method

This method checks the tweet word with AFINN word dictionary having polarity score between +5 and – 5 for each word. If the word matches with word of AFINN then polarity score is retrieved for determining whether the word is positive or negative. Then final sum of score determines the subjectivity of the (positiveness or negativeness) of sentences.

B. Preprocessing

To extract the meaningful content preprocessing is done on tweets to remove URL, Hash Tag, User Tag and words are tokenized for part of speech.

C. Emoji Detection

Calculating Emoji's polarity score increases impact of Sentiment Analysis. emoji Unicode and their polarity scores are combined and constructed a dictionary to find emoji score which ranges from -1 to +1. When emoji score is positive, customers are happy.

D. Negation Detection

Special symbols, such as (), [], {}, $, %, are removed from the tweets after emoji's polarity score calculated. In many works, negations are handled by reverse polarity to find score. For example, "not excellent" represents the negative meaning. But really it does not convey the negative meaning but it can be good or nice but not excellent which is ultimate meaning of the tweet. In this proposed work polarity shifting used. For example, "good" has polarity score 3 and "not good" polarity is - 4 + 3 =1. Likewise "not bad" is gives polarity 4 -3 =1. That is, not have either +4 or -4 depending on usage.

E. Slang Word Detection

Slang words are a short form of real words such as "Oh My God" is represented as "OMG" in tweets. Generally, sentiment list do not have such kind of words. The slang word detection in the proposed work, inputs the tokens of tweets (Sentences) for slang word detection to expand such words for getting meaning of tweets. Slang word's dictionary constructed by combining slang words with their respective expansion.

F. Modifier Detection

Modifiers word list such as very, most, ugly, truly, etc are considered in the proposed work. These modifier words having fixed score 5, increase the impact on the sentiment terms.

G. Stop Words Removal

The single lettered words such as a, an, and he, she, the, to are removed from tweet. These stop words are not significant for sentiment score calculation.

H. Intensifier Detection

The intensifier detection finds exclamatory symbol, word with repeated characters and capitalized words are represented in intensifier. These intensifier words are more expressive than normal words but polarity score is doubled to these words.

I. Degree Modifier

The words like "reallyyy", "verrryyy", are viewed as more expressive than the normal intensifier words. These kinds of words are normalized and spell checked. The presents of such words, the polarity score is doubled.

J. Conjunction

Some tweets have two-sentence or three-sentence. These sentences of tweets are concatenated by "but, however, while, although" giving impact on Sentiment Analysis. For example, "This mobile's camera quality is nice but Battery life is poor". In this sentence "This mobile's camera quality is nice" conveys positive sentiment; "Battery life is poor" conveys negative sentiment. Here concatenation word "but" passes impact to sentiment having score 4 for the sentence.

Retrieval Number: J96840881019/19©BEIESP
DOI: 10.35940/ijitee.J9684.0881019
Journal Website: www.ijitee.org

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

K. ReTweet & Likes

When retrieving tweets from Twitter, ReTweet count and likes count also retrieved. If ReTweeted count is above 0 then it is a positive sentiment with polarity score +1 if not polarity score is -1. Likewise, if Likes count is above 0 then the tweet is liked by others; so Likes polarity score is +1, if the Likes count is above for negative tweet then the Likes' polarity score is represented as -1. If ReTweet & Likes count increase for positive tweet then ReTweet & Likes polarity score is increased. If ReTweet & Likes count increase for the negative tweet then ReTweet and Likes polarity score is decreased.

L. Total Score

For each tweet, emoji's score, negation's score, modifier score, intensifier word score, degree modifier score, concatenation's score, pos-tagged word score, text blob score are calculated and summed to get the total score of a given tweet.

M. N-Gram

N-gram is a combination of words. Combination of two words is 2-gram words, a combination of three words 3-gram words, likewise n combination words n-gram words. The proposed work has used 3-gram words used.

N. Sentiment Analysis & Classification

In the proposed work sentiment score calculated and tweets are classified as highly positive, moderately positive, lightly positive, highly negative, moderately negative, and lightly negative.

For, 0< score <=3, then tweet is lightly positive. For score 4<= score<=6, it is moderately positive.

If polarity score is equal or above 7 then tweet is highly positive.

For, -3<=score<0, then tweet is lightly negative.

For, -6<= score<=-4, then tweet is moderately negative, if sentiment<=-7, is represented as highly negative.

O. Opinion Extraction by TF - IDF (Term Frequency-Inverse Document Frequency)

For extracting opinion words, TF-IDF method is used in the proposed work. N-gram words are used as input to this method. Here words are transformed into vector values by python TF-IDF package.

P. Proposed Work Architecture

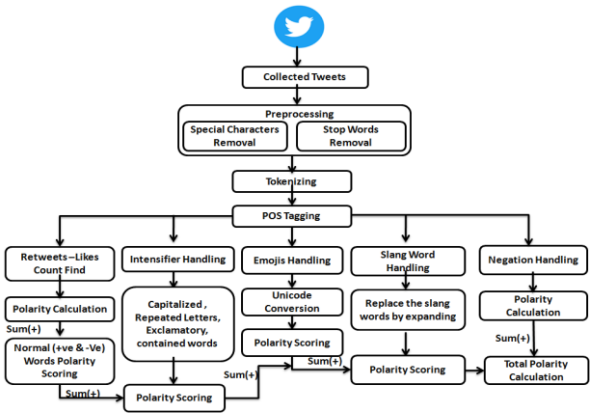

Fig.1. Polarity Calculation

In Fig.1 tweets collected, preprocessed and scores calculated separately for Retweets-Likes, intensifier words, emoji, slang words, and negation words. Then finally total polarity score of tweet calculated by adding all the score's of above mentioned words.

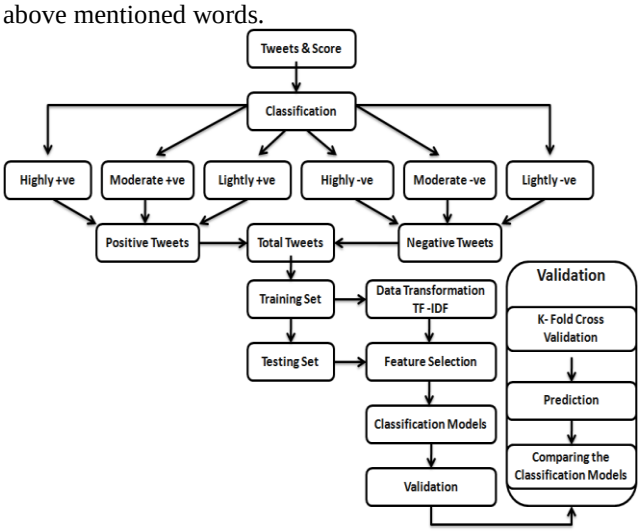

Fig.2. Best classification model selection

In Fig.2 tweets with polarity score is classified as highly +ve, moderately +ve, lightly +ve, highly –ve, moderately –ve, and lightly –ve. Then all positive tweets added as positive tweets, and all negative tweets added as negative tweets. Then total tweets with the polarity score divided as training and testing dataset. Sentiment Analysis performed using TF-IDF and features selected from the tweets. These opinion words will be input for the classification models. Then the models are evaluated.

Q. Performance Evaluation

The Statistical Method has used six kinds of parameters such as confusion matrix, accuracy, precision, recall, f1-score, and ROC-AOC area, to evaluate the performance in the work to choose best classification model. This evaluation is helpful to find effectiveness of classification algorithms.

A. Confusion Matrix

Confusion matrix provides the information about actual and predicted values given by the classification algorithm. Performance of the classification algorithm is represented & evaluated using matrix.

B. Accuracy Rate

Accuracy Rate is the total number of predictions correctly predicted. It determined using the following equation.

Accuracy Rate = (TP + FN) / (TP + TN + FP + FN)

C. Precision

Precision is the number of true positive class divided by the number of true positive and false positive.

Precision = TP / (TP + FP)

D. Recall

Recall is the number of true positive class divided by the total number of true positive and true negative.

Recall = TP / (TP+FN)

E. F1- Score

F1- Score is the harmonic mean of precision and recall. F1-Score is calculated by the below formula.

F1- Score = 2 *(Precision * Recall) / (Precision + Recall).

F. ROC (Receiver Operating Characteristic) Curve

ROC shows the trade-off between the true positive rate and the false positive rate. Area under the ROC curve is a measure of the accuracy of the classification model. Typically this represented as AUC. ROC curve measure plots the curve using the true positive and false positive predicted classes.

IV. RESULT AND DISCUSSION

Three experiments are performed on Big Data of Amazon products reviews. Customers mostly buy Electronic Items, Fashion Items, and Books in Amazon. Sentiment Analysis is done on these dataset by machine learning algorithms Naïve Bayes, SVM, Decision Tree, Logistic Regression and Random Forest models.

Experiment I

This experimental analysis is done on real time data of reviews about books. In this experiment 2672 records collected for Sentiment Analysis. The table below shows the performance evaluation of K-Fold cross validation on classification model. Fig.3 shows the number of tweets collected in this work. Fig.4 shows the ROC_AUC on classification models. In this result, accuracy of SVM is 98.69%, it is better than the other classification models. Again precision value of SVM is obtained 99.98% which is better than other models. SVM has obtained F1-score 99.15%, ROC_AUC area value is 99.74, which is better than other classification models. In Recall, SVM is not giving highest value when compared to other models but 99.32% is better result.

TABLE.I. SENTIMENT MEASURES ON BOOKS

Classification Models	K-Fold Accuracy %	Precision %	Recall %	F1 Score %	ROC Area %
Multinomial Naïve Bayes	98.69	98.32	100	99.15	98.93
SVM	98.69	98.98	99.32	99.15	99.74
Decision Tree	93.73	97.86	93.86	95.82	93.59
Logistic Regression	94.26	93.02	100	96.38	99.43
Random Forest Classifier	97.91	97.66	99.66	98.65	99.23

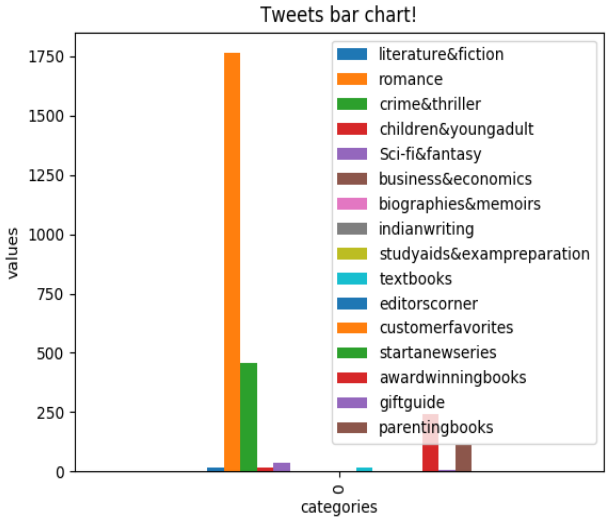

Fig 3. Number of Tweets for Books

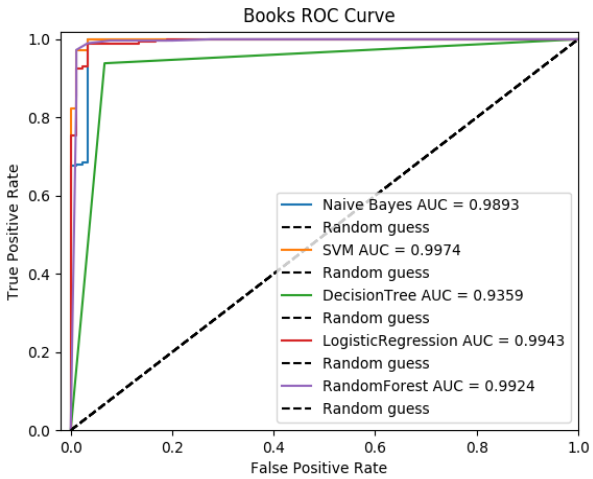

Fig 4. ROC_AUC Curve and Area of Books

From the Fig.3.can identify that in the Book's Tweets Romance type books highly liked by the customers.

Experiment II

This experimental is analyzed reviews of electronic items. In this experiment 3326 records are collected for Sentiment Analysis. In this experiment, SVM has produced 93.46% which is better than other classification models. Again precision value of SVM is 92.74%; it is not high than the Decision Tree model but 92.74% considered better value. The recall values of Naïve Bayes and Logistic Regression are better than other models. In recall SVM not obtained highest value but obtained 99.32%, it is also near to 100%. SVM has obtained F1-score 95.92% and ROC_AUC area value 95.13%, which is better than other classification models.

Retrieval Number: J96840881019/19©BEIESP
DOI: 10.35940/ijitee.J9684.0881019
Journal Website: www.ijitee.org

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

TABLE.II. SENTIMENT MEASURES ON ELECTRONIC ITEMS

Classification Models	K-Fold Accuracy%	Precision %	Recall %	F1 Score%	ROC Area %
Multinomial Naïve Bayes	91.36	89.97	100	94.72	91.40
SVM	93.46	92.74	99.32	95.92	95.13
Decision Tree	89.27	95.05	90.88	92.92	87.30
Logistic Regression	87.96	86.76	99.66	92.77	94.28
Random Forest Classifier	92.15	91.56	98.99	95.13	91.15

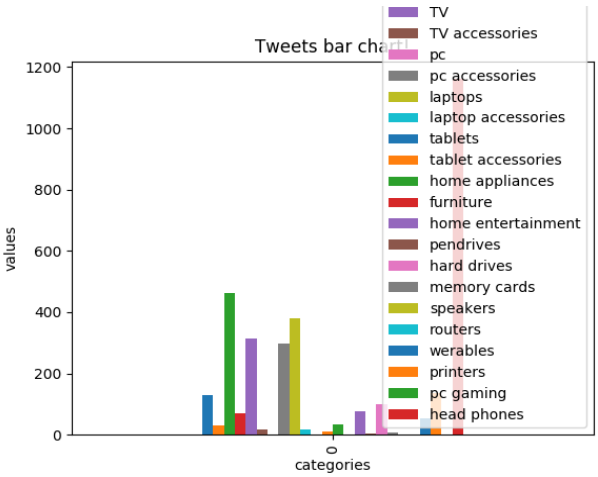

Fig 5. Number of Tweets for Electronic Items

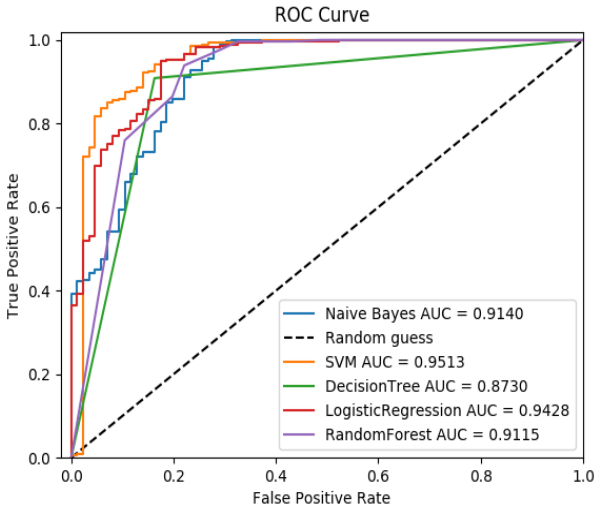

Fig 6. ROC_AUC Curve and Area of Electronic Items

From the Fig5.can identify that in the Electronic Items' Tweets Head Phones highly liked by the customers.

Experiment III

This experimental is on Fashion items. Nearly 658 records are collected for Sentiment Analysis. In this result, accuracy and precision of SVM is 96.97% which is better than other classification models. Recall value of SVM model is 98.31% which is better than other models. SVM has produced F1-score 97.48%, ROC_AUC value 99.62% which are better than other classification models.

TABLE.III. SENTIMENT MEASURES ON WEARABLE THINGS

Classification Models	K-Fold Accuracy %	Precision %	Recall %l	F1 Score %	ROC Area %
Multinomial Naïve Bayes	96.97	96.67	98.31	97.48	99.15
SVM	96.97	96.97	98.31	97.48	99.62
Decision Tree	93.94	96.49	93.22	94.83	94.11
Logistic Regression	96.67	96.67	98.31	97.48	99.49
Random Forest Classifier	96.67	96.67	98.31	97.48	99.47

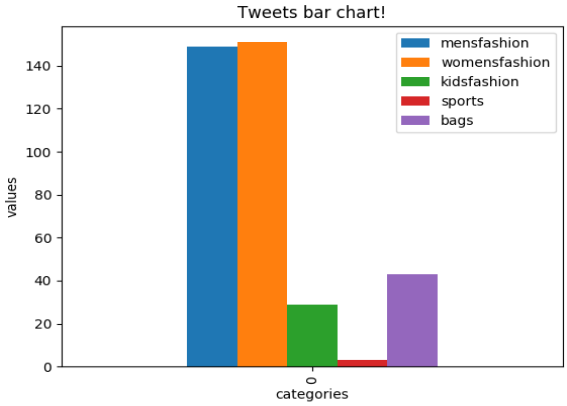

Fig 7. Number of Tweets for Fashion Items

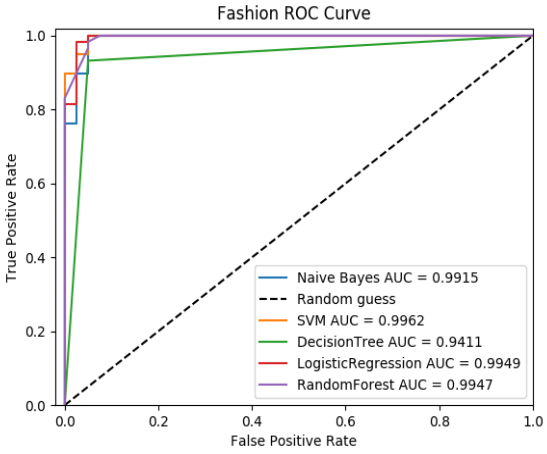

Fig 8. ROC_AUC Curve and Area of Fashion Items

From the Fig.7.can identify that in the Fashion Item's Tweets Men's Items highly liked by the customers.

From the experimentation, it is observed that SVM has given better results. Existing work done only on Re Tweets and likes but not concentrate on other kind of words.

Retrieval Number: J96840881019/19©BEIESP
DOI: 10.35940/ijitee.J9684.0881019
Journal Website: www.ijitee.org

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

No research work concentrates on all kind of user's feeling and their written format. This proposed work has taken all the kind of words
written by the users. This gives the best impact on the Sentiment Analysis and the classification. This proposed work is done the performance evaluation process using many statistical performance metrics in order to choose one best classification model out of four famous classification models.This proposed work also has some limitations that it does not include the quoted words such as single quoted words and double quoted words. This proposed work is done in python platform this more time consuming to execute and produce the result for run the huge amount of tweets.

V. CONCLUSION

In this proposed work real time tweets about the Amazon products and more parameters of tweets are included in Opinion Mining/Sentiment Analysis on E-Commerce Big Data. Five machine learning models are used on three experiments and seen only SVM has given better results than other models and hence it is observed SVM is a better classification model for Opinion Mining/Sentiment Analysis on Big Data.

ACKNOWLEDGMENT

This Proposed Research Work "Sentiment Analysis on Social Media Big Data with Multiple Tweet Words", has been written with the financial support of RUSA - Phase 2.0 grant sanctioned vide Letter No. F. 24-51 / 2014-U, Policy (TNMulti-Gen), Dept.of Edn. Govt. of India, Dt.09.10.2018.

REFERENCES

1 Mr Amritkumar Tupsoundarya1, Prof.Padma, S. Dandannavar, Sentiment Expression via Emoticons on Social Media: Twitter, International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; Volume 6 Issue VI, June 2018.
2 Liang Wu, Fred Morstatter, Huan Liu, SlangSD: Building and Using a Sentiment Dictionary of Slang Words for Short-Text Sentiment Classification, arXiv: 1608.05129v1 [cs.CL] 17 Aug 2016.
3 Naw Naw, Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers, International Journal of Scientific and Research Publications, Volume 8, Issue 10, October 2018 407, ISSN 2250-3153.
4 Wareesa Sharif, Noor Azah Samsudin, Mustafa Mat Deris, Rashid Naseem, Muhammad Faheem Mushtaq, Effect of Negation in Sentiment Analysis,International Journal of Computational Linguistics Research Volume 8 Number 2 June 2017.
5 Rini Wongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli Rudy, News Article Text Classification in Indonesian Language, Procedia Computer Science 116 Elsevier (2017) 137–143.
6 Dibakar Ray, Lexicon Based Sentiment Analysis of Twitter Data, International Journal of Research in Applied science & Engineering Technology, ISSN: 2321-9653; IC Value: 45.98; Volume 5 Issue X, October 2017
7 Rizal Setya Perdana and Aryo Pinandito, Journal of Telecommunication, Electronic and Computer Engineering, e-ISSN: 2289-8131 Vol. 10 No. 1-8. April- 2018.
8 Valmeekam Karthik, Dheeraj Nair, Anuradha J, Procedia Computer Science Volume 132 Elsevier 2018, Pages 167-173.

AUTHORS PROFILE

S. Uma Maheswari MCA., M.Phil. She is currently doing Ph.D in Computer Science, Alagappa University. Her main areas of interest are Big Data Analytics, Sentiment Analysis.

Dr. S. S. Dhenakaran M.Sc., PGDCA., PGDOR., MCA., M.Phil., Ph.D. He is presently working in Department of Computer Science, Alagappa University; He has plentiful experience in teaching and research. His main areas of research are Information Security and Data Mining.