International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies

http://TuEngr.com

PAPER ID: 11A13D

PERFORMANCE OF BIG DATA ANALYSIS OF SENTIMENTS IN TWITTER DATASET USING SVM MODELS

Preethi Nanjundan [1], K. Maheswari [2*], Jayabrabu Ramakrishnan [3], Dinesh Mavalur [4], Azath Mubarakali [5], S. Ramkumar [6]

[1] *Department of Data Science, Christ (Deemed to be University), Lavasa, Pune, INDIA.*
[2] *Department of Computer Applications, School of Computing, Kalasalingam Academy Research and Education, Tamil Nadu, INDIA.*
[3] *Department of Information Technology, College of Computer Science and Information Technology, Jazan University, SAUDI ARABIA.*
[4] *Department of Information Technology, College of Computing and Informatics, Saudi Electronic University, SAUDI ARABIA.*
[5] *Department of CNE, College of Computer Science, King Khalid University, SAUDI ARABIA.*
[6] *School of Computing, Kalasalingam Academy of Research and Education, Tamil Nadu, INDIA.*

ARTICLEINFO

Article history:
Received 06 March 2020
Received in revised form 09 June 2020
Accepted 24 June 2020
Available online 30 June 2020

Keywords:
Sentiment analysis; Machine learning; SVM linear grid; Twitter dataset; SVM model; SVM radial grid; Kappa value; SVM classifier; Negative sentiment; Neutral sentiment; Positive sentiment.

ABSTRACT

Sentiment analysis uses supervised and machine learning algorithms. The analysis can be done on movie reviews, twitter reviews, online product reviews, blogs, discussion forums, Myspace comments, and social networks. The twitter data set is analyzed using a support vector machine (SVM) classifier with various parameters. The content of the tweet is classified to find whether it contains fact data or opinion data. The deep analysis is required to find the opinion of the tweets posted by the individuals. The sentiment is classified in to positive, negative and neutral. From this classification and analysis, an important decision can be made to improve productivity. The performance of SVM radial kernel, SVM linear grid and SVM Radial Grid was compared and found that SVM linear grid performs better than other SVM models.

Disciplinary: Computer and Information Technology.

1. INTRODUCTION

Sentiment analysis involves in research field of text mining analysis. Many people post their views, opinion, and ideas in an unstructured format. The views are taken from the views of public, customer, social media, entertainment, sports, climate analysis, and Industrial organization, etc. Millions and billions of people and the public are using social network websites such as Facebook, Twitter, Google Plus and so on (Pandey & Iyer, 2009, Da Silva, 2014). The social media generates a huge volume of sentiment data in various forms such as tweet id, status updates, reviews, author, content, tweets type, and tweets status update. As the data size is going larger and larger, it is necessary to analyze and categorize the sentiment reviews or opinions of the various people to predict (Ortigosa et al., 2014; Pandey & Iyer, 2009).

Machine learning Techniques is one of the frequently used techniques in sentiment data analysis to classify the tweets or author comments as positive, negative, or neutral based on the value of the tweet (Boiy & Moens, 2009). SVM is a classifier algorithm that separates hyperplane (Zainuddin et al., 2016; Tripathy et al., 2015; Abbasi et al., 2008).

It is defined as

$$f(x) = w^T x + b \quad (1),$$

where w is the weight factor and b is the bias of the function.

The optimal hyperplane is defined as

$$|w^T x + b| = 1 \quad (2),$$

x is a training example. The distance between x and w^T, $b$is calculated as

$$Distance = \frac{|w^T x+b|}{\|w\|} \quad (3).$$

The margin M is the distance with closeness,

$$M = \frac{2}{\|w\|} \quad (4).$$

The kernel function is a dot product of data inputs. The frequently used kernel functions are Linear, Polynomial, RBF and sigmoid. The linear kernel function is

$$Kernel(X_i, X_j) = X_i \cdot X_j \quad (5)$$

2. REVIEW OF LITERATURE

Nadia et.al.,[1] developed a method that automatically classifies the sentiment of Twitter data by using lexicons and classifier ensembles. A variety of public tweet sentiment datasets is experimented using SVM, Naive Bayes, Random Forest, multinomial Naive Bayes, and logistic regression to improve the classification accuracy. Medhat et al. (2014) survey the recent techniques used in sentiment analysis and the fifty-four articles were classified and summarized. By analyzing the article, the author was given a clear picture of sentiment Classification and feature selection that was used in the field of research. To solve the sentiment classification SVM and Naive Bayes were commonly used in the Machine Learning algorithms.

Haddi et al. (2013) discover the responsibility of text pre-processing in sentiment analysis data, and the experimental results were done using support vector machines (SVM) with appropriate feature selection. Bifet and Frank (2010) used Twitter data provided by Firehouse API, which gave all messages from every user which are publicly available in real-time. The author experimented with multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. Zainuddin et al

(2016) developed a method to classify the twitter-based sentiment using Principal Component analysis which is combined with Sentiwordnet lexicon-based method and incorporated with Support Vector Machine. Liao (2017) proposed a sophisticated neural network approach for analyzing Twitter data to perform sentiment analysis.

Tripathy et al., (2015) compare the experimental result of Naive Bayes and SVM on the polarity movie dataset. The training and testing data is classified based on positive and negative reviews by applying the algorithms. Bifet and Frank, (2010) discuss the challenges faced during the sentiment classification of Twitter dataset flow and the author proposed the method called a sliding window kappa statistic to evaluate the time-changing data streams. Tan (2008) performed the sentiment classification on Chinese documents by applying the various feature selection methods such as IG, CHI, DF, MI, and other classification techniques KNN, SVM, Naive Bayes, and centroid classifier. Abbasi (2008) proposed the feature selection method entropy weighted genetic algorithm (EWGA) and the experimental result was done with the benchmark movie dataset. The results were compared with SVM and EWGA to indicate the rise in the level of performance.

Go et al. (2009) worked with machine learning algorithms such as SVM, Maximum Entropy, and Naive Bayes with an accuracy of 80% and above, and the experiment was done using tweets with emoticon data. The preprocessing is a must to increase accuracy. Agarwal et al. (2011) proposed the method to repeatedly detect the sentiment analysis of twitter data and the experimental result shows the more abstract features are captured, in the removal of noise and more robust about bias. Shreevats & Gustav (2010) observed the sentiment analysis using Twitter data and introduced the POS specific prior polarity features, to discover the use of a tree kernel.

Aramaki et al. (2011) contributed the novel investigation of stacked SVM based classification techniques to categorize some of the user attributes and they performed the complete analysis of components and features. Bifet & Frank (2010) proposed the technique to take out the twitter data related to influenza using twitter API and classify the influenza patients using the Support vector machine. The author portrays the Twitter data texts reflect the real world, and that NLP techniques can be applied to extract only tweets that contain useful information. Boiy & Moens (2009) proposed the sliding window kappa statistic to estimate the time-changing data flow and survey the twitter dataset using the machine algorithms for various data flows.

Haddi et al. (2013) presented the machine learning to analyze the sentiment data-based review. blog and forum text data available in the WWW. The dataset is manually classified based on the tweet statements as negative, positive, and neutral. Ortigosa et al. (2014) investigate the responsibility of text preprocessing in sentiment analysis and experimental results explain with the suitable feature selection and demonstration using SVM. Moraes et al. (2013) proposed a new methodology for sentiment analysis for the Facebook dataset to mine the user's sentiment categories and to notice major emotional modification. Pandey & Iyer, (2009) presents the experimental result by comparing the SVM algorithm and ANN concerning the document-level to analyze sentiment. Li et al. (2017) contribute to the use of two different classifier approaches such as neutral-polar classifier and positive-negative or polarity classifier for the experimental analysis. Shihab (2020) discusses data security/reliability in big data environments.

3. METHODOLOGY

Support Vector Machines (SVM) is a supervised learning method. This is associated with learning algorithms that analyze data in the data set used for classification. SVM model separates and constructs hyperplane. This hyperplane can be used for classification. SVM is a maximum margin classifier. The mapping is done through kernel functions. Some of the kernel functions used in this work are given below. The linear kernel functions are used to deal with large space data vectors and it is defined as

$$k(x, x') = (x, x') \tag{6}$$

The Gaussian radial basis function is a general purpose kernel which is defined as

$$k(x, x') = exp(-sigma\|x - x'\|^2) \tag{7}$$

The radial basis function is

$$\mathrm{K}(x_i, x_j) = exp\left(-\gamma|X_i - X_j|^2\right) \tag{8}$$

An important issue of the SVM classifier is the selection of various parameters. The parameter tuning influences the effectiveness and efficiency of the classifier. The cost parameter C finds the penalty for misclassifications. Figure 1 details the framework of the proposed work.

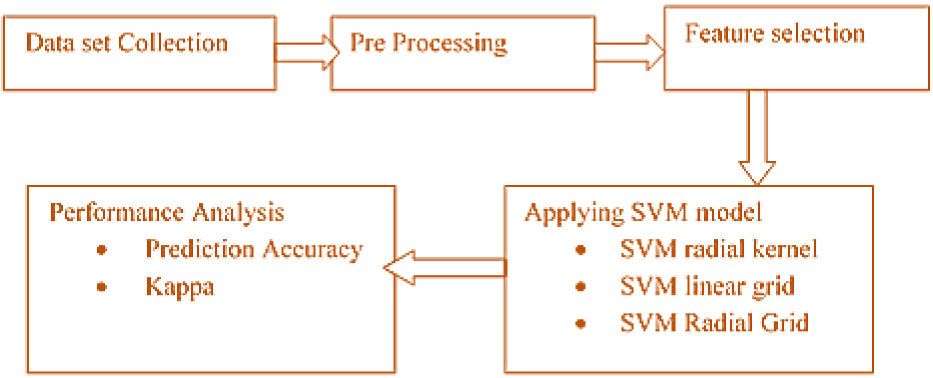

Figure 1: Proposed framework.

4. EXPERIMENTAL RESULTS

There are three different classes of sentiments in the twitter data set. They are

- Positive
- Negative
- Neutral

Positive Sentiment refers to the positive thinking nature of the person. The positive emotions sentiments are surprise, love, affection, happiness, joy, smile, etc. The positive tweet reviews, opinions, and sentiments will create a happy environment and good for the individual as well as for the society and country. Negative Sentiment reflects the negative nature of the individual. The negative sentiments will provide sadness, worry, jealousy, and hate, etc. The negative sentiments of a person will not be happy and it affects society. Neutral Sentiments reflects no participation. The person is not satisfied and could not take any decision. This is also very important for analysis.

This emotion will also affect society without creating more impacts.

Twitter data is obtained from publically available internet. The 1000 tweets were taken for analysis and split into 70% Training set and 30% test set. The twitter dataset is in the form of a CSV file. CSV file is looking like a table structured spreadsheet program of Microsoft Excel. A tweet is a user's opinion and is expressed emotionally by different people. The twitter dataset used in this work is labeled into three classes viz. neutral, negative, and positive. The data may be redundant, inconsistent, unwanted blank spaces, special characters, and non-related information. Pre-processing is essential to clean the data and bring it into a structured one. The following pre-processing was performed (Haddi et al., 2013):

- Removed all punctuations, @, _ symbols and numbers
- A removed sequence of repeated characters
- Replaced all the emotions with their Sentiment value.
- Replaced all missing values with NAN
- Removed Stop Words
- Removed unnecessary white spaces

Machine learning methods can be applied to sentiment analysis. The supervised classification is made for sentiments in the data set. The data set is divided into the training set and test set in machine learning algorithms, there are many algorithms which acts as the main role in sentiment analysis.

- Naive Bayes
- maximum entropy
- support vector machines

SVM & RBF kernel Support vector machine is a non-probabilistic algorithm that is used to separate data linearly and nonlinearly. Here dataset is a set of attributes with its values.

Dataset $\{X_i, X_j\}$, X_i is a set of data present in the dataset as tuple and X_j is a class label of data in the form of tuples. Class labels are 0, 2, and 4 for neutral, negative, and positive category respectively. The objective of SVM is to classify or separate the sentiments as neutral, negative, and positive training by finding $n-1$ hyperplane. Quadratic Programming (QP) problem is essential to solving linear data using SVM model. A classifying hyperplane is written as $w^T x + b = 0$, weight vector w of n number of attributes, X_i is a set of data present in the dataset as tuple and b are biased. A linear classifier has the form $f(x) = w^T x + b$.

Feature selection is a selection of relevant variables or features available in the dataset. It is also known as attributing selection. This is the initial process that is carried out before model construction. Identifying relevant features using the feature selection technique simplifies the task. The strongly correlated attribute provides the idea of classification on what basis. Like Feature selection, Feature extraction can also be used. Feature extraction generates new features from existing features, whereas feature selection creates and returns a new subset of the existing features. If the dataset contains many features and a few samples, then there is a necessity to select features. The Boruta feature selection method was performed in this work. The advantage of feature selection is:

- Models can be simplified
- Training time is minimized
- The dimensionality nuisance is minimized,
- Overfitting is reduced

Boruta is a feature selection method that was proposed to find whether the feature is either strong or weak. The gsub() function was used to replace any unwanted expression into wanted one. Missing values were checked and are replaced by NAN. Blank spaces are cleaned. The syntax of the Boruta package is similar to the syntax of the regression lm method. Boruta provides a well significant call of variables in a data set. In this work, there are 5 attributes, 3 attributes were rejected and 2 attributes were confirmed. Some tentative attributes will be created by Boruta package which is also having importance when the original features are not able to take any decision. The plotting of the Boruta variable is shown in the importance chart. The attributes of the data set were given in the x-axis and its importance on the y-axis. The minimal, average, and maximum *Z* score of shadow attributes are tentative attributes generated by Boruta package. The tentative attributes will be classified as confirmed or rejected by comparing the median *Z* score of the attributes with the median *Z* score of the best shadow attribute. The feature selection of attributes is shown in Figure 2.

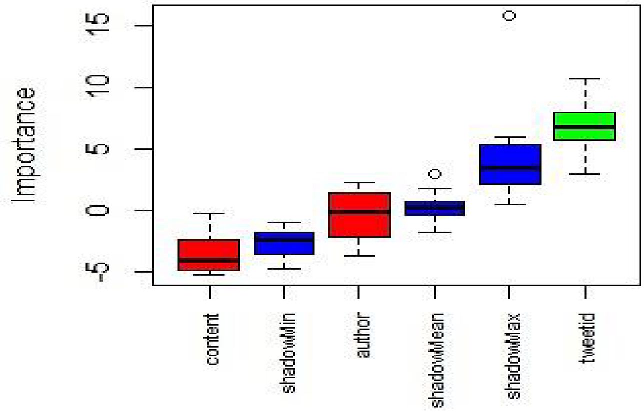

Figure 2: Feature Selection of Attributes.

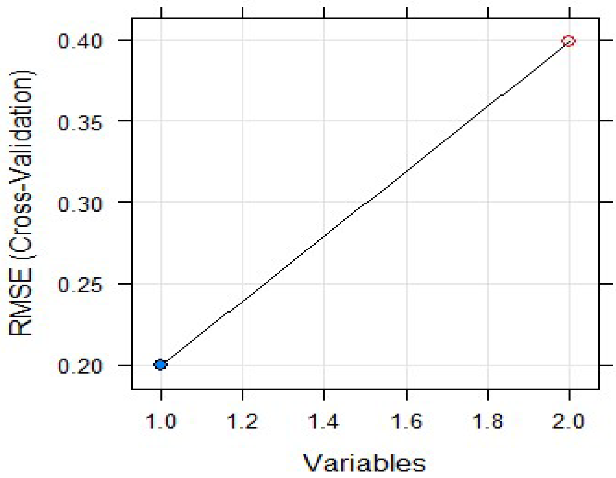

Figure 3: Random Forest Selection.

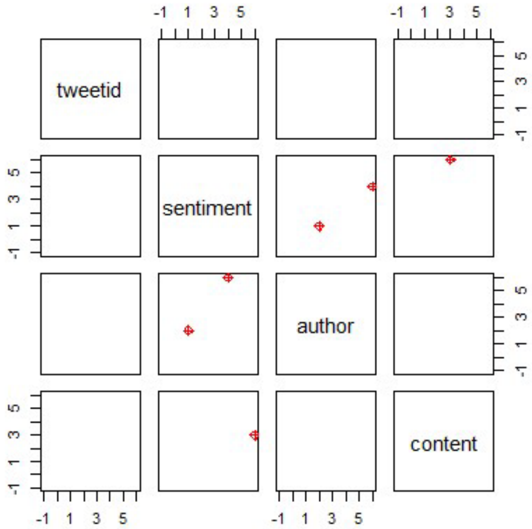

Figure 4: Random Forest Selection of Attributes

The random forest selection function through the rfFuncs option was performed. The top one attribute out of five attributes was selected by this method. The retrain was plotted and is shown in Figure 3. The order of importance of attributes in feature selection by Boruta is shown in Figure 4.

SVM model linear kernel was implemented to classify the sentiments in the twitter data set. Feature selection was performed by Boruta and two attributes confirmed important which are tweetid and sentivalue. The SVM model was plotted with a linear kernel method and is shown in Figure 5. The confusion matrix is used to calculate the accuracy and summarize the performance of data. It gives a clear idea about the efficiency of the model. It provides the user with how much is achieved and how much went wrong. Figure 6 shows the confusion matrix of the model.

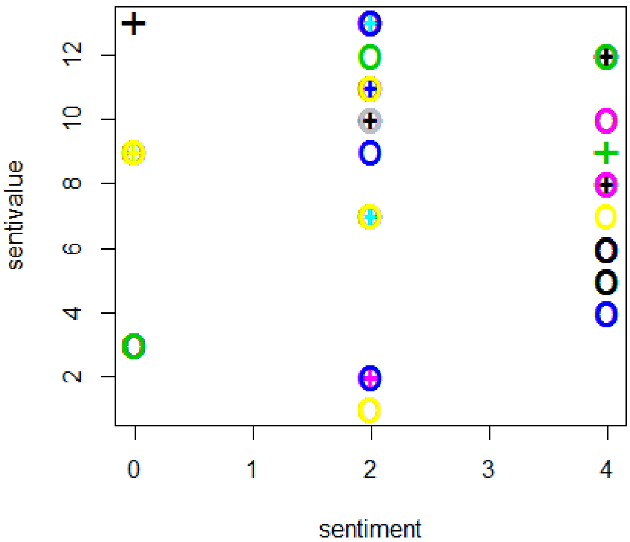

Figure 5: SVM classification.

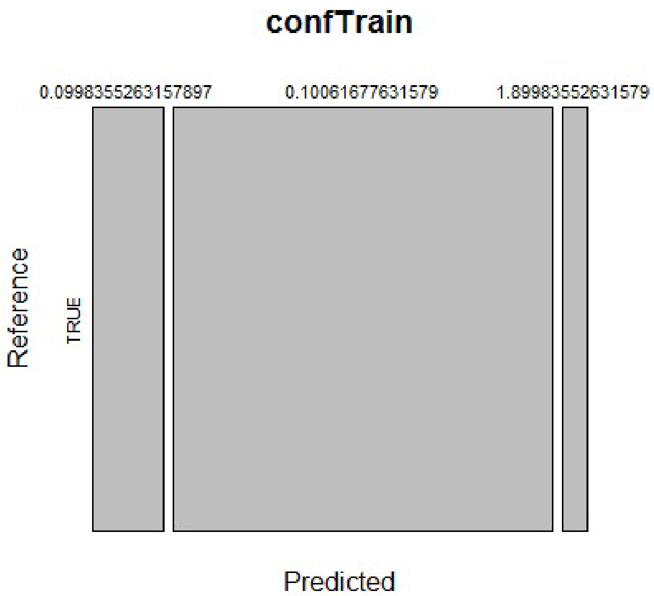

Figure 6: Confusion Matrix of SVM.

The performance of the above model was not satisfied, the improper prediction was performed in the SVM model, and therefore the minimum accuracy of 15% was achieved. So, the tuning of the parameter is necessary for improving performance. The next SVM model of c-classification with radial kernel was implemented. SVM model c – classification with radial kernel was applied for the same 1000 tweets in the twitter data set with c type classification. This model classifies the data set sentiment value as shown in Table 1.

Table 1: Classification sentiments table with radial kernel.

pred	NEGATIVE	NEUTRAL	POSITIVE
4	578	247	175

This model has 701 training samples and 299 testing samples with four predictors (pred) and three classes. The 10 fold cross-validated was performed for three times for the sample sizes 631. The tuning parameter was kept constant with a value of 1. Table 2 shows the accuracy and kappa values.

Table 2: Classification Prediction Accuracy

Accuracy	Kappa
0.8497393	0.7063484

Kappa processes the percentage of data values in the diagonal elements of the Table. These values are adjusted for the agreement than expected. Kappa value is calculated by the observed level of agreement

$$K_0 = K_{11} + K_{22} \quad (9).$$

This value needs to be compared to the value that you would expect if the two raters were independent,

$$K_e = K_1 K_1 + K_2 K_2 \quad (10).$$

The Kappa value is defined to be

$$Kappa = K_0 - K_e / 1 - K_e \quad (11).$$

The kappa result can figure out

$Kappa < 0.20$	Poor agreement
$Kappa$ = 0.20-0.40	Fair agreement
$Kappa$ = 0.40-0.60	Moderate agreement
$Kappa$ = 0.60-0.80	Good agreement
$Kappa$ = 0.80-1.00	Very good agreement

The predicted accuracy of three classes sentiment counts was given in Table 3.

Table 3: Prediction Accuracy

pred	0	2	4
neutral	78	0	0
negative	0	168	45
positive	0	1	7

The predicted accuracy of repeated cross-validation and cost is shown in Figure 7. The accuracy is 84% and the kappa is 70%. The kappa value indicates that there is a good agreement between individuals.

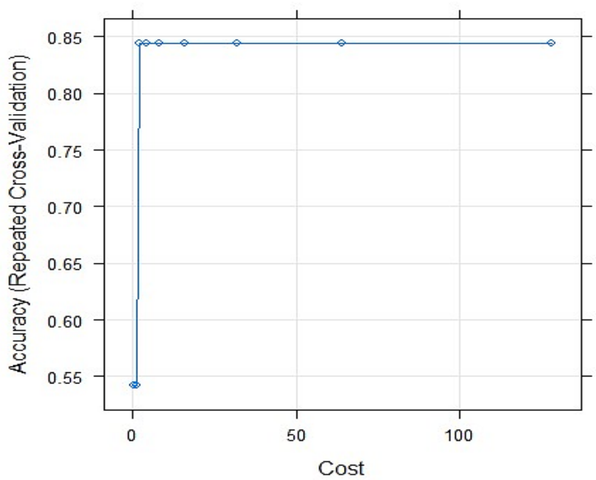

Figure 7. SVM Radial Kernal classification Accuracy

SVM_Linear_Grid is another model performed to increase the accuracy with 76 samples, 4 predictors, and 3 classes. The Resampling of the dataset is Cross-Validated with 10 fold and repeated three times with sample sizes 69, 67, and 68. The accuracy and kappa values of SVM_linear_grid are shown in Table 4.

Table 4: Prediction Accuracy of SVM_linear_grid

C	Accuracy	Kappa
0.00	NaN	NaN
0.01	0.8936508	0.7973095

Accuracy was used to select the optimal model using the largest value. The final value used for the model was C = 0.01. The accuracy is improved from 84% to 89%. The accuracy of various c values and cross-validation is shown in Figure 8. The SVM linear grid plot was applied in the same data set. The accuracy of predicted classification with SVM linear grid method is 89%.

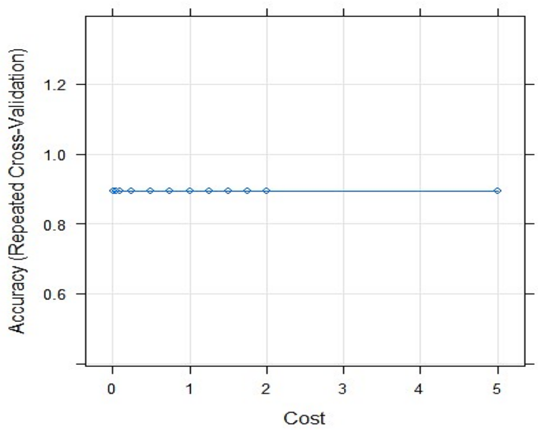

Figure 8. SVM Linear grid Accuracy

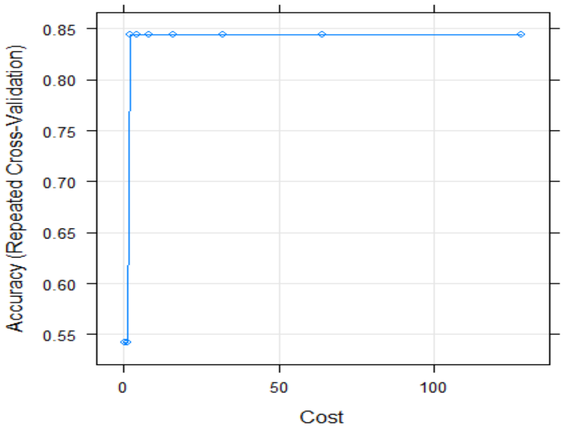

Figure 9: SVM Radial Basis Function Kernel Accuracy

Support Vector Machines with Radial Basis Function Kernel with 76 samples, 4 predictors, and 3 classes of resampling of 10-fold cross-validation and is repeated 3 times was applied for varying cost value. The tuning parameter 'sigma' was held constant at a value of 0.003044357. Accuracy was used to select the optimal model using the largest value. The final values used for the model were sigma = 0.003044357 and C = 2. The accuracy obtained was 84%. Figure 9 plots the accuracy.

SVM_Radial_Grid method is applied with 76 samples of 4 predictors and 3 classes. Support Vector Machines with Radial Basis Function Kernel summary sample size is 69, 68, 69, 68, 68. The prediction accuracy of the SVM Radial grid is shown in Tables 5 and 6.

Table 5: Prediction Accuracy of SVM Radial grid

prediction	0	2	4
0	7	0	0
2	3	16	5
4	0	0	0

Table 6: Prediction Accuracy of SVM_linear_grid

Accuracy	Kappa
0.8419	0.5108

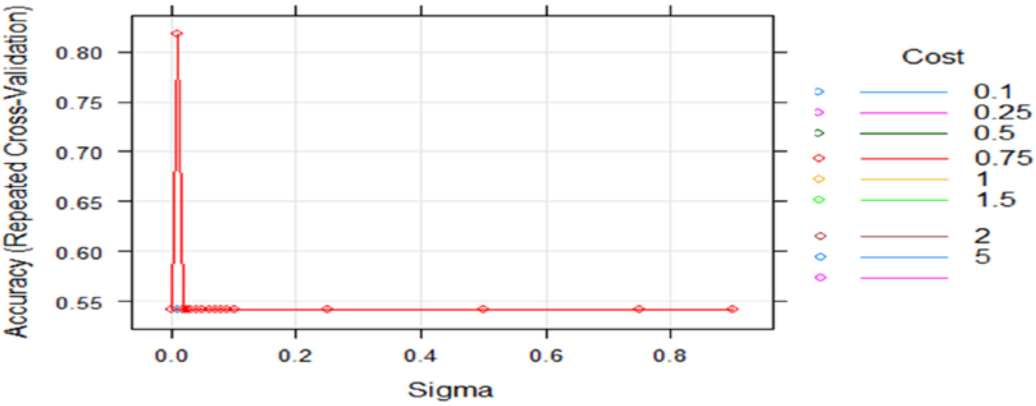

Figure 10: SVM Radial Grid Accuracy

Accuracy was used to select the optimal model using the largest value. The final values used for the model were sigma = 0.01 and C = 1.5. The accuracy of 84% was maintained.

The performance of various SVM model statistics by three classes is given in Table 7.

Table 7: Comparative Analysis of SVM Models

Pred	0	2	4
Sensitivity			
SVM radial kernel	1.0000	1.0000	0.13462
SVM linear grid	0.7000	1.0000	0.20000
SVM Radial Grid	0.7000	1.0000	0.0000
Specificity			
SVM radial kernel	1.0000	0.6538	0.99595
SVM linear grid	1.0000	0.5333	1.00000
SVM Radial Grid	1.0000	0.4667	1.0000
Pos Pred Value			
SVM radial kernel	1.0000	0.7887	0.87500
SVM linear grid	1.0000	0.6957	1.00000
SVM Radial Grid	1.0000	0.6667	NaN
Neg Pred Value			
SVM radial kernel	1.0000	0.9884	0.84536
SVM linear grid	0.8750	1.0000	0.86667
SVM Radial Grid	0.8750	1.0000	0.8387
Prevalence			
SVM radial kernel	0.2609	0.5652	0.17391
SVM linear grid	0.3226	0.5161	0.16129
SVM Radial Grid	0.3226	0.5161	0.1613
Detection Rate			
SVM radial kernel	0.2609	0.5619	0.02341
SVM linear grid	0.2258	0.5161	0.03226
SVM Radial Grid	0.2258	0.5161	0.0000
Detection of Prevalence			
SVM radial kernel	0.2609	0.7124	0.02676
SVM linear grid	0.2258	0.7419	0.03226
SVM Radial Grid	0.2258	0.7742	0.0000
Balanced Accuracy			
SVM radial kernel	1.0000	0.8240	0.56528
SVM linear grid	0.8500	0.7667	0.60000
SVM Radial Grid	0.8500	0.7333	0.5000

The true positive rate is test sensitivity and the true negative rate is test specificity. Sensitivity measures the correctly identified positives. Specificity measures the correctly identified true negatives. The performance analysis of the algorithms can be estimated by using two measures, sensitivity S and positive predictive value PPV. They are very much useful for the estimation of performance. Sensitivity S is measured as

$$S = TP/(TP + FN) \quad (12).$$

Specificity S is measured as

$$SP = TN/(TN + FP) \quad (13),$$

where TP is a number of truly relevant features recognized by an algorithm, FN is a number of truly relevant features that are not recognized by an algorithm and FP is a number of non-relevant features that are incorrectly recognized as relevant. Positive predictive value PPV is measured as

$$PPV = TP/(TP + FP) \quad (14).$$

The negative predictive value PPV is measured as

$$NPV = TN/(TN + FP) \quad (15).$$

The SVM classifier with SVM kernel, SVM linear grid, SVM Radial Grid were performed for twitter data set and performance was analyzed with various tuning parameters. SVM linear grid achieves better performance than the other. The basic SVM model without parameter tuning achieves minimum level accuracy. The overall performance is shown in Table 8. From the experimental results, it is observed that, the performance of the SVM linear grid model shown better results which are shown in Figure 11.

Table 8: Comparative Analysis of Accuracy of SVM Models

Model	Accuracy	Kappa
SVM	0.15	0.12
SVM radial kernel	0.8497393	0.7063484
SVM linear grid	0.8936508	0.7973095
SVM Radial Grid	0.8419	0.5108

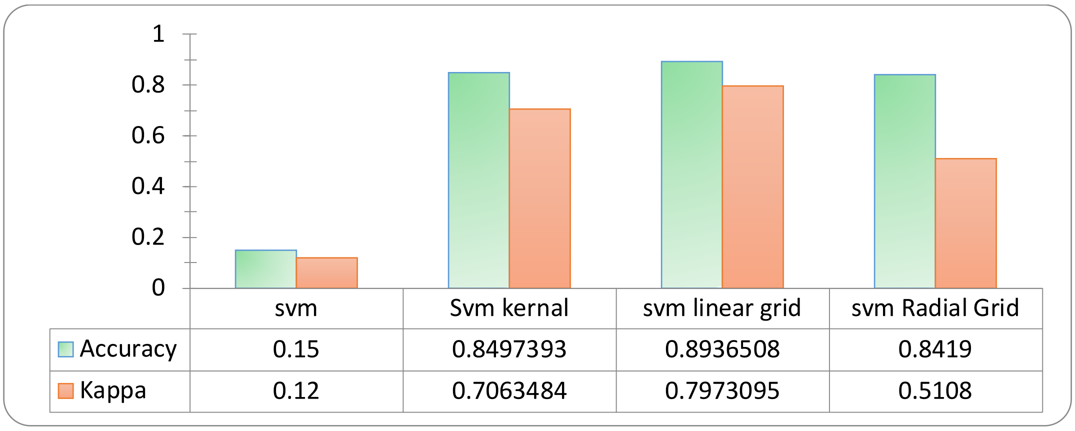

Figure 11: Accuracy of SVM Models.

The performance measures were selected to compare the SVM classifier with various parameter tuning model, the highest accuracy of 89.36% of SVM linear grid, 84.97% of SVM kernel, 84.19 of SVM radial grid was achieved. In SVM, the values of C and gamma were selected using 5-fold cross-validation and 10-fold cross-validation.

5. CONCLUSION

The experimental results of the SVM classifier attain satisfying results in the classification when compared with Li et al. (2017). There are many performance measures of the SVM model is available. In this work, the parameter tuning of various SVM model was proposed. SVM model Cross-Validated with 10 fold and repeated 3 times with sample sizes 69, 67 and 68 outperforms than the other models. The accuracy achieved was satisfied because twitter posts are emotional ones. The people post their emotions when they are not feeling comfortable. Neutral data plays a more important role because people could not take any decision. The concentration towards the neutral people will reflect either positive or negative. This idea is suited for election when the people are not ready to vote for any particular option. Future work will focus on implementing different algorithms with a different dataset.

6. AVAILABILITY OF DATA AND MATERIAL

Information can be made available by contacting the corresponding author.

7. REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, *26*(3), 1-34.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (30-38).

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. *Proceedings of the conference on empirical methods in natural language processing*, 1568-1576.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. *International conference on discovery science*, Springer, 1-15.

Boiy, E., & Moens, M.F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval*, *12*(5), 526-558.

Da Silva, N. F., Hruschka, E.R., & Hruschka Jr, E.R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170-179.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12).

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*, 26-32.

Liao, S., Wang, J., Yu, R., Sato, K., & Cheng, Z. (2017). CNN for situations understanding based on sentiment analysis of twitter data. *Procedia computer science*, *111*, 376-381.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

Moraes, R., Valiati, J.F., & Neto, W.P.G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621-633.

Ortigosa, A., Martín, J. M., & Carro, R.M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*, *31*, 527-541.

Pandey, V., & Iyer, C. (2009). Sentiment analysis of microblogs. CS229: Machine learning final projects, Cite seer.

Shreevats, A., & Gustav, M. (2010). Classifying latent user attributes in twitter, Proceedings of the 2nd international workshop on Search and mining user-generated content, 37-44.

Shihab, L.A. (2020). Technological Tools for Data Security in the Treatment of Data Reliability in Big Data Environments. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, 11*(9), 11A9M, 1-13.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, *34*(4), 2622-2629.

Tripathy, A., Agrawal, A., & Rath, S.K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*, *57*, 821-829.

Zainuddin, N., Selamat, A., & Ibrahim, R. (2016). Twitter feature selection and classification using support vector machine for aspect-based sentiment analysis. *International conference on industrial, engineering and other applications of applied intelligent systems*, Springer, 269-279.

Dr. Preethi Nanjundan is an Assistant Professor in Christ(Deemed to be University) , Lavasa campus, Pune. She received her Ph.D degree in Semantic Web in 2014 from Bharathiar University, Coimbatore, India. Her research interests are Semantic web, Machine learning, Deep Learning etc. preethi.n @ christuniversity.in

Dr. K. Maheswari is an Assistant Professor, the Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, India. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing and Text Analysis.

Dr Jayabrabu Ramakrishnan is at Department of Information Technology, College of Computer Science and Information Technology, Jazan University, Saudi Arabia. He obtained a Ph.D in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India. His area of research includes Data Mining, Intelligent Agents and Information Systems. Jayabrabu @ gmail.com

Dr. Dinesh Mavalur Balasubramanyam is an Assistant Professor at Department of Information Technology, College of Computing and Informatics, Saudi Electronic University, Saudi Arabia. His research interests span both data science and network science. Much of his work has been on Improving the Understanding, Design, and Performance of Parallel and Networked Computer Systems, mainly through the application of Data Mining, Statistics, and Performance Evaluation. d.mavaluru @ seu.edu.sa

Dr. Azath Mubarakali is an Assistant Professor at Department of CNE, College of Computer Science, King Khalid University, Saudi Arabia. He had his Bachelor's, Master's, and Doctorate degrees from Anna University. His research interest includes Network Security, Software Defined Networks, Block Chain Technology, Networking, Wireless networks and Mobile Computing. Mailmeazath @ gmail.com

Dr. S. Ramkumar: is an Assistant Professor, the Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil. His field of interest is biosignal processing, AI, HCI, BCI, and machine vision. He authored the textbook Java for Beginners and Magic Book for Quantitative Aptitude and Reasoning (Master the Tricks in 10 days). ramkumar.drl2013 @ gmail.com