SVM Optimization for Sentiment Analysis

Munir Ahmad, Shabib Aftab, Muhammad Salman Bashir, Noureen Hameed, Iftikhar Ali, Zahid Nawaz

Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

Abstract—Exponential growth in mobile technology and mini computing devices has led to a massive increment in social media users, who are continuously posting their views and comments about certain products and services, which are in their use. These views and comments can be extremely beneficial for the companies which are interested to know about the public opinion regarding their offered products or services. This type of public opinion otherwise can be obtained via questionnaires and surveys, which is no doubt a difficult and complex task. So, the valuable information in the form of comments and posts from micro-blogging sites can be used by the companies to eliminate the flaws and to improve the products or services according to customer needs. However, extracting a general opinion out of a staggering number of users' comments manually cannot be feasible. A solution to this is to use an automatic method for sentiment mining. Support Vector Machine (SVM) is one of the widely used classification techniques for polarity detection from textual data. This study proposes a technique to tune the SVM performance by using grid search method for sentiment analysis. In this paper, three datasets are used for the experiment and performance of proposed technique is evaluated using three information retrieval metrics: precision, recall and f-measure.

Keywords—Sentiment analysis; polarity detection; machine learning technique; support vector machine (SVM); optimized SVM; grid search technique

I. INTRODUCTION

In order to compete in market, it is essential for the organizations and companies to get aware of the consumer's opinion regarding the products and services, they provide. Now days micro-blogging websites are the rich sources of textual data including opinions and reviews about products, services, brands, movies, music, news and politics, etc. These sources can be used to extract public opinion about anything i.e. from comments regarding particular product or brand to the views about next election etc. In past, this type of information was obtained through telephonic and door to door surveys which were really time consuming and complex task. So the need of automatic technique for opinion extraction from textual data brought the researchers to this domain. Today, there are fundamentally three approaches available to extract the opinion from text: Lexicon driven [1], machine learning based [2] and finally the hybrid of both [3]. The lexicon-based approach depends upon its dictionary of weighted words to generalize the polarity of any given text. It does not require any form of prior training as its every aspect is pre-defined and pre-programmed. Machine learning approach can be further categorized as supervised, un-supervised and semi supervised technique. Supervised technique needs to get trained in order to perform

classification. For training purpose pre-classified/pre-labeled data (training data) is used and then it can be capable to classify the real input data (test data) [2], [4], [5]. The hybrid approach integrated both approaches (lexicon driven and machine learning) and usually brings more accurate results. SVM is one of the widely used supervised machine learning algorithms for sentiment analysis. It has proved to be highly effective in categorization of traditional texts. This research intends to tune the performance of SVM for sentiment analysis using grid search technique. Optimized Sentiment Analysis Framework (OSAF) is used in this research which is an extension of Sentiment Analysis Framework (SAF) [5]. OSAF consist of four phases: Dataset, Pre-processing, Classification and Results. Tuning is performed in the classification phase. Three datasets were used in this research, two from twitter and one from IMDB. In proposed technique, Grid-Search keeps tweaking the SVM's parameters and compares the accuracy of the obtained results with the pre-labeled data and chooses the parameters with best result. On the other hand cross-validation takes the accuracy further up to a notch by changing the testing data ratio to the training data in accordance with a pre-defined range until it obtains the highest possible accuracy. To evaluate the performance of proposed technique three information retrieval metrics are used precision, recall and f-measure.

Remaining paper is organized as follows. Section II is about related work in this domain. Section III elaborates the materials & methods used for this research. Section IV presents results & discussion. Section V finally concludes the paper.

II. RELATED WORK

Many researchers have been working to extract the sentiment from textual data by using data mining techniques. Some of the selected studies are discussed here. In [6], J48 and MLP were used to classify five datasets. Performance was evaluated in terms of TP rate, FP rate, Precision, Recall, F-measure and ROC Area. According to results, MLP performed better on each dataset. Neural Network also showed better learning capability and was suggested by the authors as good alternate option for classification. Authors in [7] presented an application for Arabic sentiment analysis of twitter data. Machine learning techniques: NB and SVM were used to analyze 1000 tweets for polarity detection. Feature vector approach was used with machine learning classifiers to improve the accuracy. Some problem areas were also identified by the researchers regarding training data such as multiple occurrences of tweets, opinion spamming and dual opinion tweets. It was mentioned that issues like these can

compromise the level of accuracy. Authors in [4] performed sentiment classification of Arabic tweets by using three machine learning algorithms: Naïve Bayes, Decision Tree and Support Vector Machine. Authors in this research used a framework for classification which includes: Term Frequency-Inverse Document Frequency (TF-IDF) and Arabic stemming as sub tasks. One dataset was used for all three techniques to evaluate the performance. Accuracy was measured in terms of three information retrieval metrics: precision, recall, and f-measure. Authors in [8] presented a feature vector technique by dividing the feature extraction process in two steps, where process starts with the extraction of twitter specific features and then integrated with feature vector. After that these features are removed from the tweets and then again the feature extraction process is completed, like the case with normal text. These extracted features are also added to the feature vector. The accuracy of the proposed feature vector technique is same for Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. In [9], students' academic performance was analyzed and predicted by using three data mining techniques: Decision tree (C4.5), Multilayer Perception and Naïve Bayes. These techniques were applied on the student's data of 2 undergraduate courses from two semesters. According to the results, Naïve Bayes showed the prediction accuracy of 86% which was higher than MLP and Decision tree. This type of prediction can help the teachers to detect those students in early stage, who are expected to get F grade. So ultimately, with the teacher's special care to those students, the academic performance can be improved. In [10], authors predicted the rainfall in Malaysia by using five classification algorithms: Naïve Bayes, Decision Tree, Support Vector Machine, Neural Network and Random Forest. A comparative analysis was performed to identify the technique (s), which can give good result with low training data. Results showed that Decision Tree and Random Forest both have potential to bring higher F-measure after getting trained with lower amount of training data. Support Vector Machine and Naive Bayes brought lower F-measure, when got trained with little amount of data. Neural Network showed less performance in prediction even after getting trained with large amount of training data. In [11], authors performed a comparative analysis of various data mining techniques upon the practitioner's decision of potential therapy change for 40 patients of posterior cancer. J48, MLP, Naïve Bayes, Radial Basis Function and K-Nearest Neighbor were used for classification. According to results, Radial Basis Function performed well and brought high accuracy and Kappa score with low error rate. In [5], performance of SVM was analyzed for polarity classification of textual data. The authors have proposed Sentiment Analysis Framework which consists of 4 phases: Dataset, Pre-processing, Classification and Results. Three datasets were used for the experiment, two from twitter and one from IMDB reviews. Performance of SVM was evaluated for each dataset by keeping in view three different ratios of training data and test data: 70:30, 50:50 and 30:70. Precision, recall and f-measure were used for performance evaluation. In [12], the authors have used SVM for sentiment analysis of twitter data. In this experiment, default parameters were selected in Weka along with 10 fold cross validation technique. Performance of SVM was analyzed on two datasets of pre-labeled tweets. Accuracy was evaluated In terms of Precision, Recall and F-Measure. In [13], authors conducted a systematic literature review on sentiment analysis. Latest research regarding use of SVM for sentiment analysis was the focus of this SLR and published research papers from year 2012 to 2017 were considered. Total of 901 articles were collected and by following a thorough systematic approach, 8 research articles were selected for critical review. Research objectives were identified in the form of research question and during the critical review the answers to those questions were provided.

III. Materials and Methods

This research aims to optimize the performance of SVM for sentiment analysis using grid search technique. To analyze the performance of proposed technique, results are compared with three pre-labeled datasets, two from twitter [14], [15] and one from IMDB reviews [16]. An Optimized Sentiment Analysis Framework (OSAF) is used in this research (Fig. 1), which is an extension of Sentiment Analysis Framework (SAF) [5]. The proposed framework consists of four phases: Dataset, Pre-processing, Classification and Results. Dataset phase deals with the insertion of data into the WEKA (Waikato Environment for Knowledge Analysis) environment on which the classification is going to be performed. Pre-processing Phase deals with the conversion of strings into vectors, which is a pre requisite process for the processing of SVM. This phase has further five steps: 1) Term Frequency-Inverse Document Frequency (TF-IDF), 2) Stemming, 3) Stop Words, 4) Tokenizing and 5) Words to Keep. Classification phase deals with the working of SVM in WEKA using K-fold cross validation grid search technique. Result phase produces the results in the form of tables and graphs. For performance evaluation, three information retrieval metrics are used: precision, recall and f- measure.

A. WEKA

This study used WEKA for the implementation of SVM grid search technique. WEKA is one of the widely used data mining software, developed in Java language at the University of Waikato, New Zealand [17]. The reason behind the widely acceptance is it's easy to use GUI interface for different functionalities such as data analysis, classification, predictive modeling and visualizations. Further advantages of this software include its general public license and its portability.

B. Data Sets

Three datasets are used in this research (Table I). First dataset [14] consisted of tweets related to following four topics: 'Apple', 'Google', 'Microsoft' and 'Twitter'. It contains 571 positive, 519 negative, 2331 neutral and 1689 irrelevant tweets. In the second dataset [15], tweets are related to U.S. airlines and categorized as 2362 'positive', 9178 'negative' and 3099 'neutral'. Third dataset [16] was taken from the Internet Movie Database (IMDB) reviews and contains 1000 positive and 1000 negative comments.
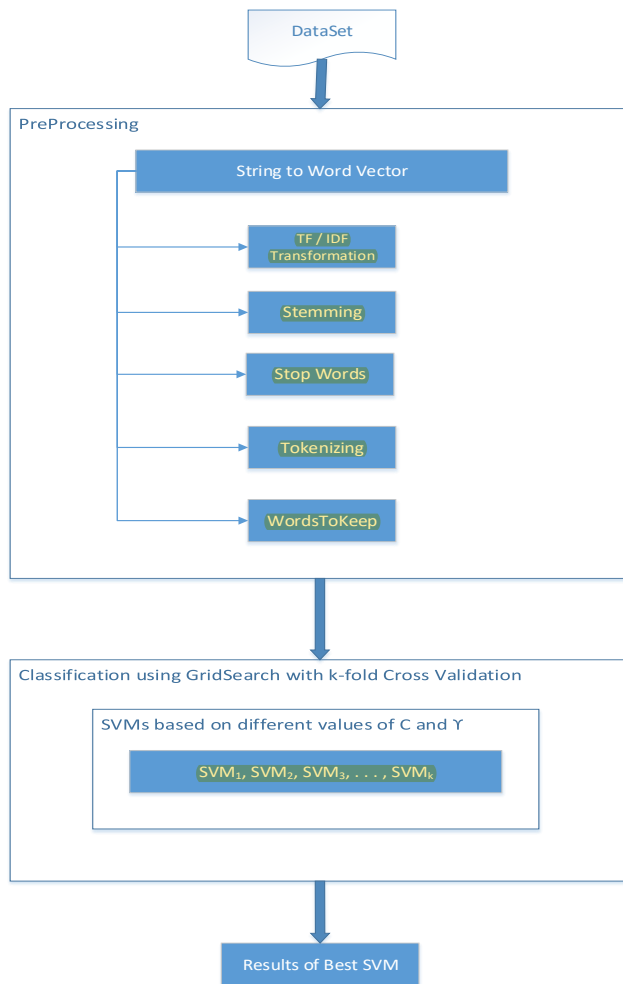
Fig. 1. Optimized Sentiment Analysis Framework (OSAF).

TABLE I.   DATASETS DETAIL

Source	Positive	Negative	Neutral	Irrelevant	Total
Twitter [14]	571	519	2331	1689	5110
Airline [15]	2362	9178	3099	-	14639
IMDB [16]	1000	1000	-	-	2000

Dataset phase deals with the arrangement of relevant data and its transformation to CSV/ARFF format to use in WEKA Workbench [17]. Simple CLI can be used to convert text files into ARFF format using "WEKA.core.converters.TextDirectoryLoader" function.

C. Pre-Processing

It is the most important phase of the framework. Purpose of this phase is to normalize the data by converting the strings into vectors for the classification process. Following sub tasks are performed in this phase.

1) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF provides important information in pre-processing phase by evaluating the frequency of useful words, which essentially makes the sentiment detection process easy. Frequency of these terms plays an important role in

identification of important information as explained by [18]. For example, frequently appearing words in a text document can be 'Good', 'Bad', 'Happy' or 'Sad' etc. Identification and frequency of these words can play a vital role in the process of opinion mining. Term Frequency (TF) is the number of occurrences of a term in a given document and can be calculated with following equation:

$$W_d(t) = TD(t, d) \qquad (1)$$

Where TD corresponds to frequency of term t in a given document d. TF-IDF contains the inverse document frequency (IDF) that reverts higher weight-age for rare conditions while maintaining lower weight-age for common conditions as explained by [19]. IDF can be calculated with following equation:

$$IDF_t = \log\left(\frac{N}{N_t}\right) \qquad (2)$$

Where N represents number of documents and N_t represents the number of terms. When both TF and IDF parameters are set to true, the results are calculated using the following equation:

$$W_t = TF(t, d).IDF_t \qquad (3)$$

In WEKA, TF & IDF transformations are also available along with other filters.

2) Stemming

The process of stemming is immensely useful in many areas of computational linguistics and information retrieval as it reduces all words with the same stem/ base to a common form [20], for example, the word 'working' will be stemmed in to 'work' and so on. Word stemming is one of the essential features of pre-processing in text mining [21]-[25]. In this study, "IterativeLovinsStemmer" is selected in WEKA as the word stemmer in the pre-processing phase. It is based on the LovinsAlgortihm which was the first Stemming algorithm by Lovins JB in 1968 [26].

3) MultiStopWords

The Concept of stop words was originally introduced by [27]. These are common high frequency words like "A", "the", "of", "and", "an". This data is not use full and also does not affect the performance of classification; thus, it has to be removed. There are several methods available for stop word removal as explained by [20], [21], [23], [28], [29]. In this research, "MultiStopwords" was selected for stop words criterion for the pre-processing phase in WEKA.

4) N-GramTokenizer

"N-GramTokenizer" was selected as the Tokenizer in WEKA for pre-processing of data. It first breaks the text into words whenever one of the listed specified characters is detected in it. Afterwards it emits N-Grams of each word of the specified length.

5) WordstoKeep

1000 words were short-listed in "wordstokeep" parameter to narrow down the results within a limited amount of time.

After applying these parameters, pre-processing was performed on all three datasets. Then the processed datasets were forwarded to the classifier.

D. *Classification*

In supervised machine learning approach, first the algorithm has to get trained with pre classified data (training data) with which it makes rules for classification and then on the basis of these rules it classifies the input data (test data). For performance analysis of any supervised machine learning algorithm, pre classified data is provided as test data and then the results of the algorithm can be compared with this pre labeled data. Same strategy is used is this study to analyze the performance of proposed grid search technique. Pre labeled datasets are obtained from social forums: Twitter and IMDB. For classification, Support Vector Machine (SVM) with grid search and K-fold cross validation technique is used. Grid-Search is basically a model for hyper parameter optimization. Hyper parameter tuning is an important task in SVM to extract more accurate results [30]-[33]. In Grid-Search, different models having different parameter values are trained and then evaluated using cross validation. For an RBF kernel, there are two parameters: C and Υ. It cannot be ascertained in advance that which values of C and Υ are best suited for a given problem, so an optimized model is required which can identify the ideal pair of values for these parameters to achieve maximum accuracy. The process of 10-k cross validation is performed on each model of C and Υ and the pair with optimum results is selected. Cross validation is a method used to test multiple models under a particular classifier with the subset of input data as explained by [35]. For K-fold cross validation, the training data is first divided into k subsets of same size. One subset is tested using the classifier on the remaining k-1 subsets. The cross validation procedure can prevent the overfitting problem [34], [36], a binary classification problem is shown in Fig. 2 to illustrate this issue. Filled circles and triangles are the training data while hollow circles and triangles are the testing data. The testing accuracy of the classifier in Fig. 2(a) and (b) is not good as it overfits the training data. On the other hand, the classifier in Fig. 2(c) and (d) does not overfit the training data and gives better accuracy with cross validation.
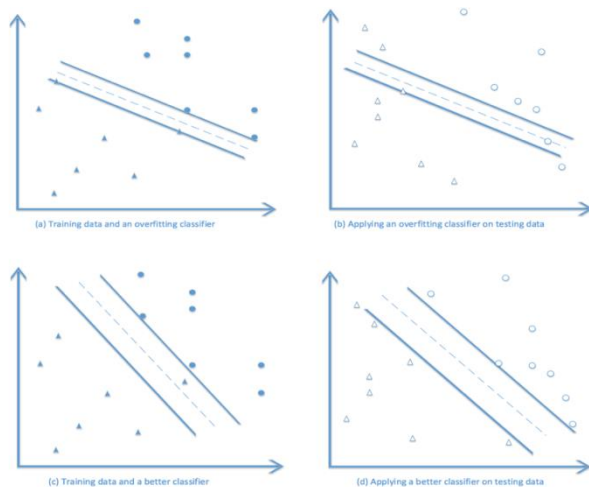

Fig. 2.   Over-fitting with cross validation [36].

IV.   RESULTS AND DISCUSSION

Performance of proposed technique is evaluated for all of three selected datasets and for this purpose three information retrieval parameters are used: Precision, Recall and F-Measure.

The precision can be calculated using TP and FP rate as shown below:

$$Precision = \frac{TP}{(TP + FP)} \qquad (4)$$

TP is used for the sentences which are correctly classified whereas FP is used for sentences, which are wrongly classified.

Recall can be calculated as shown below:

$$Recall = \frac{TP}{(TP + FN)} \qquad (5)$$

FN is used for non-classified sentences and TP are the sentences, which are correctly classified (as explained above).

F-measure can be computed as below:

$$F\text{-measure} = \frac{Precision * Recall * 2}{(Precision + Recall)} \qquad (6)$$

A. *Results with the First Dataset*

First dataset is taken from [14] and consisted of tweets about following topics: 'Apple', 'Google', 'Microsoft' and 'Twitter'. According to results the average Precision, Recall and F-Measure scores are 0.745, 0.752 and 0.747, respectively.

By keeping in view the f-measure score for each class, it can be analyzed that the proposed technique performed well for irrelevant class with the score of 0.87. Detailed results are available in Table II whereas shown graphically in Fig. 3.

TABLE II.   RESULTS OF FIRST DATASET

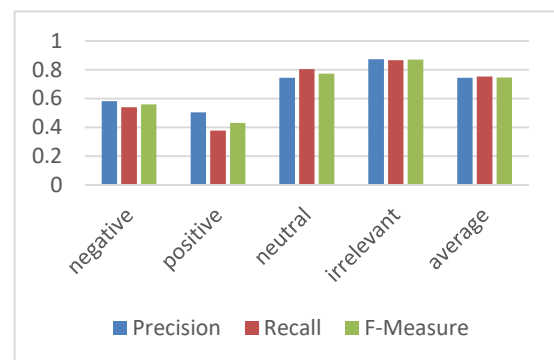| Class	Precision	Recall	F-Measure
Negative	0.583	0.539	0.561
Positive	0.504	0.378	0.432
Neutral	0.745	0.804	0.773
Irrelevant	0.872	0.867	0.87
Average	0.745	0.752	0.747

Fig. 3.   Results of first dataset.

B. Results with the Second Dataset

Second dataset is obtained from [15] and contained the tweets regarding US airlines. The results show that the average Precision, Recall and F-Measure values are 0.786, 0.79 and 0.787 respectively (Table III and Fig. 4). Highest f-measure is reported in negative class which is 0.868.

TABLE III.   RESULTS OF SECOND DATASET

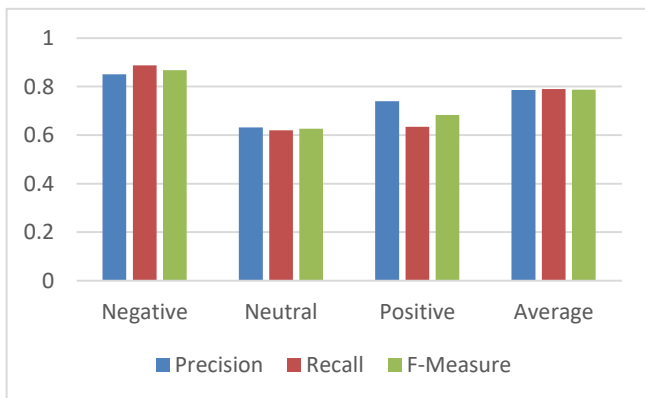| Class	Precision	Recall	F-Measure
Negative	0.85	0.887	0.868
Neutral	0.632	0.62	0.626
Positive	0.74	0.634	0.683
Average	0.786	0.79	0.787

Fig. 4.   Results of second dataset.

C. Results with the Third Dataset

Third dataset is taken from [16] and consisted of IMDB reviews. The results show that Precision, Recall and F-Measure values on average are 0.841, 0.841 and 0.841 respectively. Highest f-measure is reported in negative class with the score of 0.843 (Table IV, Fig. 5).

TABLE IV.   RESULTS OF THIRD DATASET

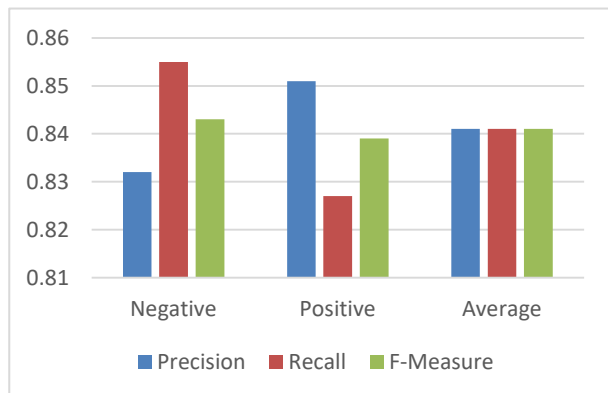| Class	Precision	Recall	F-Measure
Negative	0.832	0.855	0.843
Positive	0.851	0.827	0.839
Average	0.841	0.841	0.841

Fig. 5.   Results of third dataset.

D. Performance Evaluation of Grid Search SVM

To evaluate the performance of grid search technique, f-measure scores of this research are compared with f-measure scores achieved with 70:30 in [5]. In that study, Sentiment Analysis Framework (SAF) was proposed and SVM was used along with various proportions of training and test data. The selected study is an ideal benchmark for comparison as the datasets used in both the studies are same. For comparison, f-measure score is considered as it is the average of precision and recall. OSAF used in this research is the extension of SAF. F-measure scores of both the studies are compared in Table V and graphically represented in Fig. 6. It can be seen that the grid search technique explored in this research has improved the performance in each class of all three data sets.

TABLE V.   COMPARISON OF SAF AND OSAF

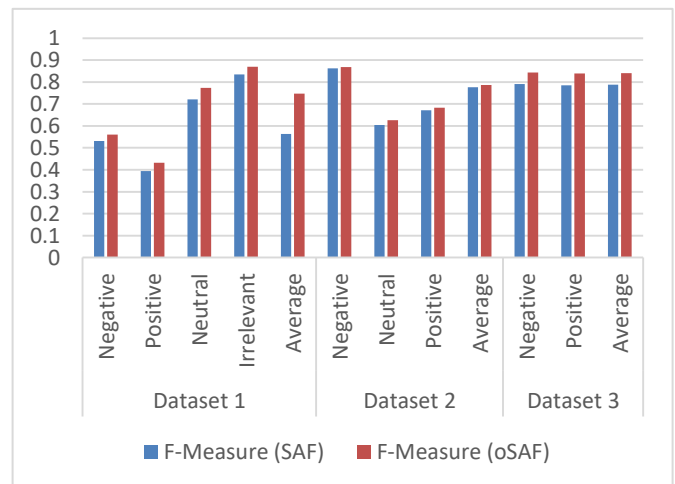| Dataset	Class	F-Measure (SAF)	F-Measure (OSAF)
Twitter [14]	Negative	0.532	0.561
	Positive	0.394	0.432
	Neutral	0.721	0.773
	Irrelevant	0.835	0.870
	Average	0.563	0.747
Twitter [15]	Negative	0.862	0.868
	Neutral	0.604	0.626
	Positive	0.671	0.683
	Average	0.776	0.787
IMDB [16]	Negative	0.791	0.843
	Positive	0.785	0.839
	Average	0.788	0.841

Fig. 6.   Comparison of SAF and OSAF.

V.   CONCLUSION AND FUTURE WORK

In this research, Grid- search technique for SVM is explored in Weka. SVM is one of the widely used supervised machine learning algorithms for sentiment analysis. Optimized Sentiment Analysis Framework OSAF is proposed in this study for text classification which used the SVM grid search technique along with 10k cross validation. Grid search method continually changes parametric values of SVM: gamma and cost, until the highest polarity detection accuracy

is achieved for the given dataset. This characteristic helps to tune the performance of SVM in classification. Cross validation on the other hand keeps changing the ratio of training and test data until finds the particular proportion which further enhance the accuracy to the maximum point. Finally performance of the explored technique was evaluated by comparing the f-measure scores with the published research, which have used the SVM with same datasets. It has been observed that grid search technique performed well and it is further suggested to use this technique on further datasets to explore its accuracy.

REFERENCES

[1] M. Ahmad, S. Aftab, S. S. Muhammad, and U. Waheed, "Tools and Techniques for Lexicon Driven Sentiment Analysis : A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 1, pp. 17–23, 2017.

[2] M. Ahmad, S. Aftab, and S. S. Muhammad, "Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.

[3] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, 2017.

[4] M. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," vol. 6, no. 6, pp. 1067–1073, 2016.

[5] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[6] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," Int. J. Comput. Appl., vol. 54, no. 13, pp. 21–25, 2012.

[7] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," Fourth Work. Comput. …, no. November, pp. 47–56, 2012.

[8] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2013.

[9] A. Mueen, "Modeling and Predicting Students ' Academic Performance Using Data Mining Techniques," no. November, pp. 36–42, 2016.

[10] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1148–1153, 2016.

[11] N. Mallios, E. Papageorgiou, M. Samarinas, and K. Skriapas, "Comparison of machine learning techniques using the WEKA environment for prostate cancer therapy plan," in Proceedings of the 2011 20th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2011, 2011, pp. 151–155.

[12] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," Int. J. Comput. Appl., vol. 177, no. 5, pp. 25–29, 2017.

[13] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment Analysis using SVM : A Systematic Literature Review," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 182–188, 2018.

[14] N. J. Sanders, "Sanders-twitter sentiment corpus," Sanders Anal. LLC., 2011.

[15] E. Ikonomovska, "Airline dataset." .

[16] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," 2004.

[17] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench," in Morgan Kaufmann, Fourth Edition, 2016, pp. 553–571.

[18] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008.

[19] T. Tokunaga and I. Makoto, "Text categorization based on weighted inverse document frequency," Spec. Interes. Groups Inf. Process Soc. Japan (SIG-IPSJ, 1994.

[20] C. C. Aggarwal and C. X. Zhai, Mining text data. 2013.

[21] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 04, vol. 4, p. 168, 2004.

[22] N. Kobayashi, K. Inui, and Y. Matsumoto, "Opinion Mining from Web Documents: Extraction and Structurization," Trans. Japanese Soc. Artif. Intell., vol. 22, pp. 227–238, 2007.

[23] H. Jeong, D. Shin, and J. Choi, "FEROM: Feature extraction and refinement for opinion mining," ETRI J., vol. 33, no. 5, pp. 720–730, 2011.

[24] G. Mishne, "Experiments with mood classification in blog posts," Proc. ACM SIGIR 2005 Work. Stylist. Anal. Text Inf. Access, p. 19, 2005.

[25] S. Stymne, "Pre- and Postprocessing for Statistical Machine Translation into Germanic Languages," Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Student Sess., no. June, pp. 12–17, 2011.

[26] J. B. Lovins, "Development of a stemming algorithm," Mech. Transl. Comput. Linguist., vol. 11, no. June, pp. 22–31, 1968.

[27] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM J. Res. Dev., vol. 2, no. 2, pp. 159–165, 1958.

[28] D. K. Ly, K. Sugiyama, Z. Lin, and M.-Y. Kan, "+Product Review Summarization from a Deeper Perspective," Proc. 11th ACM/IEEE-CS Jt. Conf. Digit. Libr., no. July, pp. 311–314, 2011.

[29] N. Archak, A. Ghose, and P. G. Ipeirotis, "Show me the Money ! Deriving the Pricing Power of Product," Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '07, pp. 56–65, 2007.

[30] D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino, "Hyperparameter design criteria for support vector classifiers," Neurocomputing, vol. 55, no. 1–2, pp. 109–134, 2003.

[31] K. Duan, S. Keerthi, and A. Poo, "Evaluation of simple performance measures for tuning SVM hyper parameters. Technical report," Neurocomputing, vol. 51, pp. 41–59, 2001.

[32] J. Wainer and G. C. Cawley, "Empirical Evaluation of Resampling Procedures for Optimising SVM Hyperparameters," J. Mach. Learn. Res., vol. 18, no. 15, pp. 475–509, 2017.

[33] E. Duarte and J. Wainer, "Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters," Pattern Recognit. Lett., vol. 88, pp. 6–11, 2017.

[34] D. M. Hawkins, "The Problem of Overfitting," J. Chem. Inf. Comput. Sci., vol. 44, no. 1, pp. 1–12, 2004.

[35] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," Encycl. Database Syst., pp. 532–538, 2009.

[36] C. W. Hsu, C. C. Chang and C. J. Lin, "A Practical Guide to Support Vector Classification," BJU Int., vol. 101, no. 1, pp. 1396–400, 2008.