Evolutionary Intelligence (2021) 14:725–731
https://doi.org/10.1007/s12065-020-00429-1

SPECIAL ISSUE

An efficient approach for sentiment analysis using machine learning algorithm

A. Naresh[1] · **P. Venkata Krishna**[2]

Received: 21 November 2019 / Revised: 6 May 2020 / Accepted: 20 May 2020 / Published online: 3 June 2020

Abstract
Sentimental analysis determines the views of the user from the social media. It is used to classify the content of the text into neutral, negative and positive classes. Various researchers have used different methods to train and classify twitter dataset with different results. Particularly when time is taken as constraint in some applications like airline and sales, the algorithm plays a major role. In this paper an optimization based machine learning algorithm is proposed to classify the twitter data. The process was done in three stages. In the first stage data is collected and preprocessed, in the second stage the data is optimized by extracting necessary features and in the third stage the updated training set is classified into different classes by applying different machine learning algorithms. Each algorithm gives different results. It is observed that the proposed method i.e., sequential minimal optimization with decision tree gives good accuracy of 89.47% compared to other machine learning algorithms.

Keywords Semantic analysis · Machine learning algorithms · Preprocessing · Accuracy · Optimization

1 Introduction

A series of tools, techniques and methods used to detect and extract information is defined as sentimental analysis. The information extracted includes attitudes and opinion of the users. The main objective is to know whether the user has negative, positive and neutral opinion towards a product or something else. The numbers of papers on sentimental analysis have been increased drastically nowadays. It is one of the fastest growing research areas. It is observed that the public opinions, their motivation are political in nature. Later in mid-2000 the companies have mainly focused on reviews of the product which are available in web. This is very useful in the prediction of financial markets [1], reaction of public to terrorist attacks [2].

But the natural language processing has been addressing many problems to applicability of sentimental analysis [3]. The efforts have been advanced from simple polarity detection to complex emoticons, differentiating negative emotions [4]. Decision making is one of the important factors "what other people are thinking". For example while planning to vote in elections. By using internet it is possible to search the individual's opinion from personal network. Furthermore, many individuals are making their opinions through internet such as from comments, blogs in social networking sites.

Large portion of our society is connected through social networks directly or indirectly. Anyone can express their opinions without any fear of undesirable consequences. To derive the opinion of users from lengthy comments, informal language such as emoticons, abbreviation will complicate the analysis. In this paper, machine learning algorithms were applied with optimization technique's to extract features and opinions. The paper is organized as follows. Section 2 describes the related work. Section 3 describes proposed algorithm. Section 4 describes algorithm, Sect. 5 the results and evaluation and finally Sect. 6 conclusion.

✉ A. Naresh
pandu5188@gmail.com

P. Venkata Krishna
pvk@spmvv.ac.in

[1] Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

[2] Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, India

1.1 sequential minimal optimization (SMO)

This is mainly used in training a support vector classifier which uses polynomial or RBF kernels. All the missing values will be replaced and also transforms nominal to binary attributes. It is more efficient algorithm to solve SVM problem when compared to generic quadratic program algorithms. It is considered as a method of decomposition where a multivariable optimization problem is decomposed into a series of sub problems. Each sub problem is optimized individually with small number of variables. Only one variable will be optimized and remaining variables are treated as constants. Consider a_i is a variable then a1 will be optimized and a_2, a_3, a_4 … a_n are treated as constants. The condition is

$$Y_i a_i = 0 \tag{1}$$

Throughout the iterations which means whenever one multiple is updated at least another multiple should be adjusted to satisfy the condition. In each step SMO selects two elements ai and a_j to jointly optimize then finds optimal values for those two parameters and remaining all others are fixed.

Finally updates 'a' vector accordingly. The choice of two points is made by a heuristic approach where optimization is done analytically.

The advantage is it doesn't use kernel matrix in the memory to store. Since there are no matrix operations it doesn't require other packages hence easy to implement and increase space complexity.

1.2 Decision tree

The classification is built in the form of a tree structure. It breaks down the large dataset into smaller subsets and a relevant decision tree is developed incrementally. The final tree consists of decision nodes and leaf nodes. A decision node has two or more branches and leaf node is represented as a decision or classification. The topmost node in a tree is called as a root node. It can handle both numerical and categorical data. it uses entropy to calculate homogeneity of a sample. The entropy is zero if the sample is completely homogeneous and the entropy is one if the sample is equally divided.

$$\text{Entropy(s)} = \sum_{i=0}^{n} -xi\, log2\, xi \tag{2}$$

2 Related work

In [5] knowledge based approach and machine learning techniques were used for classification. It ensemble the applications common sense computing. This is also referred as concept level analysis. In [6] a machine learning techniques have been used for polarity detection and to improve its accuracy. It combines lexical-based and machine learning techniques. It classifies the Facebook messages based on its polarity.

In [7] a lexicon based approach has been used for classification of twitter data which works on static polarity. In [8] a natural language processing method is used to extract the features from the reviews. A score value is given to all the features which help to analyze the reviews with high accuracy. In [1] a CNN with multiple filters have been used to extract features with varying the window size. It is an unsupervised learning method.

In [9] the features have been extracted from the text and then classified by using supervised learning classifiers such as support vector machine. In [10] a knowledge based approach is used to classify the sentiments from the social media.

An iterative algorithm was proposed in [11] to predict sentiment polarities in twitter data sets. The algorithm was performed in two stages. In the first stage, sentiment reversal was done to analyze the tweets and retweets. The tweets were constructed into a tree by splitting into tweet (parent node) and retweet (child node). Both will have different polarities i, positive, negative and neutral. In the second stage, the relationship between the diffusion and reversals was done to extract the patterns. Overall performance has been improved by 8.53% compared to other methods.

Multivariate vehicle regression models were applied on the values of stock market and social media to predict monthly sales of the vehicles in [12]. Three kinds of datasets were taken as inputs to analyze and predict the sales. They are values of stock market, scores of sentiments and hybrid model. By observation, it is noted that by applying regression models on three datasets, hybrid model is giving more accurate results compared to other datasets.

In [13] seven machine learning techniques were applied to classify the airline dataset into three sentiment classes i.e., positive, negative and neutral. The input data is preprocessed and the tweets were represented as vectors to perform deep learning concepts. The dataset was divided into 80% of train data and 20% as test data. By observation it is noted that AdaBoost is giving good accuracy of 84.5%. The main drawback is the limited numbers of tweets were considered. It would be good to get high accuracy if the number of tweets were increased.

3 Proposed methodology: sequential minimal based optimization decision tree (SMODT)

The features are extracted by an optimization algorithm and then classified the sentiments using machine learning algorithms for better accuracy. The following are the steps

followed. The step by step procedure is represented in the Fig. 1.

1. Data collection.
2. Data preprocessing.
3. Feature extraction by optimization.
4. Classification.

3.1 Data collection

The input data is collected from the twitter data source using API. Application program interface is used to collect the input data. It is an user interface between user and source. The source is a twitter website which consists of tweets of the users. The airline dataset is taken in the paper. It consists of many attributes like tweet_id, tweet sentiment, comments, tweet_created, tweet_count, negative reason, location, time zone etc., The Airline twitter dataset was taken as input.

3.2 Data preprocessing

After collecting the data, preprocessing is done by removing noisy, irrelevant data from the dataset. The collected tweets consist of many emoticons, hash tags, special symbols along with the comment (opinion). Those symbols are considered as noisy and irrelevant hence removed from the dataset. Preprocessing of data is necessary to eliminate noisy, inconsistent, incomplete data. The data before preprocessing and after preprocessing is shown in the Tables 1 and 2.

3.2.1 Removing of URL

To classify and to analyze the sentiment of the tweets, URLs will not be considered usually. For example 'your chat is not working on your site: http://ecstacy.com as she is feeling excited". In this URL by considering the word 'not' it can be referred to a negative sentiment but the word excited refers to a positive sentiment. Hence it will be considered as neutral. In order to remove such kind of sentiments the URLs are removed.

3.2.2 Removal of special characters

During the assignment of polarity the special characters like [], {}, () are removed to resolve logical and incompatibilities.

3.2.3 Removing re tweets

Copying of another user tweets and posting it into another account is considered as re tweeting. It is abbreviated as RT, in order to avoid redundancy, re tweets are removed.

3.2.4 Removal of hash symbols

These labels act as keywords and helpful in search of particular messages and posts. For example #guilty pleasures will produce tweets list which are related to #guilty pleasures. These are not needed in polarity detection. Hence considered as irrelevant and removed.

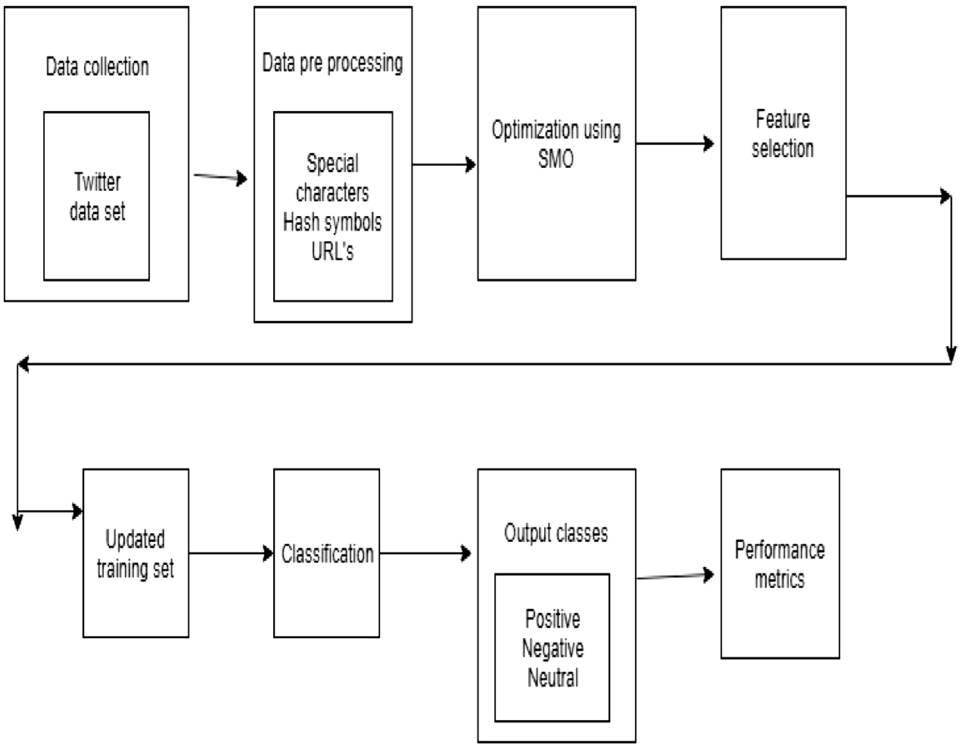

Fig. 1 Classification process

Table 1 Before preprocessing

Tweet_id	sentiment	Sentiment-confidence	Tweet
570306133677760513	Neutral	1.0	@VirginAmerica What @dhepburn said
570301130888122368	Positive	0.3486	@VirginAmerica plus you've added commercials to the experience… tacky
570301083672813571	Neutral	0.6837	@VirginAmerica I didn't today… Must mean I need to take another trip!
570301031407624196	Negative	1.0	it's really aggressive to blast obnoxious ""entertainment"" in your guests' faces & they have little recourse
570301031407624196	Negative	1.0	@VirginAmerica it's really aggressive to blast obnoxious ""entertainment"" in your guests' faces &
570300817074462722	Negative	1.0	it's a really big bad thing about it
570300616901320704	Positive	0.6559	Well, I didn'tâ€¦but NOW I DO!
570289584061480960	Positive	1.0	This is such a great deal! Already thinking about my 2nd trip
570289584061480960	Positive	1.0	This is such a great deal! Already thinking about my 2nd trip
570300248553349120	Neutral	0.634	Really missed a prime opportunity for Men Without Hats parody, there

Table 2 After preprocessing

Tweet_id	Sentiment	Tweet
1	Positive	It was amazing, arrived early, too good
2	Positive	It is a great deal
3	Positive	Must mean to take another trip
4	Negative	Really bit bad thing about it
5	Positive	Thinking about second trip
6	Negative	Could not fully sit in my seat
7	Negative	Would u miss me will be there together
8	Positive	This is a great news
9	Neutral	Will flights be leaving for Dallas tonight
10	Neutral	New marketing song. Let's know what you think
11	Positive	It was amazing, arrived early, too good

3.3 Feature generation

In this technique the preprocessed data is analyzed. The classification is done based on an optimization method called SMO (sequential minimal optimization) [14] and decision tree [15].

The input data is divided into smallest sub quadratic programming problems and they are solved analytically. It is mainly applicable for large datasets. The large datasets consumes more time to process. By applying SMO the total time taken can be optimized as the memory required is linear in SMO. In SMO matrix computation is avoided. Thus it is faster even for larger datasets.

The testing time is less when SMO is combined with Decision Tree.

The total data i.e., all the tweets are divided evenly according to the decision tree. The left sub tree belongs to one class and right sub tree belongs to another class. All the samples were trained at the same time. The SMO was applied during training time. Thus process is repeated until the leaf node is reached. IN the first step the SMO has to deal with entire data set. In the second step the SMO has to deal with only subset of data obtained from the first step. By parallel processing the data the time taken to train and test the data can be reduced dramatically and simplify the prediction process.

Consider the total data set as 'a' and it is divided into two datasets ai and aj. Then the data can be optimized as follows,

$$\text{MaxW(a)} = \sum_{i=0}^{n} ai - 1/2 \sum_{i=0}^{n} \sum_{j=0}^{n} aiajxixjk(yi, yj) \tag{3}$$

$$\text{Subject to } \sum_{i=0}^{n} aixi = 0$$

Only a_1 and a_2 are allowed to change.

$$a_2^{new} = a_2^{old} + x_2 \left\{ E_2^{(old)} - E_1^{(old)} \right\} / k \tag{4}$$

where $E_i = f(y_i) - x_i$

$$E_i = \left(\sum_{j=0}^{n} ajxjkji - c \right) - xi \tag{5}$$

4 Algorithm

Input: Set of tweets

Step1: Initialize the training set

Step2: consider two variables heuristically

Step3: optimize the two values

MaxW(a) = $\sum_{i=0}^{n} ai - 1/2 \sum_{i=0}^{n} \sum_{j=0}^{n} aiajxixjk(yi, yj)$

Step4: Find out the new values

$a_2^{new} = a_2^{old} + x_2\left\{E_2^{(old)} - E_1^{(old)}\right\}/k$

$E_i = f(y_i) - x_i$

$E_i = \left(\sum_{j=0}^{n} ajxjkji - c\right) - xi$

Step5: Update new values.

Step6: Repeat steps 3, 4, and 5, to all values until the convergence is reached.

Step7: Update new training set.

Step8: Find the entropy value of each variable.

Step9: This process is repeated until the tree is constructed.

Step10: Calculate accuracy, prediction, recall, f-measure for the obtained classes.

Output: Accuracy, precision, recall, f-measure.

4.1 Algorithm

Result: Classification of tweets

Input: Tweets of airline dataset

While $i = 0$ *to* n
do
Initialize the training set

Consider two variables heuristically

Optimize the two values

$$\text{Max W (a)} = \sum_{i=0}^{n} ai - 1/2 \sum_{i=0}^{n} \sum_{j=0}^{n} aiajxixjk(yi, yj)$$

Find out the new values

$a_2^{new} = a_2^{old} + x_2\{E_2^{(old)} - E_1^{(old)}\}/k$

$E_i = f(y_i)\text{-}x_i$

$$E_i = \left(\sum_{j=0}^{n} ajxjkji - c\right) - xi$$

end
Calculate accuracy, precision, recall, f-measure,

Compare the obtained accuracy with different algorithms

Classify the tweets into different classes

Output the classes and performance metrics
end
Algorithm 1: Prediction of sentiments

Table 3 Performance metrics of different algorithms

Classifiers	Accuracy(%)	Precision	Recall	F-Measure
KNN	67	70.5	69.3	67.9
SVM	68	69	68.1	68.7
KNN + SVM	76	68.45	68.14	77.56
SMO	86.23	89.2	86.2	85.8
DT	80	81.4	81.4	80.95
SMO + DT	89.47	91.6	89.5	96.3

5 Performance evaluation

The collected tweets are preprocessed and classified into different classes. The best features are extracted from the preprocessed data using SMO algorithm and then machine learning algorithm is ensemble to classify tweets into different classes. The tweets are mainly classified into three classes. The collected data consists of three sentiments positive, negative and neutral. The tweets are classified accordingly. Table 4 shows the total number of tweets classified into different classes. 466 tweets are classified into positive class out of 1000 tweets, 277 tweets are classified into negative class and 255 tweets are classified into neutral class. Table 3 represents the performance metrics of different algorithms. KNN algorithm has 67% accuracy, SVM has 68%, SVM + KNN has 76%, SMO has 86.23% and SMO + DT has highest accuracy of 89.47% compared to other algorithms.

The performance metrics can be described as following.

Accuracy: Accuracy can be defined as the ratio of true positives i.e., correctly classified samples to the total number of samples. The accuracy can be represented as,

$$\text{Accuracy} = \frac{\text{Correctly Classified samples}}{\text{Total number of samples}} \tag{6}$$

Precision: It can be defined as the ration of correctly classified samples to the total number of positive samples.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{7}$$

Recall: It is the ratio of correctly predicted positive samples.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{8}$$

F-measure: It is the weighted average of recall and precision. F-measure is more useful compared to accuracy.

$$\text{F - measure} = 2 * \frac{(\text{Recall*Precision})}{(\text{Recall} + \text{Precision})} \tag{9}$$

The total number of tweets considered in the work is 1000. These are classified into three classes positive, negative and neutral. By applying SMODT algorithm 466 are classified into positive class, 277 into negative class and remaining into neutral class. The results are represented in the Fig. 2 and in Table 4. The algorithm is evaluated using python. In the proposed method sequential minimal optimization and decision tree algorithms were used. It is a hybrid approach. The hybrid approach has many advantages like optimizing total training time. It is scalable i.e., more suitable for large datasets. Since twitter datasets are large in size, by applying SMODT the classification can be simplified with good accuracy.

Machine learning algorithms have been tested to check the performance. K-Nearest Neighbor algorithm with Support Vector Machine was applied on the airline dataset along with the proposed method. But the performance was less when compared to proposed method. Hence, it can be concluded that the proposed method SMODT is giving good results.

The proposed method can be applicable in many fields.

- In market to know the customer feedback and to improve productivity. It may be seasonal or unseasonal, monthly or yearly.
- In travel's like airline to know whether the customers are satisfied with the services provided or not.

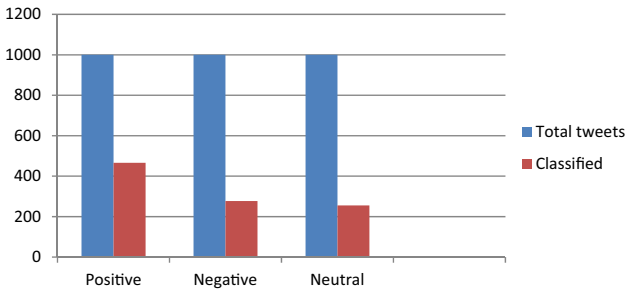

Fig. 2 Classification of tweets

Table 4 Classification of tweets

Sentiment class	Total tweets	Classified
Positive	1000	466
Negative	1000	277
Neutral	1000	255

6 Conclusion

In this paper, an optimization based machine learning algorithm is proposed to classify the tweets into different classes. The proposed model is evaluated in three stages. First stage, the data is preprocessed to remove noisy data. In the second stage the features are extracted by applying an optimization technique and in the third stage the updated training set is classified into three different classes namely positive, negative and neutral. The proposed algorithm has got maximum accuracy of 89.47% compared to other machine learning algorithms. The algorithm is faster and reduces the overall time taken to process the data. This is most suitable for larger datasets. In the future work an efficient optimization technique can be applied to improve the performance.

References

1. Gokulnath C, Priyan MK, Balan EV, Prabha KR, Jeyanthi R (2015) Preservation of privacy in data mining by using PCA based perturbation technique. In: 2015 International conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). IEEE, pp 202–206
2. Kumar PM, Gandhi U, Varatharajan R, Manogaran G, Jidhesh R, Vadivel T (2019) Intelligent face recognition and navigation system using neural learning for smart security in internet of things. Clust Comput 22(4):7733–7744
3. Manogaran G, Varatharajan R, Lopez D, Kumar PM, Sundarasekar R, Thota C (2018) A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Gener Comput Syst 82:375–387
4. Da Silva NF, Hruschka ER, Hruschka ER Jr (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66:170–179
5. Varatharajan R, Manogaran G, Priyan MK, Sundarasekar R (2018) Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. Clust Comput 21:681–690
6. Varatharajan R, Manogaran G, Priyan MK, Balas VE, Barna C (2018) Visual analysis of geospatial habitat suitability model based on inverse distance weighting with paired comparison analysis. Multimed Tools Appl 77:17573–17593
7. Balan EV, Priyan MK, Gokulnath C, Devi (2015) Fuzzy based intrusion detection systems in MANET. Proc Comput Sci 50:109–114
8. Manogaran G, Varatharajan R, Priyan MK (2018) Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimed Tools Appl 77(4):4379–4399
9. Devi GU, Balan EV, Priyan MK, Gokulnath C (2015) Mutual authentication scheme for IoT application. Indian J Sci Technol 8(26):15
10. Somani A, Suman U (2011) Counter measures against evolving search engine spamming techniques. In: 2011 3rd international conference on electronics computer technology (ICECT). vol 6, pp 214–217
11. Wang L, Niu J, Yu S (2019) SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2019.2913641

12. Pai P-F, Liu C-H (2018) Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. IEEE Access 6:57655–57662
13. Rane A, Kumar A (2018) Sentiment classification system of twitter data for US airline service analysis. In: 2018 IEEE 42nd annual computer software and applications conference (COMPSAC). IEEE, vol 1, pp 769–773
14. Pham BT, Bui DT, Prakash I, Nguyen LH, Dholakia MB (2017) A comparative study of sequential minimal optimization-based support vector machines, vote feature intervals and logistic regression in landslide susceptibility assessment using GIS. Environ Earth Sci 76:371
15. Harrison PA, Dunford R, Barton DN, Kelemen E, Mart B (2018) Selecting methods for ecosystem service assessment: a decision tree approach. Ecosyst Serv 29:481–498