American Journal of Applied Sciences

Original Research Paper

Sentiment Analysis: Comparative Study between GSVM and KNN

[1]Hany Mohamed, [2]Ayman Atia and [3]Mostafa-Sami M. Mostafa

[1]Faculty of Computer Science, Helwan University, Egypt
[2]Faculty of Computer Science, Misr International University, Helwan University, HCI-LAB, Egypt
[3]Faculty of Computer Science, Helwan University, HCI-LAB, Egypt

Article history
Received: 5-04-2018
Revised: 13-06-2018
Accepted: 1-08-2018

Corresponding Author:
Hany Mohamed
Faculty of Computer Science,
Helwan University, Egypt
Email: hany.abdelmawgood@its.ws

Abstract: Sentiment classification aims detecting general opinion of users in social media towards business products or daily life events. The classification tells whether sentiment is positive or negative. Techniques of sentiment classification are categorized into lexical analysis and machine learning techniques. In this paper, we propose a comparative study between SVM applied genetics (GSVM) against KNN algorithm in terms of speed and accuracy. We present also an experimental study of sentiment classification on different domains movie reviews, financial and amazon toys products. The experimental results shows that GSVM achieves a classification accuracy of 92% and KNN achieves 87% on movie reviews dataset. For classification speed, KNN shows a remarkable improvement (above 10% improvement) in comparison with GSVM.

Keyword: Sentiment Classification, SVM, KNN, NLP, GENETICS

Introduction

People are get used to express their sentiment in social media towards daily events in different life areas whether sports, political and so on. Sentiment analysis is defined as the process of detecting sentiment or opinion of user's statement towards daily activities (Mohamed *et al*., 2017).

Sentiment takes either positive, negative, or neutral. Positive sentiment is when people express good feeling towards movie, while negative sentiment is vise versa.

Table 1 shows a sample of positive and negative posts from Cornell movie reviews dataset (Heather Whiting *et al*., 2017).

Different approaches in sentiment classification area are presented in (Mohamed and Ezzat, 2015; Shivhare and Khethawat, 2012; Kalaivani and Shunmuganathan, 2014; Guo *et al*., 2003), which is categorized based on lexical analysis or Machine Learning (ML) techniques.

The main idea of lexical analysis is detecting effective words in posts based on common lexicon like Wordnet or Wordnet-Affect. While machine learning role is to find the class label of input text based on training data and predictive model (Shivhare and Khethawat, 2012). This problem is called text classification which is different than classification in other domains due to large number of features. SVM, NAIVE Bayes, KNN are popular techniques in sentiment classification, while they are depending on bag of words model to generate unique words from input text. a lot of generated features of this technique are irrelevant, redundant or noisy and filtration mechanism has to be applied. SVM Applied Genetics (GSVM) (Mohamed *et al*., 2017) is an enhanced technique that aimed filtering selected features to achieve better classification accuracy.

The main contribution of this paper is to present a comparative results between GSVM and K-nearest neighbor algorithm in terms of speed and accuracy. The implemented experiments are based on three different dataset which are movie review dataset (Zhang *et al*., 2011), financial dataset (Mohamed *et al*., 2017) and amazon Toys review dataset (He and McAuley, 2016).

The rest of this paper is organized as follows, Section II discuss popular approaches in emotion detection area from lexical to machine learning approaches. Section III presents methodology implemented in comparative study. Section IV shows the evaluation results. Finally, conclusion and future work listed in section V.

Background and Related Work

In sentiment analysis, there exists specific research challenges. Text Informality, Language Acronyms, Languages Mixture, Emotion icons and Relevance (Mohamed *et al*., 2017) are samples. Early works in sentiment analysis are depending on lexical resources. Preotiuc-Pietro *et al*. (2012), SentiWordNet lexicon was applied by counting positive and negative terms found in a review and the sentiment polarity was determined based on which class received the highest score.

Table 1: Sample of positive and negative tweets from movie review dataset

Tweet	Sentiment
” jaws” is a rare film that grabs your attention before it shows you a single image on screen.	Positive
One cannot observe a star trek movie and expect to see serious, science fiction.	Positive
The purpose of star trek is to provide flashy, innocent fun” snake eyes”	
is the most aggravating kind of movie: the, kind that show so	Negative
much potential then becomes unbelievably disappointing	
whether you like the beatles or not, nobody wants to see the bee,	Negative
gee’s take on some of the fab four’s best known songs	

Neviarouskaya *et al*. (2009), Construction of domain-oriented sentiment lexicon as clustering of sentiment words and extends the information-bottleneck clustering algorithm by integration more restriction for building an appropriate knowledge context of every sentiment word. Opinion-Finder, WordNet-Affect, MPQA and SenticNet are popular lexical resources that highly used in sentiment analysis rather than SentiWordNet. Point-wise Mutual Information (PMI) (Kamble1 and Deshmukh, 2016) is a criterion commonly useful for statistical language model of word associations and its related applications. This method calculates mutual information between two words to obtain numeric score as in Equation 1:

$$PM1(word1, word2) = \log_2\left(\frac{prob(word1\ \delta\ word2)}{prob(word1) * prob(word2)}\right) \quad (1)$$

Here, each word is defined based on percentage of its relation to positive PMI (word1, positive word) or negative emotion PMI (word1, negative word). Finally, Semantic Orientation (SO) is calculated using Equation 2:

$$PM1(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} = \log_2 \frac{P(y \mid x)}{P(y)} \quad (2)$$

Sentiment classification of movie reviews is proposed in (Neviarouskaya *et al*., 2009) by applying three machine learning techniques of SVM, NAIVE Bayes and character based N-gram model for sentiment classification of the reviews. The evaluation results tells that accuracy of all approaches is more than 80% and also that SVM and N-gram approaches outperformed NAIVE Bayes technique.

K-Nearest Neighbor (KNN) algorithm in combination with TF-IDF for classifying sentiment is utilized in (Guo *et al*., 2003). A key advantage of KNN is it simplicity and execution speed. KNN is based on finding the most similar objects (documents) from sample based on mutual Euclidean distance (He and McAuley, 2016). Based on given results, it is proved that KNN applied TF-IDF method has been a good choice taking into consideration that amount of unusable words in documents has a significant impact on the final quality of classification.

A new way of sentiment analysis is proposed in (Zhang *et al*., 2011), it combines Lexicon-based and Learning-based Methods. The method first adopts a lexicon based approach to perform entity-level sentiment analysis that gives high precision but low recall. Then a classifier is trained to assign polarities to the entities newly identified tweets, proved that this way gives better F-Score. Corpus collected from Twitter with annotated microblog posts (or ”tweets”) annotated at the tweet-level with seven emotions: anger, disgust, fear, joy, love, sadness and surprise. This research illustrate framework of EmpaTweet system for annotating and detecting emotion from twitter. The system uses a series of binary SVM classifiers to detect each of the seven emotions annotated in the corpus. Each classifier performs independently on a single emotion.

Another novel method introduced in (Li *et al*., 2016), by applying a pre-training method to deep neural networks based on restricted Boltzmann machines, which aims to gain competitive and stable classification performance of user emotions over short text. The result indicates that this method performed competitively competitively in terms of accuracy and robustness.

An improved NAIVE Bayes algorithm is presented in (Kang *et al*., 2011) for sentiment analysis of restaurant reviews based on unigram and bigram features. The experiments showed an accuracy that improved by a maximum of 10.2% in recall and a maximum of 26.2%.

A new method of SVM applying genetics is presented in (Mohamed *et al*., 2017), which is based on important feature selection method which is information gain by removing the irrelevant or redundant features. Information gain outperformed than other feature selection method which is calculated based on entropy (Preotiuc-Pietro *et al*., 2012; Neviarouskaya *et al*., 2009; Kamble1 and Deshmukh, 2016). Entropy is a common way in information retrieval area to measure impurity, while impurity refers to class distribution within dataset, High impurity leads to high classification accuracy. Entropy is calculated as in Equation 3:

$$Entropy = \sum -P_i \; Log \; P_i \quad (3)$$

where, P_i is the probability of class *i*, the higher Entropy leads to better accuracy and high information content. Information Gain (IG) then is calculated to check which features are considered the most important in our classification problem. IG is calculated as in Equation 4:

$$IG = entropy(parent) - [average\ entrop(children)] \quad (4)$$

The experiments in (Mohamed *et al*., 2017) shows an improvement of classification accuracy (89.9%) rather than support vector machine technique (88.6%).

Methodology

Preprocessing Steps

Sentiment analysis process start with tweet tokening to split the text into a sequence of words. To assure accuracy, all characters are converted to lowercase. Stemming phase is then applied for removing morphological affixes from words that generated from previous step. Lancaster stemming is the used algorithm in our research process. Removing Sarcastic words is then applied to remove tokens like stop words from wordlist. We use python NLTK (Natural Language toolkit) for text processing.

KNN

KNN algorithm is based on finding the most similar objects (tweets) from sample groups using mutual Euclidean distance (Zhang *et al*., 2011).

Algorithm 1: KNN implementation – Euclidian distance
Procedure KNNAlgorithm (K)

> **Initialize**
- T <- number of tweets
- N <- Number of unique words

> **Steps**
- For each tweet in training dataset (i in T)
  - For each word in the list of unique words (j in N)
    - Compute term frequency of term i in tweet j.
    - Computer tweet frequency of term i in the dataset.
    - Compute weight of term i in tweet j (Aij)
  - Compute distance between two tweets
    - D[i] ← D[i] + A[j,i] – A[j,T – 1)] 2
    - D[i] ← Sqrt (D[i])

> **Return**
- Return k tweet that has least distance d[i]

End procedure

Algorithm 2: GSVM – Feature selection
Procedure GSVMAlgorithm (V, N)

> **Initialize**
- P <- InitializePopulation (V); Initialized populated has feature vector of selected tweets after applying text processing techniques.
- r <- 0
- N <- number of rounds

> **Steps**
-
- F(r) = ComputeFitness (P); This function is responsible for calculating F1 score of current population
- While r < N do
  - P <- Mutate (P); reconstruct population by replacing one or more features by other ones that have high information gain.
  - F(r) = ComputeFitness (P)
- End while

> **Return**
- Return best P

End procedure

The process start with preparation of weight matrix which evaluates importance of words in given dataset based on Term frequency-inverse document frequency (TF-IDF). Assuming matrix N*M, where N is defined by unique words that is generated by preprocessing stage while M represents number of collected post. Thus, matrix constructed as relational matrix between each word and each tweet. Equation 5 is used to calculate the weight value of word i tweet *j*.

$$aij = tf_{ij}\, idf_i = \frac{fi_j}{\sum_{S=1}^{N}(tf_i\ df)(a_{sj}))} * \log_2(\frac{N}{df_i}) \quad (5)$$

Where:
a_{ij} = The weight of term *i* in tweet *j*
N = The number of tweets in dataset
$t\,f_{ij}$ = The term frequency of term *i* in tweet *j*
df_i = The tweet frequency of term *i* in the dataset

while equation 6 determine vector distance between any two tweets:

$$d(x,y) = \sqrt{\sum_{r=1}^{N}(a_{rx} - a_{ry})^2} \quad (6)$$

Where:
$d(x,y)$ = The distance between any two tweet
N = The number of unique words in given dataset
arx = Weight of term *r* in tweet *x*
ary = Weight of term *r* in tweet *y*

Algorithm 1 shows an implementation of KNN.

GSVM

SVM classification algorithm uses a set of training instances and predicts new instances with two possible class label-1, 1 (Zaguai and Beizak, 2015). The process start with applying Bag-of-Word models to model frequencies or number of occurrence for each word in tweet. The main problem here that Bag of word generates hundred or thousands features in input space which is not efficient way of vectorizing features, a lot of generated features of this technique are irrelevant, redundant or noisy (Buck *et al*., 2014). Based on genetic algorithm, feature selection method is applied on these features that generated from text to select best chromosome of features with high information gain. Algorithm 2 shows implementation of GSVM, it starts with initialized chromosome of generated features. The objective function is to maximize F1 score of the best generated chromosome.

Evaluation and Results

Experiment Setup

In this research, we use Cornell movie review dataset (Li *et al*., 2016) collected from twitter; it contains 1000 positive reviews and 1000 negative reviews. The rating classifier determine whether a review was positive or negative by obtaining accurate rate specified by user, which takes either numerical rating (range from 1 to 10). Sample of published posts taken is shown Table 1.

Another case study we show in this research is amazon toys reviews (He and McAuley, 2016). This dataset contains product reviews and metadata from Amazon, including about 167,597 reviews span till July 2014. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand and image features) and links (also viewed/also bought graphs). The dataset is represented in JSON format, Fig. 1 shows a sample of these reviews.

Financial dataset is the third case study in this research, Sample of published posts taken at 03-Mar-2017 for EURUSD currency pair shown in Fig. 2, data represented in CSV format. While investors express their opinion/expectation for EURUSD after critical event hold in USA for FED in subject related to interest rate. Each record contains the following: Author Name-No. Of Followers-Tags-Time-Post-Sentiment as shown in Fig. 2.

The key software and hardware specifications of our server that we think may affect the performance are shown in Table 2.

Table 2: The key software and hardware specifications

Hardware	Details
CPU	Intel - Core i7 - 2.9 GHz
Mem	8.00 GB
Software	
OS	Windows 7
GSVM	Python 3.5 32 bit – Skitlearn
KNN	Python 3.5 32 bit – Skitlearn


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
piano.  He is having a wonderful time playing these old
hymns.  The music  is at times hard to read because we think
the book was published for singing from more than playing
from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Fig. 1: Sample of Amazon review of toys product

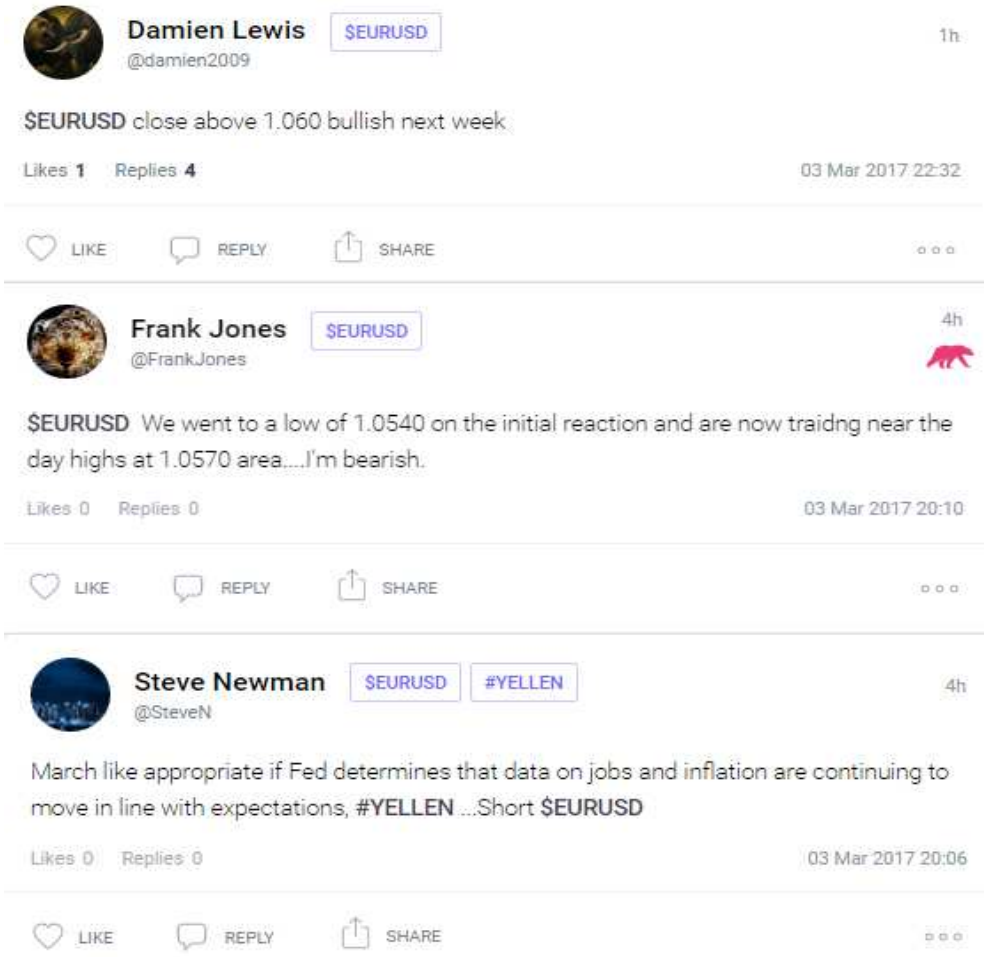

Fig. 2: Sample of financial posts

Evaluation Criteria

We consider accuracy, precision, recall, F-Measure and classification speed as evaluation criteria.

Accuracy: It is a powerful factor to evaluate ML technique, it is calculated based on ration of correct predicated sentiment related to all movie reviews.

Equation calculated as per equation 7:

$$Accuracy = TN / (TP + FP + TN) \quad (7)$$

Precision: it is calculated based on ratio of right predicated positive sentiment in related to total of positive sentiment movie reviews as per Equation 8:

$$\text{Pr}ecision = Tp / (TP + FP) \quad (8)$$

Recall: it is calculated based on the right predicated positive sentiment in related to all sentiment in actual sentiment class as per Equation 9:

$$\text{Re}call = Tp / (TP + FN) \quad (9)$$

F1 score: it is calculated based on weighted average of precision and recall as Equation 10:

$$F1 = 2 * (\text{Re}call * \text{Pr}ecisiion) / (\text{Re}call + \text{Pr}ecision) \quad (10)$$

Classification speed: it indicates the time required for sentiment classification on given dataset, the classification speed is measured by CPU time.

Results

In this experiment, we select top k values with highest performance, Table (3 and 4) show experimental results of both GSVM and KNN classifier on three different dataset. In movie review, KNN achieves best result when k=30, while it achieves best result when k=20 when classifying amazon toys reviews. When classifying financial dataset, KNN achieves best result when k=20.

Table 3: Comparative results between GSVM and KNN on movie reviews

Classifier	Criteria	k=1	k=5	k=10	k=20	k=30
KNN	Accuracy	62.0	56.5	56.7	55.6	59.50
Precision	62.5	55.0	54.1	54.0	57.43	
Recall	74.36	92.0	94.0	98.0	99.00	
F1 Score	67.2	69.2	69.9	70.3	72.69	
GSVM Accuracy			89.7			
Precision			88.3			
Recall			89.5			
F1 Score			90.5			

Table 4: Comparative results between GSVM and KNN on Amazon toys reviews

Classifier	Criteria	k=1	k=5	k=10	k=20	k=30
KNN	Accuracy	61.5	55.5	58.5	67.0	58.0
	Precision	60.5	54.5	52,5	55.3	56.5
	Recall	72.35	90.5	94.5	99.5	92.5
	F1 Score	65.2	67.0	70.1	74.3	70.5
GSVM	Accuracy			88.2		
	Precision			85.6		
	Recall			87.3		
	F1 Score			86.5		

Table 5: Comparative results between GSVM and KNN on financial dataset

Classifier	Criteria	k=1	k=5	k=10	k=20	k=30
KNN	Accuracy	77.8	81.5	80.3	87.6	78.5
	Precision	81.5	83.7	79.5	83.2	79.5
	Recall	72.5	76.0	85.6	73.0	75.8
	F1 Score	76.3	79.5	79.6	77.5	76.0
GSVM	Accuracy			86.5		
	Precision			87.5		
	Recall			88.6		
	F1 Score			89.7		

Table 6: Classificaiton speed in (ms)

Classifier	Movie reviews	Amazon toys	Financial
GSVM	35,000.00	27,500.00	28,300.00
KNN	20,000.00	19,500.00	20,750.00

As shown in all results, GSVM is much better than KNN classifier with different k values in all our case studies.

On the classification speed side, the results in Table 6 show that the classification based on KNN has faster speed than GSVM. The reason is that GSVM use multiple rounds of feature filtration of input vector space.

In general, from the above results, selection between both technique is depending on business domain and real case study. GSVM will be better option if we take classification accuracy into account. On other side, if real time if we take classification speed into account KNN will be better option.

Conclusion and Future Work

The scope of this research is to present comparative study between GSVM and KNN. We use movie review data set from twitter source as input for this experimental study. From classification accuracy perspective, the result shows that GSVM approach outperform the KNN technique. While observed that KNN takes less time in implementing classification. A future work direction is to implement parallel processing of GSVM that allow speed up of calculation. Morever, more comparative studies need to be presented with Nonstationary LDA (NSLDA) classification rule which is based on the Kalman Smoother algorithm.

Acknowledgement

The authors are grateful to the editors and the anonymous referees for their valuable comments. Their constructive comments have already improved the paper from its previous version and their knowledgeable advice is highly appreciated.

Author's Contributions

Hany Mohamed: Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Ayman Atia: Organized the study, shared in experiments planning and discussions, contributed to the writing of the manuscript and reviewing it critically for sigininfcat intellectual content. Gave the final approval of the version to be submitted.

Mostafa-Sami M. Mostafa: Designed the research plan and organized the study and contributed to the writing of the manuscript and reviewing it critically for signinficatin intellectual conten. Gave the final approval of the version to be submitted.

Ethics

The author confirm that they have thoroughly seen the content of the paper and do not find any conflict of interest and ethical issues.

References

Ghoreishi, S.F. and D.L. Allaire, 2018. Gaussian process regression for bayesian fusion of multi-fidelity information sources. Proceedings of the Multidisciplinary Analysis and Optimization Conference, AIAA AVIATION Forum, (AIAA 2018-4176), Atlanta, Georgia. DOI: 10.2514/6.2018-4176

Guo, G., H. Wang, D. Bell, Y. Bi and K. Greer, 2003. KNN model-based approach in classification. Lect. Notes Comput. Sci., 2888: 986-996.

He, R. and J. McAuley, 2016. Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In: Image-based Recommendations on Styles and Substitutes. McAuley, J., C. Targett, J. Shi and A. van den Hengel (Eds.), SIGIR.

Heather Whiting, R., P. Hansen and A. Sen, 2017. A tool for measuring SMEs' reputation, engagement and goodwill: A New Zealand exploratory study. J. Intellectual Capital, 18: 170-188. DOI: 10.1108/JIC-02-2016-0028

Imani, M. and U. Braga-Neto, 2018. Control of Gene Regulatory Networks using Bayesian Inverse Reinforcement Learning, Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics, April, 20, IEEE Xplore press, pp: 1-1. DOI: 10.1109/TCBB.2018.2830357

Kalaivani, P. and K.L. Shunmuganathan, 2014. An improved k-nearest-neighbor algorithm using genetic algorithm for sentiment classification, Proceedings of the International Conference on Circuit, Power and Computing Technologies, (ICCPCT' 14).

Kamble1, V.S. and S.N. Deshmukh, 2016. PMI Based Sentiment Analysis with SVM Cross Validation. IJESC, 6: 8579- 8582.

Kang, H., S.J. Yoo and D. Han, 2011. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems Applications, 39: 6000-6010. DOI: 10.1016/j.eswa.2011.11.107

Lan, M., C.L. Tan, J. Su and Y. Lu, 2009. Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Analysis Machine Intelligence, 31: 721-735.

Li, X., J. Pang, B. Mo, Y. Rao and F.L. Wang, 2016. Deep Neural Network for Short-Text Sentiment Classification. In: Database Systems for Advanced Applications, Gao, H., J. Kim and Y. Sakurai (Eds.), Lecture Notes in Computer Science, Springer, Cham.

Mohamed, H. and A. Ezzat, 2015. Mostafa Sami, the Road to Emotion Mining in Social Network. Int. J. Computer Applications.

Mohamed, H., A. Atia and M. Sami 2017. Mood Miner: Sentiment mining of financial market. ESOLEC, pp: 115-121.

Neviarouskaya, A., H. Prendinger and M. Ishizuka, 2009. Compositionality principle in recognition of fine-grained emotions from text, Proceedings of the 3rd International ICWSM Conference, Association for the Advancement of Artificial, Japan.

Preotiuc-Pietro, D., S. Samangooei, T. Cohn, N. Gibbins and M. Niranjan, 2012. Trendminer: An architecture for real time analysis of social media text, Proceedings of the Workshop on Real-Time Analysis and Mining of Social Streams (RAMSS; 12), Dublin.

Shivhare, S.N. and S. Khethawat, 2012. Emotion detection from text. Proceedings of the 2nd International Conference on Computer Science, Engineering and Applications (ICCSEA), Delhi, India, pp: 1-7.

Ye, Q., Z. Zhang and R. Law, 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems Applications, 36: 6527-6535.

Zaguai, F. and T.U. Beizak, 2015. Genetic Algorithm based Feature Selection in High Dimensional Text Dataset Classification. WSEAS transactions on information science and applications.

Zhang, L., R. Ghosh, M. Dekhil, M. Hsu and B. Liu, 2011. Combining Lexicon based and Learning-based Methods for Twitter Sentiment Analysis, Technical Report, HPL-2011-89.