International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

SENTIMENT ANALYSIS USING FUSION CUCKOO SEARCH TECHNIQUE FOR SOCIAL MEDIA TEXT

[1]Harshal Thakre [2]Vaibhav Pate [3]Anurag Shrivastava
[1]M.Tech Scholar, [2]Assistant Professor [3]Assistant Professor and Head
Department of Computer Science & Engineering, NIIST, Bhopal, India-23

Abstract

Sentiment analysis is one of the prominent fields of data mining that deals with the identification and analysis of sentimental contents generally available at social media. Twitter is one of such social medias used by many users about some topics in the form of tweets. These tweets can be analyzed to find the viewpoints and sentiments of the users by using clustering-based methods. However, due to the subjective nature of the Twitter datasets, metaheuristic-based clustering methods outperform the traditional methods for sentiment analysis. Therefore, this paper proposes a novel metaheuristic method (CSK) which is based on K-means and cuckoo search. The proposed method has been used to find the optimum cluster heads from the sentimental contents of Twitter dataset. The efficacy of proposed method has been tested on different Twitter datasets and compared with particle swarm optimization, differential evolution, cuckoo search, improved cuckoo search, gauss-based cuckoo search, and two n-grams methods. Experimental results and statistical analysis validate that the proposed method outperforms the existing methods. The proposed method has theoretical implications for the future research to analyze the data generated through social networks/medias. This method has also very generalized practical implications for designing a system that can provide conclusive reviews on any social issues

Introduction

The unrivalled increase in the acceptance as well as penetration of social media platforms, such as Facebook, Twitter, Google plus, etc., in a day to day life, havechanged the pattern of online communication of people. Formally, user's online access was highly restricted to professional contents such as news agencies or corporations. However, these days they can seamlessly interact with each other in a more concurrent way by creating their own content within a network of peers. According to Howard [29], "We use Facebook to schedule the protest, Twitter to coordinate, and YouTube to tell the word". Social media has emerged as a vital platform of representing people's sentiment, boosting the requirements of data mining in the field of the sentiment analysis. In the sentiment analysis, the raw data is the online text that is exchanged by users through social media [65]. Twitter, which is one of such social media, has become the prominent source to exchange the online text, providing a vast platform of sentiment analysis. Twitter is a very popular social networking website that allows registered users to post short messages, also called tweets, up to 140 characters. Twitter database is one of the largest database having 200 million users who post 400 million messages/tweets in a day [56]. At Twitter, users often share their personal opinion on different subjects such as acceptance or rejection of politicians and viewpoint about products, talk about current issues and share their personal life events. However, users post their tweets with fewer characters by using a short form of words and symbols such as emoji. Therefore, analysis of these tweets can be used to find strong viewpoints and sentiments for any topic. Twitter data has already been used by different people to predict stock market prediction Bollen, Mao, & Zeng, [11] , box office revenues for movies Asur & Huberman, [5],identify the clients with negative sentiments Thet, Na, & Khoo, [66], etc.

Sentiment analysis

Sentiment analysis is the measurement of positive and negative language. It is a way to evaluate written or spoken language to determine if the expression is favorable, Unfavorable, or neutral and to what degree, It is one of the prominent fields of data mining that deals with the identification and analysis of sentimental contents generally available at social media. Today's algorithm-based sentiment analysis tools can handle huge volumes of customer feedback consistently and accurately. Paired with text analytics, sentiment analysis reveals the customer's opinion about topics ranging from your products and services to your location, your advertisements, or even your competitors.

Why is sentiment analysis important?

Sentiment analysis is critical because helps you see what customers like and dislike about you and your brand. Customer feedback from social media, your website, your call centre agents, or any other source contains a treasure trove of useful business information. But, it isn't enough to know what customers are talking about. You must also know how they feel. Sentiment analysis is one way to uncover those feelings. The main aim of sentiment analysis is to determine the attitude of users on a particular topic. Therefore, this work proposes a novel clustering method for sentiment analysis on Twitter dataset. Sentiment analysis methods can be broadly categorized into lexicon based methods, machine learning-based methods, and hybrid methods Medhat, Hassan, & Korashy, [46] which can be further classified into sub-category as depicted in Fig. 1

LITERATURE SURVEY

Lexicon-based methods require predefined sentiment lexicon to determine the polarity of any document. However, the accuracy of lexicon-based method is reduced drastically in the presence of emoticons and short hand texts, as they are not the part of predefined sentiment lexicon[39] Emoticons are the visual emotional symbols used by the users at social medias (Hu,Tang, Gao, & Liu, [22]. Hu, Tang, Tang, and Liu [23] proposed a novel method of sentiment analysis that considers the short texts like "gudnite" and emotional symbols such as ":)", in a unified frame- work. The performance of this method does not show stability on some of the emotional signals, such as emoticons, when used on datasets from different domains (Hu et al., [22]). This problem can be resolved by examining the contributions of other emotion indication information existing in social media, like product ratings, restaurant reviews, and other emotion correlation information (Hu et al., [22]; Yusof,Mohamed, & Abdul-Rahman, [8]) such as correlation between two words in a post. Emotion indication represents the sentiment polarity of a post and further, it is classified into post level emotion indication (emoticons) and world level emotion indication (publicly available sentiment lexicons) (Hu et al., [22]). More- over, emotion correlation for posts are usually represented by a graph in which nodes represent the data points and edge represent correlation between the words Canuto, Gonçalves, and Benevenuto [15] proposed a new sentiment- based meta-level features for effective sentiment analysis. This method has a capability to utilize the information from the neighborhood effectively and efficiently to capture important information from highly noise data. Bravo-Marquez, Mendoza, and Poblete [36] introduced a novel supervised method to combine strengths, emotions, and polarities for improving the Twitter sentiment analysis process. Kontopoulos, Berberidis, Dergiades, and Bassiliades [12] proposed ontology-based sentiment analysis of tweets. In this method, a sentiment grade has been assigned for every distinct notion in the tweets. Further, Kranjc, Smailovi ´c, Podpe ˇcan, Gr ˇcar, Žnidarši ˇc and Lavra ˇc [43] [10]. However, K-means method has its own limitations like data size, shape, balance, etc. For the same, overlapping clustering methods Yokoyama, Nakayama, & Okada, [64] are being used to improve the accuracy and to reduce the limitations of K-means. Recently, sentiment analysis methods have used natural language processing (NLP) to add semantics in feature vector which improves the accuracy of the classifiers Kanakaraj & Guddeti, 2015; Saif, Ortega, Fernández, & Cantador, 2016b [34, 59]. To illustrate certain facets of natural language semantics, Altınel and Ganiz [3] proposed novel semantic smoothing kernels which is used by class term matrices, a new type of vector space models (VSM), to extract class specific semantics, Wiratunga, and Lothian [49] introduced a lexicon-based sentiment classification system which uses textual neighborhood (local context) interaction and text category (global context) for social media genres Appel, Chiclana, Carter, and Fujita [4] presented a hybridized method which uses NLP and fuzzy sets to determine semantic polarity and its intensity for posts. Furthermore, Cambria [16] discussed merits and limitations of various sentiment analysis methods such as knowledge based, statistical, and hybrid. Shah et al. [52] presented a multimedia summarization system to analyze online user generated contents (UGCs) from multiple modalities. For the same, they have used the Event Builder system for semantics understanding and Event Sensor system for sentics understanding. Chen, Xu, He, Xia, and Wang [19] introduced a document-level sentiment analysis method using sequence modeling-based neural network. Further, Sulis, Farías, Rosso, Patti, and Ruffo [55] investigated the effect of figurative linguistic phenomena in twitter to separate the tweets with tag #irony, #sarcasm and #not using psycholinguistic and emotional features. Metaheuristic-based methods have also been used for sentiment analysis. Basari, Hussin, Ananta, and Zeniarja [5] pro- posed a hybrid method based on support vector machine (SVM) and particle swarm optimization (PSO) to categorize a movie into watchable and non-watchable. Due to the above mentioned limitations of traditional as well as metaheuristic-based clustering methods, this paper introduces a novel metaheuristic method (CSK) which is being used to cluster the sentimental contents. The proposed method, which is based on cuckoo search (CS) Yang & Deb [75] and K-means Žalik, [78], optimizes the cluster-heads of sentimental datasets. Moreover, the performance of the proposed method has been compared with cuckoo search (CS), improved cuckoo search algorithm (ICS) Valian, Mohanna, & Tavakoli,[72], Gauss based cuckoo search algorithm (GCS) (Zheng & Zhou, [79], particle swarm optimization algorithm (PSO) , differential evolution (DE) , and two n-grams (basic baseline method).

Problem Domain

Due to the subjective and implemented description of various methods described in literature survey and because of nature of the Twitter datasets, metaheuristic-based clustering methods outperform the traditional methods for sentiment analysis. Therefore, we propose metaheuristic method (Cuckoo Search K-mean) which is based on K-means and cuckoo search. The proposed method has been used to find the optimum cluster-heads from the sentimental contents of Twitter dataset.

Proposed Method

The proposed method CSK (Cuckoo Search K-mean) clusters the input tweets in three phases;

(i) Preprocessing of the tweets,
(ii) Feature extraction, and
(iii) Hybrid clustering using K-means and cuckoo search. The detailed flow chart of the proposed method has been shown in Fig. 2.

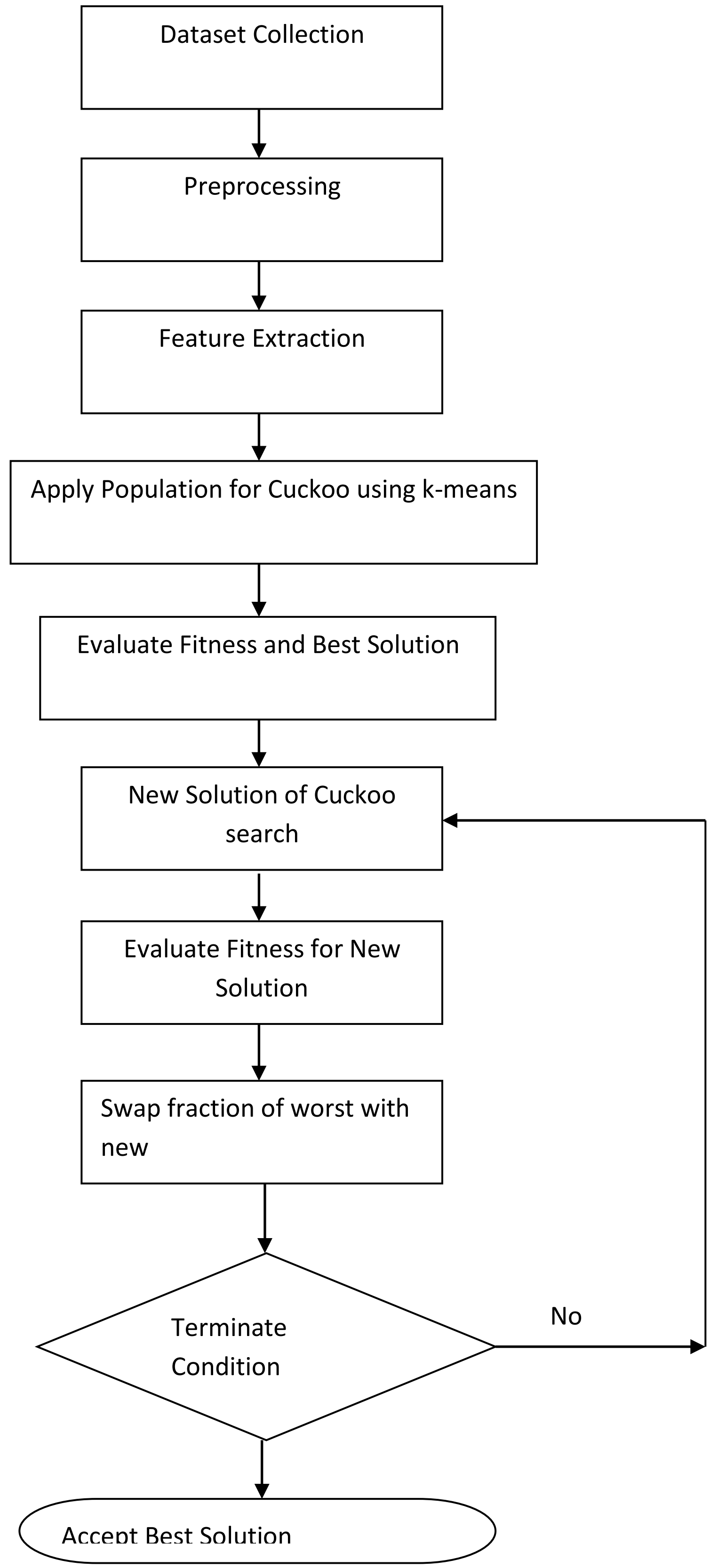

Fig 2. Flow of Proposed Work

Preprocessing

The raw tweets, collected from Twitter, have noise in terms of unwanted and fuzzy words, URLs, stopwords etc., which are needed to be reduced before feature extraction. Therefore, the proposed method uses the following preprocessing method in two phases before extracting the features:

Phase 1

This phase eliminates unwanted noise elements from the Twitter data set using the following steps

1. Eliminate all the URLs via regular expression matching. A regular expression is a textual pattern that defines a search pattern for strings/text. It can be used to search for URLs, email address etc. The list of regular expression used in this work is following

Regular expression

1. Regular expression to replace URLs from a tweet with string url tweet = re.sub (’((www\.[^\s]+)|(https?://[^\s]+))’,’url’, tweet)

2. Regular expression to remove @username from tweet

Tweet=re.sub(’(?<=^|(?<=[^a-zA-Z0-9-_.]))@([A-Za-z]+[A-Za-z0-9]+)’,’’,tweet)

3. Regular expression to remove additional white spaces from tweet tweet=re.sub(’[\s]+’,’’,tweet)

4. Regular expression to replace #word with word in tweet tweet=re.sub(r’#([^\s]+)’,r’\1’,tweet)

5. Regular expression to strip punctuations from a word

Tweet=tweet. strip(’-’’()’’\’’/’)

2. Replace “@ Username” with “usr” using regular expression matching.

3. Since “hash-tag(#)”provides some useful information, therefore remove only #, keeping the word as it is. viz. , “# Lee” is replaced with “Lee”.

4. Remove parenthesis, forward slash (/), backward slash (\ ), and dash from tweets.

5. Replace multiple white spaces with single white space.

Phase 2

In this phase, two dictionaries namely; stop word [1, 23, 24] and acronym (Acronym dictionary [23, 24] has been deployed to improve the precision of resultant Twitter dataset of Phase 1. The steps of Phase 2 are as follows:

1. Convert all the words of tweets into lowercase.

2. Remove all the stop words such as, a, is, the, etc. by comparing them with stop word dictionary [1].

3. Replace sequence of repeated characters (three or more) in a word by one character viz., “hellooooo” is converted to “Hello”.

4. Eliminate words which do not start with an alphabet.

5. Replace all the short forms in the respective full forms using acronym dictionary [1]

Feature Extraction Method

After applying the preprocessing [24] tweets are converted into the feature vector by calculating the following 11 features from the Twitter dataset.

1. Total Characteristics: It represents the total number words available in the tweets.

2. Positive Emoji: Positive emoji, such as :), ;), : D , etc., are the symbols used to express happy moments. This feature uses a positive emoticon dictionary [23] to count the total number of positive emojis in the tweets.

3. Negative Emoji: The special symbols used to express sad/ negative feelings, such as: (, : _ (, > : (, etc., are known as negative emoji. To get the total counts of negative emoji in tweets a negative emoticon dictionary [23] is used.

4. Neutral Emoji: Neutral emoji (straight-faced emoji) do not provide any particular emotion. Total neutral emoji is counted by comparing tweets with neutral emoticon dictionary [23, 24]

5. Positive Exclamation: Exclamatory words, such as hurrah! wow! etc., can be used to convey a very strong feeling/ opinion about the topic. For the same, positive exclamation dictionary [1] is used to count the positive exclamation.

6. Negative Exclamation: Negative exclamations are counted by comparing the tweet with negative exclamation dictionary [1, 23, 24]

7. Negation: To express the negative opinion, negations words like no, not, etc., are generally used. Therefore, this feature counts the negation words in the tweet by comparing it with negation words.

8. Positive Words: This feature counts the number of positive words like achieve, confidence, etc., using positive word dictionary [34, 44]. If there are two negative words (double negation) then these words are counted as single positive word.

9. Negative Words: This feature represents the total counts of negative words such as bad, lost, etc., in tweets [33, 34]

10. Neutral Words: Neutral words (okay, rarely) do not provide any particular emotion/feeling. Total counts of neutral words are obtained by comparing the tweets with neutral word dictionary [54]

11. Intense Words: Intense words, like very, much etc. are used in a sentence to make it more effective/intense. Total counts of intense words are determined by using intense word dictionary [54]

Hybrid clustering using K-means & cuckoo search

The normalized feature vector is given input to the proposed clustering method which uses K-means and cuckoo search method to cluster the data. As K-means is very popular cluster method, but it generally stuck to initial an cluster which is a major drawback of K-means method, However the generated clusters can be used for further analysis. Therefore, in this method, the generated clusters from K-means have been used in the cuckoo search method for further optimizing the cluster-heads. Since, in the cuckoo search, a random initialization of the population is required and this may increase the number of iterations to converge and also stuck to some local solution. Therefore, this method modifies the initialization process of cuckoo search which results in faster convergence and better optimum solution. In the CSK, the solutions obtained from K-means are used to initialize the population of cuckoo search, which resolve the problem of random initialization in CS. Thereafter cuckoo search is executed for obtaining the optimum result and faster convergence. Let there be n number of tweets which are to be clustered into N classes. Each tweet is represented by a feature vector having S number of features and each feature has been scaled in [0, T]. The probability distribution of each feature can be defined as follows [47]

$$P_i = \frac{O_i}{n} \tag{2}$$

Where i represents the ith feature value, i.e., $0 \leq i \leq T$, and Oi denotes the total number of tweets having ith feature value. Moreover, the total mean of each feature is calculated using Eq. (3).

$$\mu = \sum_{i=1}^{T} iP_i \tag{3}$$

Any tweet is classified into class D j for which it has minimum Euclidean distance. Therefore, the probability (w j) of occurrence of class D j (j = 1, 2 .., N) is given by Eq. (4).

$$W_j = \sum_{i \in D_j} P_i \tag{4}$$

The mean of class D j can be calculated by Eq. (5) .

$$\mu_j = \sum_{i \in D_j} \frac{ip_i}{w_j} \tag{5}$$

The inter-class variance can be generally defined as:

$$\sigma^2 = \sum_{j=1}^{N} W_j \left(\mu_j - \mu\right)^2 \tag{6}$$

To cluster the different tweets into their respective class, the inter-class variance shown in Eq. (6) should be maximized. Therefore, the objective function for the proposed hybrid cuckoo search method is to maximize the functions as defined in Eq. (6). The detailed steps of the proposed method are given in Algorithm 3.

Algorithm 3 Proposed method

Set the size of population as N.

for i = 1 to N do

Generate k clusters using the K-means algorithm.

Use k cluster-heads to initialize the population of cuckoo search

end for

Calculate the fitness of these N solutions by using objective function

while t < MaxGeneration do

Generate N new solutions using Cuckoo Search

Calculate the fitness of new solutions

Remove the old solutions with better new solutions

Replace the fraction (P_a) of worse solutions by random new solutions

end while

Print the best solution and its fitness

Twitter dataset

The Twitter dataset (twitter, 2014) has been taken from Twitter which is based on the topics of sports, saints, funny images, jokes, and college students. This dataset has 2000 tweets posted from No. 17, 2014 to Dec 10, 2014. The considered dataset is manually labelled in two classes namely; positive and negative, each containing 10 0 0 tweets. In dataset, positive tweets are represented by 1 and negative tweets by 0.

Table 1 Considered Twitter datasets

Sr No	Dataset	Number of Instances	Number of Classes	Positive	Negative	Neutral	Date Range	Topic Covered
1	Testdata.manual.2009.06.14	498	3	182	177	139	May 11, 2009 to Jun 14, 2009	Google, Obama, Kindle, China
2	Twitter-sanders-apple2	479	2	163	316		Oct 15, 2011 to Oct 20, 2011	Apple, Google, Microsoft, Twitter
3	Twitter-sanders-apple3	988	3	163	316	509	Oct 15, 2011 to Oct 20, 2012	Apple, Google, Microsoft, Twitter
4	Twitter dataset	2000	2	1000	1000		Nov 17, 2014 to Dec 10, 2014	Sports, Saint, Funny Images, etc.

Evaluation of the Feature Extraction Process the FE process as evaluated based on two traditional measures used in Sentiment Analysis and Text Classification: precision and recall. We also computed the F-measure, a combined metric that takes both precision and recall into consideration, as follows F-measure = 2 × Precision × Recall Precision + Recall (2) In order to calculate these metrics, it was necessary to manually extract the relevant features appearing in the opinions on the validation corpus. This task was performed as follows: for each sentence containing opinions, all the implicit and explicit features evaluated by the user were identified and stored on a separate file. This list was then compared to the list of automatically extracted features, and the precision, recall and F-measure rates were calculated. We did not consider the computational cost of our solution as relevant because, in most SA applications, the feature extraction process is an off-line activity that is not frequently repeated [15].

Experiment Tool

Orange Canvas: Orange is a library of C++ core objects and routines that includes a large variety of standard and not-so-standard machine learning and data mining algorithms, plus routines for data input and manipulation. Orange is also a scriptable environment for fast prototyping of new algorithms and testing schemes. It is a collection of Python-based modules that sit over the core library and implement some functionality for which execution time is not crucial and which is easier done in Python than in C++. This includes a variety of tasks such as pretty-print of decision trees, attribute subset, bagging and boosting, and alike. Orange is also a set of graphical widgets that use methods from core library and Orange modules and provide a nice user's interface. Widgets support signal-based communication and can be assembled together into an application by a visual programming tool called Orange Canvas.

Experimental Setup

Fig 3 Experimental Setup

Result Analysis

The Twitter dataset has been pre-processed to remove the undesired words and characters. From the pre-processed dataset, 11 features have been extracted as shown in Table 3 along with their mean and standard deviation values for each dataset. The statistical mean shows the central tendency of each dataset. From the table, it is observed that each dataset is unbiased and contains different types of words which may affect the clustering accuracy. Further, standard deviation shows that each feature has sufficient variation in tweets. Moreover, the proposed method has been compared with three existing methods namely; two word-level n-grams (support

vector machine-trigram (SVM-tri) and Naive Bayes-trigram (NB-tri)), cuckoo search (CS), the considered n-grams are weighted using term frequency (tf) and the value of n has been selected using cross-validation (rotation estimation) (Kohavi et al., 1995). The parameter settings for all the considered methods have been presented in Table 4. To measure the performance of the proposed method, three parameters have been considered namely; accuracy, computational time, and fitness function value. Table 4 shows the comparative results of the proposed method and existing considered methods in terms of all the above three parameters. For fair comparison, each method has been executed 30times and Table 4 represents the mean values of accuracy, computational time and fitness function values. From the table, it is visualized that the proposed method gives the best accuracy among all the considered methods. Moreover, the proposed method also outperforms in the mean fitness function value. Further, the proposed method is computationally efficient as compared to other existing methods except DE method. However, the main concern is the accuracy of the system, where proposed method outperformed. To test the significant difference between the proposed method and considered methods, a statistical comparison is performed for accuracy, computational time, and fitness function value using student's t -test (Owen, 1965) with a confidence level of 95%. In this experiment, student's t -test is applied for the null hypothesis that there is no significant difference in the parameter values for 30 runs with respect to proposed method and existing methods. Moreover, to compare the performance of all the considered methods and proposed method, Graph analysis (McGill, Tukey, & Larsen, 1978) is carried out. The Graph Plot graphically represents the empirical distribution of the data. The Graph Plot for existing and proposed methods are shown in Figs a. In the box plot, the x-axis represents the name of the methods and the corresponding parameters under consideration on the y-axis. From the Graph plots, it is observed that proposed method gives the better and consistent results for all the considered performance parameters except computational time where DE outperforms. To show the convergence behavior of all the considered methods and proposed method convergence plot have also been plotted in Fig. In the convergence plot, the x and y-axis represent the number of iterations and fitness function values respectively. From the convergence plots, it is observed that proposed method converges quickly as compared to all the considered methods and gives the better results.

Table 3 Parameter settings for all the considered datasets

Sr	Parameter	CS	SVM	Naive Bayes	CSK
1	Probability (Pa)	0.64	0.07-0.8	0.25	0.28
2	Step scaling factor (α)	0.04	0.01-0.8	0.04	1
3	Number of iterations	450	450	450	450
4	AUC	-	0.985	1	1
5	CA	-	0.970	0.97	1.025
6	F1	-	0.96	0.96	0.96
7	Precision	-	0.955	0.952	0.96
8	Recall	-	0.970	0.971	1

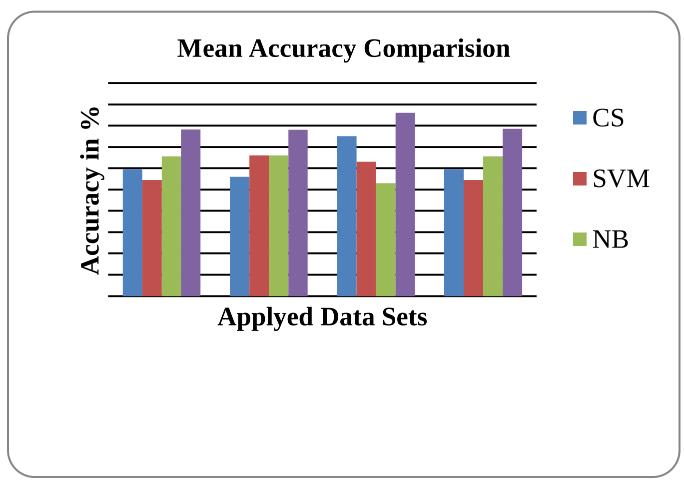

Fig 6 Mean Accuracy Graph

Table 4 Comparison of proposed method with the existing methods in terms of mean accuracy, mean computational time, and mean fitness function value.

Sr No	Data Set	Method	Mean Acc	Mean Computation Time	Mean Fitness value
1	Testdata.manual.2009.06.14	CS	59 .54%	293	0 .2506
2		SVM	54 .54%	332	-
3		NB	65 .54%	316	0.2789
4		New CSK	78.24%	299	0.2884
Sr No	Data Set	Method	Mean Acc	Mean Computation Time	Mean Fitness value
1	Twitter-sanders-apple2	CS	56%	293	0 .2406
2		SVM	66%	332	-
3		NB	66%	316	0.2289
4		New CSK	78%	284	0.28
Sr No	Data Set	Method	Mean Acc	Mean Computation Time	Mean Fitness value
1	Twitter-sanders-apple3	CS	75%	274	0 .2530
2		SVM	63%	300	-
3		NB	53%	272	0.2789
4		New CSK	86%	241	0.279
Sr No	Data Set	Method	Mean Acc	Mean Computation Time	Mean Fitness value
1	Twitter dataset	CS	59 .54%	293	0 .2570
2		SVM	54 .54%	332	-
3		NB	65 .54%	316	0.2801
4		New CSK	78.58%	270	0.2884

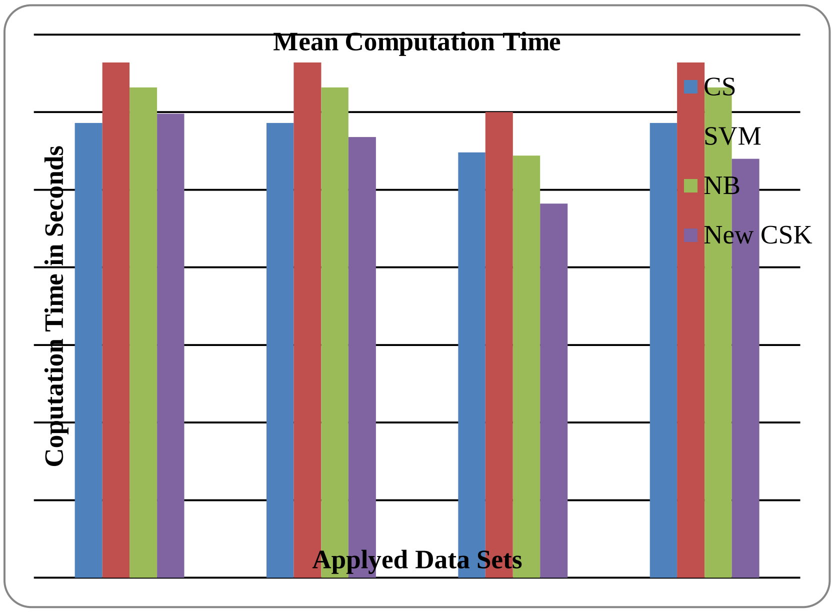

Fig 7 Mean Computation Time Graph

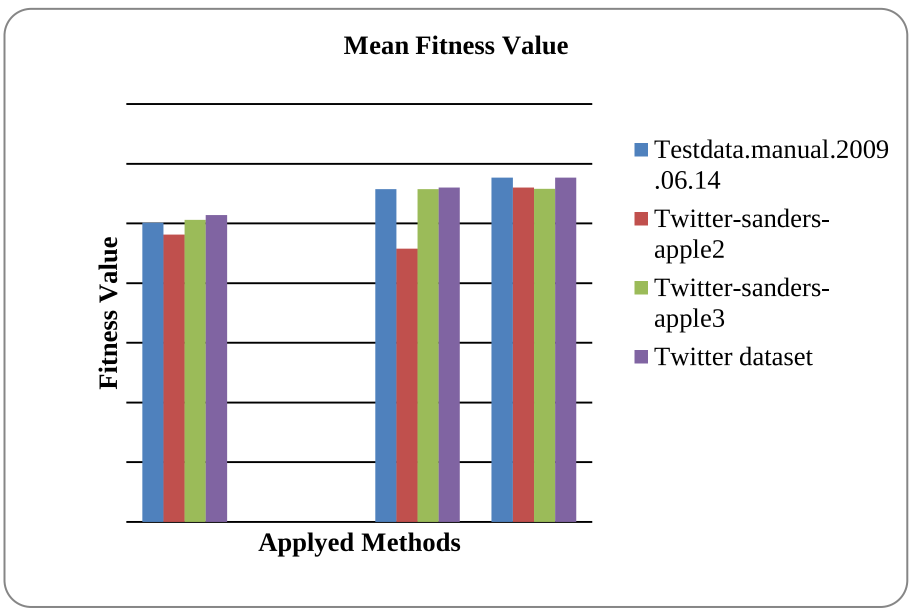

Fig 8 Mean Fitness Value Graph

Comparisons represent that applied values in different algorithms are showing variable results and CSK method is showing high mean value, less computation time, and high mean fitness value for all applied datasets.

References

[1] Acronym dictionary, (2015). www.netlingo.com/acronyms.php.
[2] Agarwal, B. , Mittal, N. , Bansal, P. , & Garg, S. (2015). Sentiment analysis using common-sense and context information. Computational Intelligence and Neuro- science, 2015, 30.
[3] Altınel, B. , & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowledge-Based Systems, 108 , 50–64 . 2016
[4] Appel, O. , Chiclana, F. , Carter, J. , & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems .
[5] Asur, S. , Huberman, B. , et al. (2010). Predicting the future with social media. In International conference on web intelligence and intelligent agent technology (WI-IAT): 1 (pp. 4 92–4 99). IEEE .
[6] Basari, A. S. H. , Hussin, B. , Ananta, I. G. P. , & Zeniarja, J. (2013). Opinion mining of movie review using hybrid

method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453–462.

[7] Bello-Orgaz, G. , Menéndez, H. D. , & Camacho, D. (2012). Adaptive k-means algorithm for overlapped graph clustering. International Journal of Neural Systems, 22 .

[8] Bharti, S. , Vachha, B. , Pradhan, R. , Babu, K. , & Jena, S. (2016). Sarcastic sentiment detection in tweets streamed in real time: A big data approach. Digital Communications and Networks, 2 (3), 108–121.

[9] Bharti, S. K., Babu, K. S., & Jena, S. K. (2015). Parsing-based sarcasm sentiment recognition in twitter data. In 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) (pp. 1373–1380). IEEE.

[10] Boiy, E. , Hens, P. , Deschacht, K. , & Moens, M.-F. (2007). Automatic sentiment analysis in on-line text. In ELPUB (pp. 349–360).

[11] Bollen, J. , Mao, H. , & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2, 1–8.

[12] Bravo-Marquez, F., Frank, E., & Pfahringer, B. (2016). Building a twitter opinion lexicon from automatically-annotated tweets. Knowledge-Based Systems

[13] Bravo-Marquez, F. , Mendoza, M. , & Poblete, B. (2013). Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In Proceedings of the second international workshop on issues of sentiment discovery and opinion mining (p. 2). ACM.

[14] Brown, C. T. , Liebovitch, L. S. , & Glendon, R. (2007). Lévy flights in dobe ju/hoansi foraging patterns. Human Ecology, 35, 129–138 .

[15] Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, Vol. 31 , 102–107 .

[16] Canuto, S. , Gonçalves, M. A. , & Benevenuto, F. (2016). Exploiting new sentiment-based meta-level features for effective sentiment analysis. In Proceedings of the ninth ACM international conference on web search and data mining (pp. 53–62). ACM .

[17] Carstens, L. (2016). Using argumentation to improve classification in natural language problems.

[18] Carvalho, P. , Sarmento, L. , Silva, M. J. , & De Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! It's so easy ;-). In Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion (pp. 53–56). ACM.

[19] Chen, T., Xu, R., He, Y., Xia, Y., & Wang, X. (2016). Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Computational Intelligence Magazine, 11 (3), 34–44.

[20] Chiang, M. M.-T. , & Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads. Journal of Classification, 27 (1), 3–40 .

[21] Coletta, L. F. S. , da Silva, N. F. F. , Hruschka, E. R. , & Hruschka, E. R. (2014). Combining classification and clustering for tweet sentiment analysis. In Intelligent systems (BRACIS), 2014 Brazilian conference on (pp. 210–215). IEEE.

[22] Danielsson, P.-E. (1980). Euclidean distance mapping. Computer Graphics and Image Processing, 14 , 227–248 .

[23] Emoticon dictionary, (2015). http://www.netlingo.com/smileys.php.

[24] Exclamation word dictionary, (2015). http://www.Vidarholen.net/contents/interjections/

[25] Fernández-Gavilanes, M. , Álvarez-López, T. , Juncal-Martínez, J. , Costa-Montenegro, E. , & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. Expert Systems with Applications, 58, 57–75.

[26] González-Ibánez, R. , Muresan, S. , & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2 (pp. 581–586). Association for Computational Linguistics .

[27] Gupta, D. K. , Reddy, K. S. , Ekbal, A. , et al. (2015). Pso-asent: Feature selection using particle swarm optimization for aspect based sentiment analysis. In Natural language processing and information systems (pp. 220–233). Springer.

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

[29] Howard, P. N. (2011). The Arab springs cascading effects. Pacific Standard, 23.

[30] Hu, X. , Tang, J. , Gao, H. , & Liu, H. (2013a). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on world wide web (pp. 607–618). ACM .

[31] Hu, X. , Tang, L. , Tang, J. , & Liu, H. (2013b). Exploiting social relations for sentiment analysis in microblogging. In Proceedings of the sixth ACM international conference on web search and data mining (pp. 537–546).

[32] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31 (8), 651–666 .

[33] Jeffrey Breen (2015). Positive word dictionary. twitter-sentiment-analysis-tutorial-201107/data/opinion-lexicon-English/positive-words.txt. 2011 (accessed December 15.

[34] Jeffrey Breen (2015). Negative word dictionary. twitter-sentiment-analysis-tutorial-201107/data/opinion-lexicon-English/Negative-words.txt accessed December 15.

[35] Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic sarcasm detection: A survey. arXiv preprint arXiv:1602.03426,.

[36] Joshi, A. , Sharma, V. , & Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing: 2 (pp. 757–762)

[37] Kanakaraj, M., & Guddeti, R. M. R. (2015). Nlp based sentiment analysis on twitter data using ensemble classifiers. In Signal processing, communication and networking (ICSCN), 2015 3rd international conference on (pp. 1–5). IEEE.

[38] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Neural Networks, 4, 1942–1948.

[39] Khan, A. Z., Atique, M., & Thakare, V. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89.

[40] Kogan, J., Teboulle, M., & Nicholas, C. (2005). Data driven similarity measures for k-means like clustering algorithms. Information Retrieval, 8. , 331–349.

[41] Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai: 14 (pp. 1137–1145).

[42] Kontopoulos, E., Berberidis, C. , Dergiades, T. , & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications, 40, 4065–4074 .

[43] Kranjc, J., Smailovi ´c, J. , Podpe ˇcan, V. , Gr ˇcar, M. , Žnidarši ˇc, M. , & Lavra ˇc, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the clowdflows platform. Information Processing & Management, 51, 187–203.

[44] Liu, B. , Hu, M. , & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342–351).

[45] McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32, 12–16.

[46] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5 (4), 1093–1113.

[47] Mendenhall, W., Beaver, R., & Beaver, B. (2012). Introduction to probability and statistics. Cengage Learning.

[48] Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. Information Processing & Manage- ment, 51 (4), 4 80–4 99.

[49] Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. Knowledge-Based Systems.

[50] Nugent, R., Dean, N., & Ayers, E. (2010). Skill set profile clustering: The empty k-means algorithm with automatic specification of starting cluster centres. Educational data mining 2010.

[51] Owen, D. (1965). The power of student's t -test. Journal of the American Statistical Association, 60, 320–333.

[52] Pandarachalil, R., Sendhilkumar, S. , & Mahalakshmi, G. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. Cognitive Compu- tation, 7, 254–262.

[53] Pavlyukevich, I. (2007). Lévy flights, non-local search and simulated annealing. Journal of Computational Physics, 226 , 1830–1844 .

[54] Psychological feelings, (2015). http://www.psychpage.com/learning/library/assess/feelings.html.

[55] Qiu, G., Liu, B., Bu, J., & Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In IJCAI: 9 (pp. 1199–1204).

[56] Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. Data & Knowledge Engineering, 74, 1–12.

[57] Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In Proceedings of international conference on empirical methods in natural language processing (pp. 1524–1534).

[58] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016a). Contextual semantics for sentiment analysis of twitter. Information Processing & Management, 52 (1), 5–19. 2016

[59] Saif, H., Ortega, F. J., Fernández, M., & Cantador, I. (2016b).Sentiment analysis in social streams, 2016,.

[60] Sarcasm, (2016). http://examples.yourdictionary.com/examples-of- sarcasm.html.

[61] Saraswat, M., Arya, K., & Sharma, H. (2013). Leukocyte segmentation in tissue images using differential evolution algorithm. Swarm and Evolutionary Com- putation, 11, 46–54.

[62] Shah, R. R., Yu, Y., Verma, A., Tang, S., Shaikh, A. D., & Zimmermann, R. (2016). Leveraging multimodal information for event summarization and concep- t-level sentiment analysis. Knowledge-Based Systems

[63] Stopwords dictionary, (2015). http://www.ranks.nl/stopwords.

[64] Storn, R. , & Price, K. (1997). Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimiza- tion, 11 , 341–359 .

[65] Sulis, E. , Farías, D. I. H. , Rosso, P. , Patti, V. , & Ruffo, G. (2016). Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. Knowledge-Based Systems.

[66] Tang, H. , Tan, S. , & Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36, 10760–10773 .

[67] Thet, T. T., Na, J.-C. , & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science.

[68] Testdata.manual.2009.06.14, (2015). http://help.sentiment140.com/for-students/.

[69] Twitter dataset, (2014). https://drive.google.com/file/d/0BwPSGZHAP _ yoN2pZcVl1Qmp1OEU/view?usp=sharing .

[70] Twitter-sanders-apple, (2015). http://boston.lti.cs.cmu.edu/classes/95- 865- K/HW/HW3/ .

[71] Uysal, A. K. , & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50 , 104–112 .

[72] Valian, E. , Mohanna, S. , & Tavakoli, S. (2011). Improved cuckoo search algorithm for feedforward neural network training. International Journal of Artificial Intelligence & Applications, 2 , 36–43 .

[73] Wilkinson, R. , & Hingston, P. (1991). Using the cosine measure in a neural network for document retrieval. In

Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval (pp. 202–210).

[74] Xia, R. , Xu, F. , Yu, J. , Qi, Y. , & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Information Processing & Management, 52 , 36–45 .

[75] Yang, X.-S. , & Deb, S. (2009). Cuckoo search via lévy flights. In World congress on nature & biologically inspired computing (pp. 210–214). IEEE .

[76] Yokoyama, S. , Nakayama, A. , & Okada, A. (2009). One-mode three-way overlapping cluster analysis. Computational Statistics, 24 , 165–179 .

[77] Yusof, N. N. , Mohamed, A. , & Abdul-Rahman, S. (2015). Reviewing classification approaches in sentiment analysis. In International conference on soft com- puting in data science (pp. 43–53). Springer.

[78] Žalik, K. R. (2008). An efficient k'-means clustering algorithm. Pattern Recognition Letters, Vol. 29 , 1385–1391 .

[79] Zheng, H. , & Zhou, Y. (2012). A novel cuckoo search optimization algorithm based on Gauss distribution. Journal of Computational Information Systems, 8, 4193–4200 .

[80] Zhu, J. , Wang, H. , & Mao, J. (2010). Sentiment classification using genetic algorithm and conditional random fields. In 2nd IEEE international conference on information management and engineering (ICIME) (pp. 193–196). IEEE.

[81] L. Ferreira, N. Jakob, and I. Gurevych, "A comparative study of feature extraction algorithms in customer reviews," in Proceedings of the 2nd Annual IEEE International Conference onSemantic Computing (ICSC '08), pp. 144–151, Santa Clara, Calif, USA, August 2008.

[82] G. Qiu, B. Liu, J. Bu, and C.Chen, "Opinion word expansion and target extraction through double propagation," Computational Linguistics, vol. 37, no. 1, pp. 9–27, 2011.

[83] M.Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the 10th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), pp. 168–177, August 2004.

[84] C.-P.Wei, Y.-M.Chen, C.-S.Yang, andC.C.Yang, "Understanding what concerns consumers: a semantic approach to product feature

extraction fromconsumer reviews," Information Systems and E-Business Management, vol. 8, no. 2, pp. 149–167, 2010.

[85] V.Hatzivassiloglou and K. R.McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Associa

[86] Sureka, V. Goyal, D. Correa, and A. Mondal, "Generating domain-specific ontology from common-sense semantic network for target specific sentiment analysis," in Proceedings of the 5th International Conference of the Global WordNet Association

[87] (GWC '10), Mumbai, India, January–February 2010.

[88] Yusof, N. N. , Mohamed, A. , & Abdul-Rahman, S. (2015). Reviewing classification approaches in sentiment analysis. In International conference on soft com- puting in data science (pp. 43–53). Springer .

[89] S. Mukherjee and S. Joshi, "Sentiment aggregation using concept net ontology," in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP '13), pp. 570–578, 2013.

[90] Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. Polarity analysis of texts using discourse structure. In Proceedings of the 20th ACM Conference on Information and Knowledge Management, 1061-1070, 2011

[91] Alexander Hogenboom, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. Using rhetorical structure in sentiment analysis. Communications of the ACM, 58(7):69-77, 2015.

[92] Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn A. Walker, and M. Inffes Torres. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. Knowledge-Based Systems, 69:124-133, 2014.

[93] Rosalind W. Picard. Affective Computing. The MIT Press, Cambridge, Massachusetts, 1st edition, 2000.

[94] Soujanya Poria, Erik Cambria, Grffegoire Winterstein, and Guang-Bin Huang. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Knowledge- Based Systems, 69:45 - 63, 2014.

[95] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3):813{830, 2016.

[96] Felipe Bravo-Mffarquez, Marcelo Mendoza, and Barbara Poblete. Meta-level sentiment models for big social data analysis. Knowledge-Based Systems, 69:86-99, 2014.

[97] Erik Cambria and Amir Hussain. Sentic computing: A common-sense-based framework for concept-level sentiment analysis. Springer International Publishing Switzerland 2015.

[98] Erik Cambria, Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. Big social data analysis. In Big Data Computing. CRC Press - A Chapman & Hall Book. Editor: Rajendra Akerkar, 401-414, 2014.

[99] Erik Cambria, Haixun Wang, and Bebo White. Guest editorial: Big social data analysis.Knowledge-Based Systems, 69:1 - 2, 2014.

[100] Heeryon Cho, Songkuk Kim, Jongseo Lee, and Jong-Seok Lee. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classifcation of product reviews. Knowledge-Based Systems, 71:61-71, 2014.

[101] Sheng Huang, Zhendong Niu, and Chongyang Shi. Automatic construction of domainspeci ffc sentiment lexicon based on constrained label propagation. Knowledge-Based Systems, 56:191: 200, 2014.