ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Review on User Behaviour Analysis using KNN/SVM vide Tweetson Big Data

Pushpa, Gaurav Garg

M.Tech(pursuing), Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayanand University at Rohtak, Haryana, India

Assistant Professor, Dept. of CSE, Advanced Institute of Technology & Management, Palwal, Haryana under the Affiliation of Maharshi Dayan and University, Rohtak, Haryana, India

ABSTRACT: User behavior analysis could be a method throughout that the polarity (i.e. positive, negative or neutral) of a given text is set, generally there area unit which approaches to deal with this problem, particularly, machine learning approach or lexicon primarily based approach, in this scheme we will present the paper deals with behavior analysis using KNN and SVM for tweets.In particular, the SVM and K-Nearest Neighbor classifiers were run on this dataset. The results show that SVM gives the highest precision while KNN (K=10) gives the highest Recall.The ability to use public sentiment in social media is more and more thought-about as a crucial tool for market understanding, client segmentation and stock value prediction for strategic promoting coming up with and maneuvering. This evolution of technology adoption is energized by the healthy growth in huge knowledge framework, that caused applications supported Behavior Analysis (BA) in huge knowledge to become common for businesses. However, scarce works have studied the gaps of militia application in huge knowledge. The contribution of this paper is two-fold: (i) this study reviews the state of the art of various approaches. Together with behavior polarity detection, various options (explicit and implicit), sentiment classification techniques and applications of militia and (ii) this study reviews the suitableness of militia approaches for application within the huge knowledge frameworks, in addition as highlights the gaps and suggests future works that ought to beexplored, various studies are foretold to be distended into approaches that use measurability, possess high ability for supply variation,velocityand veracity to maximize value mining for the benefit of the users, whereas Big Data refers to collection of large datasets containing massive amount of data. Big Data is generated from various sources such as social networking sites like Facebook, Twitter etc. and the data that is generated can be in various formats like structured, semi-structured or unstructured format. Social media monitoring is growing day by day therefore analysis of social data plays a vital role in knowing user behavior. These behavior of users country wise helps in getting information about various current trends and can be used further in deciding usefulness of some tasks, products and themes. In this paper we would be analyzing tweets for user behavior. Tweets are available in JSON format which is to be converted into a structured data. By analyzing all the user social data about a particular topic we would give the output of how users behave for particular topic in certain country and city.

KEYWORDS: K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Machine Learning, HDFS (Hadoop Distributed File System), Big Data.

I. INTRODUCTION

Twitter as a Data Source Twitter is an online micro blogging service that allows users to expose their thoughts in 140 characters. A tweet is short and informal which contains Internet slang words and emoticons. Though, it is easier to analyse tweets, users tend to express their thoughts and opinions straight forward because the length limit. This increases the chances of achieving high Opinion Mining analysis or User Behaviour analysis accuracy.

Big Data: McKinsey Global Institute estimates that consumers around the world stored more than six Exabyte (one Exabyte equal to 1,048,576 terabytes) of new data on devices such as Personal Computers (PCs) and notebooks, which is equivalent to more than 4,000 times the information stored in the U.S. Library of Congress. “Big data” is a term

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

refers to datasets whose size is beyond the ability of typical technologies or software tools to capture, store, manage, and analyse. The term in many instances refers to the use of predictive analytics (like data mining) or other certain advanced analytics methods to extract value from data, and rarely to a particular size of a data set. Big data has its own characteristics that determine the value and potentials of the dataunder the following consideration :

Volume which measures the quantity or the size data, which characterizes whether it can actually be called big data or not.

Variety represents and measures the richness of the content; data may be in forms like structured or non-structured ext, images video, audio, etc.

Velocity measures the speed at the data been generated and processed.As shown in Figure 1, the above characteristics can determine the level of complexity of big data

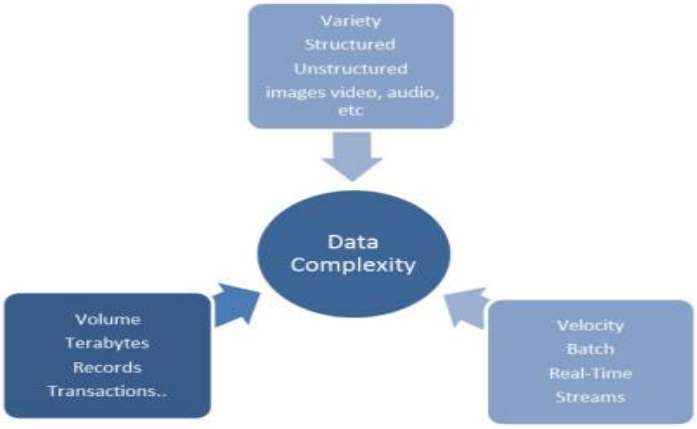

Figure 1 : The above characteristics can determine the level of complexity of big data whereas the above figure depicts the three characteristics of Big Data i.e. Volume, Variety and Velocity.

Hadoop:is being used widely in companies like Yahoo, Facebook, etc. Hadoop is an open source Apache-based framework for reliable, scalable, distributed computing that allows for the distributed processing of large datasets across clusters of computers using a simple programming model called MapReduce. In this work, we use Hadoop as the main component in our architecture Hadoop is a reliable framework, so it has been designed to automatically deal withhardware failures. Hadoop Map-Reduce and HDFS are designed by Google Map-reduce and Google file system.

HDFS: The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework. It is an open source implementation of the Google File System (GFS) that is capable of storing and accessing large-scale data-intensive applications among the Hadoop cluster.

Apache Hive: Apache Hive is a data warehouse software built on top of Hadoop that is capable of analyzing and querying large datasets stored in Hadoop's HDFS using an SQL-like language called HiveQL. By using Hive, you can easily perform complex queries to retrieve information from big data. For the Sentiment Analysis, we developed multiple Hive User Defended Functions (UDF) that are used inside Hive queries for extracting opinions

MapReduce: MapReduce is a programming model introduced by Google in 2004 to support distributed and parallel computing on large data sets on Hadoop clusters. Each MapReduce program is composed of two main functions: map() function, which does the data processing, sorting, and filtering tasks. The second function is reduce(), which will aggregate and summarize all the outputs from the maps. Figure 2 shows how the MapReduce model works; first, the input data gets split into maps. In parallel, the maps will do the processing and produce output in key/pair format to the reducers. The reducers then aggregate the output pairs per each key and write the result into files, figure 2 below

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

depicts the same.

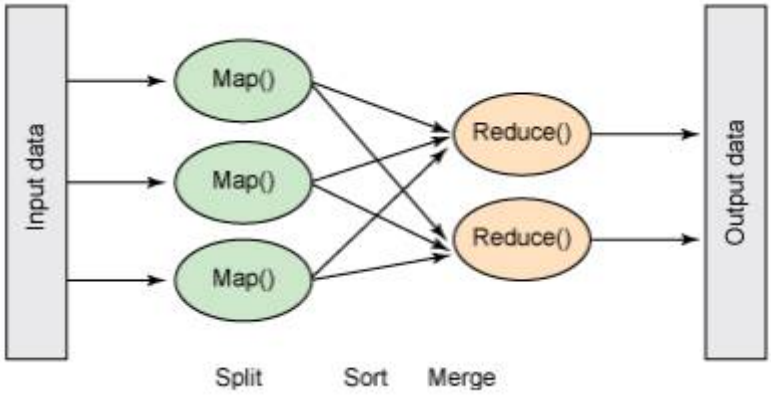

Figure 2: Map-Reduce architecture depicting the Map, Shuffle and Reduce technique with parallel processing

K-nearest neighbour classifier (KNN) This classifier is a simple one which chooses the K number of nearest neighbours in the training documents and classifies an unannotated document based on these K neighbours. Specifically, it calculates the similarity between the unlabeled document and the remaining documents in the training dataset. After that, the labels of the most K similar documents are considered. The final label of the new document isdetermined using majority voting or weighted average of the labels of these K neighbours below figure depicts the algorithm for the same.

k-Nearest Neighbor
Classify ($\mathbf{X}, \mathbf{Y}, x$) // $\mathbf{X}$: training data, $\mathbf{Y}$: class labels of $\mathbf{X}$, x: unknown sample
for $i = 1$ **to** m **do**
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set I containing indices for the k smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$

Figure 3: KNN algorithm to calculate the similarities between more than one document.

Support Vector Machines (SVM) It is an effective traditional text categorization framework. The main idea of SVM is to find the hyper plan, which is represented as a vector that separates document vectors in one class from document vectors in other classes. SVM shows very good performance and higher accuracy in many studies directed towards sentiment analysis in many languages. The work reported in shows that SVM did well with the English language when compared to other classifiers.Although the SVM can be applied to various optimization problems such as regression, the classic problem is that of data classification. The basic idea is shown in figure 1. The data points are identified as being positive or negative, and the problem is to find a hyper-plane that separates the data points by a maximal margin.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

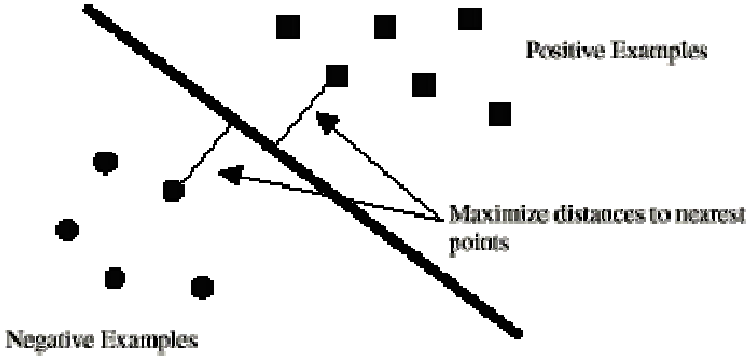

Figure 4: Data Classification using machine learning for sentimental analysis based on positive, negative and neutral context/tweets or datasets.

The above figure only shows the 2-dimensional case where the data points are linearly separable. The mathematics of the problem to be solved is the following:

‘

$$\min_{\vec{w},b} \frac{1}{2}\|w\|,$$

$$s.t \quad y_i = +1 \Rightarrow \vec{w}\cdot\vec{x}_i + b \geq +1$$

(1)

$$y_i = -1 \Rightarrow \vec{w}\cdot\vec{x}_i - b \leq -1$$

$$s.t\ y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1, \quad \forall i$$

The identification of the each data point x_i is y_i, which can take a value of +1 or -1 (representing positive or negative respectively). The solution hyper-plane is the following:

(2)

$$u = \vec{w}\cdot\vec{x} + b$$

The scalar b is also termed the bias.A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

(3)

$$\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha_N} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \quad \forall i$$

The variables α_i are the Lagrangian multipliersfor corresponding data point x_i.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

II. RELATED WORK

Natural Language Processing, Machine Learning, Information Theory and Coding, and Text mining are some of the branches of computer science that are used for sentiment analysis. These approaches, methods and techniques will help us categorize and organize and structure this unstructured data, which is in the form of tweets, into positive, negative or neutral sentiment.

Sentiment analysis can be classified into two types:

1. **Subjectivity/objectivity identification**
2. **Feature/aspect based sentiment analysis**

Machine Learning Techniques: Machine learning techniques can be classified on the basis of

a. **Supervised Machine Learning Techniques:** This basically uses a training data set for categorization of the document or text and has two different algorithms which have achieved great success1. They are as follows :

i. KNN.

ii. Support Vector Machines .

b. **Unsupervised Machine Learning Techniques:** When classification is done without the help of a training data set. Some examples of these techniques are Point wise Mutual Information (PMI) and Semantic Orientation.

Text Mining Techniques : Text mining process has four stages:

a. Texts Collection
b. Pre-processing
c. Analysis
d. Validation

Natural Language Processing: The techniques or tasks of Natural Language Processing play a major role in Sentiment analysis. The different tasks like Part Of Speech tagging, Speech Recognition, N-gram algorithms, Markov model, sentiment lexicon acquisition and parsing techniques can express opinion on document level, sentence level and aspect.levels.

Hybrid Approaches To perform the sentiment analysis according to our needs, we can use a combination of any of the above approaches : combination of any of the two or more techniques mentioned above can be used for more accurate results for explicit and implicit sentiment analysis. For identifying Twitter messages, we use SVM and N-gram algorithms. Generation of an implicit opinion for proper semantic orientation can be done with the combination of NLP and Machine Learning techniques with semantic approach.Combination of any of the NLP techniques with/without semantic approach, machine learning techniques can be made for generation of proper semantic orientation as and when needed for analyses of objective sentences that carry sentiment

III. LITERATURE SURVEY

For example Hassan, Yulan, and Alani [1] studied Semitic sentiment analysis of Twitter. The authors used three different Twitter datasets for their experiments. They proposed the using of Semitic features in Twitter sentiment classification and explored three different approaches for incorporating them into the analysis with replacement, augmentation, and interpolation. In Researchers have proposed many different approaches for sentiment analysis. In general, there aretwo main methods, the first one is using machine learning techniques or supervised techniques which arepresented in this paper, and the other one is unsupervised techniques. Many studies have focused on thesentiment analysis for the English language and other Indo-European languages

Pang and Lee [6] usedmachine learning techniques for sentiment classification. They employed three classifiers (Naïve Bayes,Maximum Entropy classification, Support Vector machine). Their data source was the Internet MovieDatabase (IMDB); they selected only reviews where the author rating was expressed either with stars orsome numeral value.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Dave, Lawrence, and Pen nock [10] proposed an approach, which begins with training a classifier using a corpus of self-tagged reviews available from major web sites. They decided to use n-grams on two tests and the result showed that this way is better than traditional machine learning. Many researches were introduced to analyze sentiment and extracting opinions from the World Wide Web. This proved to be important due to the large amount of data contributed by users in websites such as social networks (Facebook, Twitter, etc.).

There are few studies for sentiment analysis for the Arabic language. For example, Abdul Majeed and Diab presented a newly developed manually annotated corpus of Modern Standard Arabic (MSA) together with a new polarity lexicon [2]. They ran their experiments on three different pre-processing settings based on tokenized text from the Penn Arabic Treebank (PATB). They adopted two-stage classification approach, in the first stage they built a binary classifier to sort out objective from subjective cases. For the second stage, they applied binary classification that distinguishes positive from negative cases.

In [4], the same researchers in [2] reported efforts to bridge the gap between Arabic researches by presenting AWATIF; a multi-genre corpus for Modern Standard Arabic for Subjectivity and Sentiment Analysis (MSA SSA). They extend their previous work by showing how annotation studies within subjectivity and sentiment analysis can both be inspired by existing linguistic theory and cater for genre nuances. Alams, and Ahmed [3] target three languages (English, Arabic, and Urdu) in their work. They described a method for automatically extracting specialist terms called local grammar. The authors compared the behavior of single and compound tokens in specialist and general language corpora todetermine whether a token is behaving like a sentiment term or not. Elhawary and Elfeky [5] extract business reviews scattered on the web written in the Arabic language. They built a system that comprises two components: a reviews classifier that classifies any web page whether it contains reviews or not, and sentiment analyzer that identifies the reviews' text if it (positive, negative, neutral or mixed).

IV. PROPOSED WORK

1. To Extract twitter real time data using Twitter4J.
2. To study existing techniques for user behaviour like SVM and KNN
3. To predict behaviour of user tweets whether neutral, positive or negative in context of certain topic.
4. To predict the general behaviour of users of particular location according to city or country wise and in context of a particular topic.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

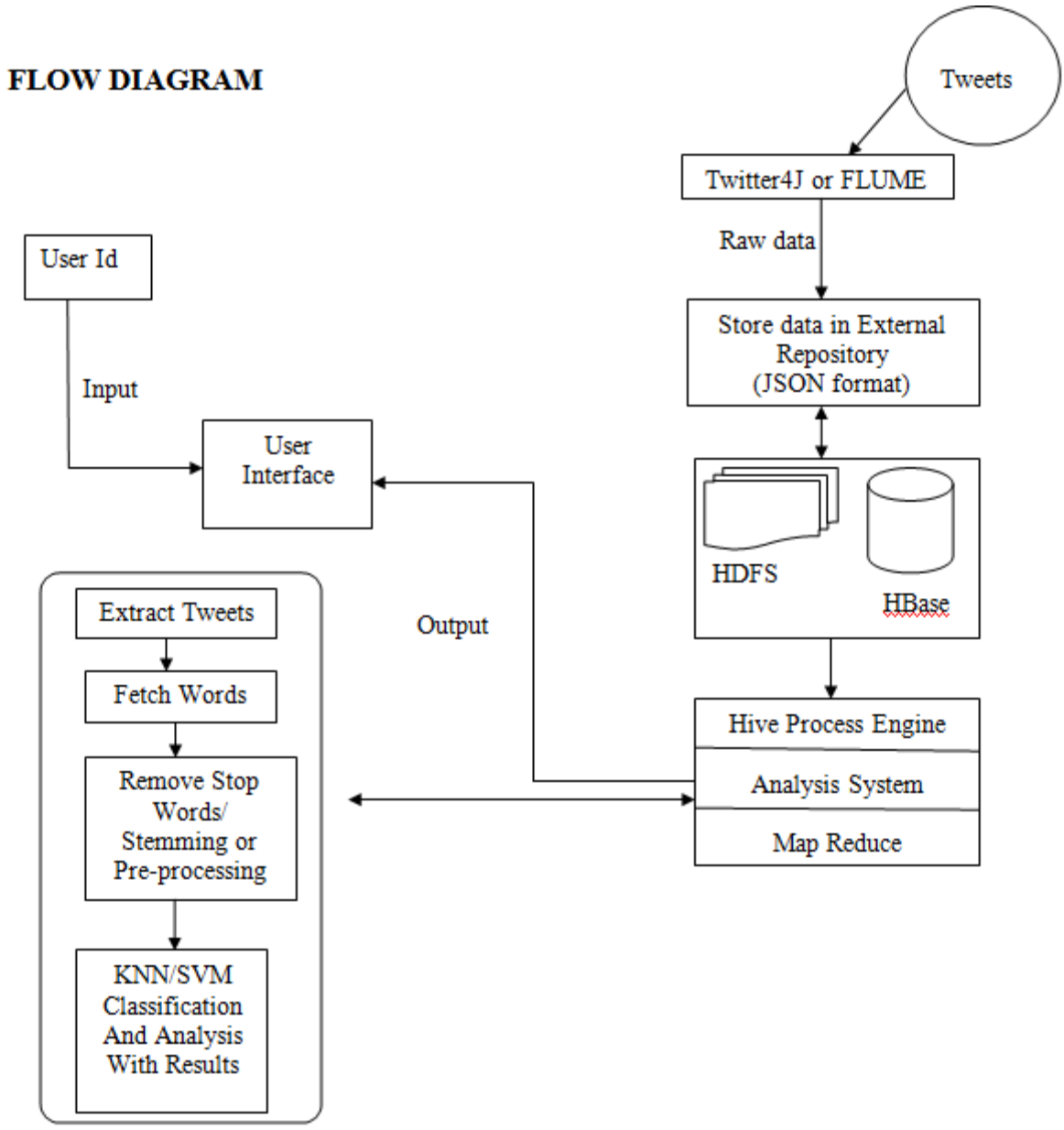

Figure 5: Flow Diagram of Proposed System

REFERENCES

1 Hassan. S., Yulan.H., and Alani. H., "Semantic sentiment analysis of Twitter."The Semantic Web–ISWC.Springer, pp. 508-524, 2012.
2 Abdul-Mageed. M., Diab.M., and Korayem M., "Subjectivity and sentiment analysis of modernstandard Arabic." Proceedings of the 49th Annual Meeting of the Association for ComputationalLinguistics: Human Language Technologies. Vol. 2. 2011.
3 Almas Y., and Ahmad K., "A note on extracting sentiments in financial news in English, Arabic &Urdu." The Second Workshop on Computational Approaches to Arabic Script-based Languages.2007.
4 Abdul-Mageed M., and Diab M., "AWATIF: A multi-genre corpus for Modern Standard Arabicsubjectivity and sentiment analysis." Proceedings of LREC, Istanbul, Turkey, 2012.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

5 Elhawary M. and Elfeky M., "Mining Arabic Business Reviews." Data Mining Workshops(ICDMW), P. 1108-1113, 2010.
6 Pang B., and Lee L. "A sentimental education: Sentiment analysis using subjectivity summarizationbased on minimum cuts." Proceedings of the 42nd annual meeting on Association for ComputationalLinguistics. 2004.
7 El-Halees A., "Arabic Opinion Mining Using Combined Classification Approach." Proceedings of theInternational Arab Conference on Information Technology, ACIT. 2011.
8 Rushdi-Saleh M., Martín-Valdivia M., Ureña-López L. &Perea-Ortega J.M., Bilingual Experimentswith an Arabic-English Corpus for Opinion Mining, 2011.
9 Al-Subaihin A., Al-Khalifa H., and Al-Salman A.M., "A proposed sentiment analysis tool formodern arabic using human-based computing." Proceedings of the 13th International Conference onInformation Integration and Web-based Applications and Services.ACM, 2011.
10 Dave K., Lawrence S., &Pennock D.M., "Mining the peanut gallery: Opinion extraction andsemantic classification of product reviews." In Proceedings of the 12th international conference onWorld Wide Web, pp. 519-528.ACM, 2003.
11 Nasukawa T., and Jeonghee Y., "Sentiment analysis: Capturing favorability using natural languageprocessing." Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003.
12 Self reference to the authors, names were removed as per Journal instructions] " SentimentAnalysis." June 2012.
13 Rao D., and Ravichandran D., "Semi-supervised polarity lexicon induction." Proceedings of the 12thConference of the European Chapter of the Association for Computational Linguistics.Associationfor Computational Linguistics, 2009.
14 Dasgupta S., and Ng V., "Mine the easy, classify the hard: a semi-supervised approach to automaticsentiment classification." In Proceedings of the Joint Conference of the 47th Annual Meeting of theACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP:Volume 2, pp. 701-709. Association for Computational Linguistics, 2009.
15 Sindhwani V., and Melville P., "Document-word co-regularization for semi-supervised sentimentanalysis." 8th IEEE International Conference on Data Mining (ICDM'08), pp. 1025-1030, 2008.
16 Rapidminer, http://rapid-i.com/, last access on 31-Jan-201.
17 Goldberg, Anderwo B., and Zhu X., "Seeing stars when there aren't many stars: graph-based semisupervisedlearning for sentiment categorization." Proceedings of the First Workshop on Graph BasedMethods for Natural Language Processing. Association for Computational Linguistics, 2006.
18 Kumar A., and Sebastian T.M., "Sentiment Analysis on Twitter." IJCSI International Journal ofComputer Science, Issue 9.3, pp. 372-378, 2102.
19 Malouf R, and Mullen.T. "Taking sides: User classification for informal online politicaldiscourse." Internet Research 18.2: pp. 177-190, 2008.
20 Glance N., Hurst M., Nigam K., Siegler M., Stockton R., &TomokiyoT.,"Deriving marketingintelligence from online discussion. In Proceedings of the 11th ACM SIGKDD internationalconference on knowledge discovery in data mining, pp. 419-428, 2005.