Dynamic Sentiment Analysis Using Multiple Machine Learning Algorithms: A Comparative Knowledge Methodology

Manmeet Kaur, Krishna Kant Agrawal and Deepak Arora

Abstract Human can easily understand or interpret the meaning of language. However, a machine has no natural language to deduce the hidden emotions. Without knowing the context of the word, it cannot simply infer whether a piece of text conveys joy, anger or frustration. Here, sentiment analysis came into picture. Sentiment analysis is the analysis of feelings, attitude and opinions of human emotions extracted from text. It uses natural language processing (NLP) for classifying the text into positive, negative or neutral category. Many businesses nowadays take feedback of the product from the customers to improve the quality or service of the product. Earlier feedbacks were taken by the call center executives but today a vast amount of data is available on the Internet. People share their views regarding products, services, people, etc. Sentiment analysis makes the task easier by extracting the relevant words from the sentences and classifying it in different categories. In this paper, we have described the essential steps used in the process of the sentiment analysis and few fields that work under its umbrella. A comparative analysis of machine learning algorithm like Naive Bayes, SVM, maximum entropy is done along with the few algorithms like artificial neural network and K-nearest neighbor, which can be used in sentiment analysis.

Keywords Machine learning · Support vector machine · Naive Bayes
Maximum entropy · Sentiment classification · Building resource
Transfer learning · Emotion detection · Chi square · Information gain

M. Kaur (✉) · K. K. Agrawal · D. Arora
Department of Computer Science and Engineering, Amity University, Noida, Uttar Pradesh, India
e-mail: meetsoni2006@yahoo.co.in

K. K. Agrawal
e-mail: kkagrawal@outlook.com

D. Arora
e-mail: darora@lko.amity.edu

M. L. Kolhe et al. (eds.), *Advances in Data and Information Sciences*, Lecture Notes in Networks and Systems 38, https://doi.org/10.1007/978-981-10-8360-0_26

1 Introduction

The Internet today has become the vast sea of information and knowledge (a source of information). A large-scale enterprise to a small-scale firm uses the Internet for their business. People provide feedback of the products, rate them, share feelings about the politics and give review about the movie. It is providing a platform where everything is presented at one place. These data are useful for firms in order to improve the quality of the product to enhance their business. An automated machine learning system operates on this abundance data to perform sentiment analysis which eliminates the manual work of human by classifying the text in positive, negative or neutral category.

As shown in Fig. 1, sentiment analysis can be done using lexicon-based approach, machine learning approach and hybrid approach which have been described in next section in detail. The most used approach is machine learning in which the classification of text can be done using many supervised and unsupervised algorithms. Supervised learning makes use of the training documents or corpus, e.g., Naïve Bayes, SVM, maximum entropy. Unsupervised learning is used when no such labeled training documents are available, e.g., neighbor classification.

2 Background

A lot of work has been done in the field of sentiment analysis and its related field. Sentiment extraction can be done through different approaches, namely lexicon-based approach, machine learning approach, hybrid/combined approach [1].

2.1 Lexical-Based Approach

In the lexical-based approach, the text input is separated into several words, which is called tokenization. Once the word is tokenized, the numbers of positive, negative and neutral words are counted which are kept in the library of the words called Bag of Words (BoW). Further to classify the sentiment values of the document, lexicon is used.

In the process of classifying the sentiments, the system takes the count of the word which is being used in the statement and measures the value of goodness or badness of each of them, thus summing up and deciding on the sentiment of the text as a whole. This technique disregards the order of the words; e.g., it is not a good movie, is a negative sentence, but in such approach, this statement can be misinterpreted as a positive one. Lexical analysis has a drawback, with the exponential growth of the size of the dictionary (number of words); its performance (in terms of time complexity and accuracy) degrades drastically [1].

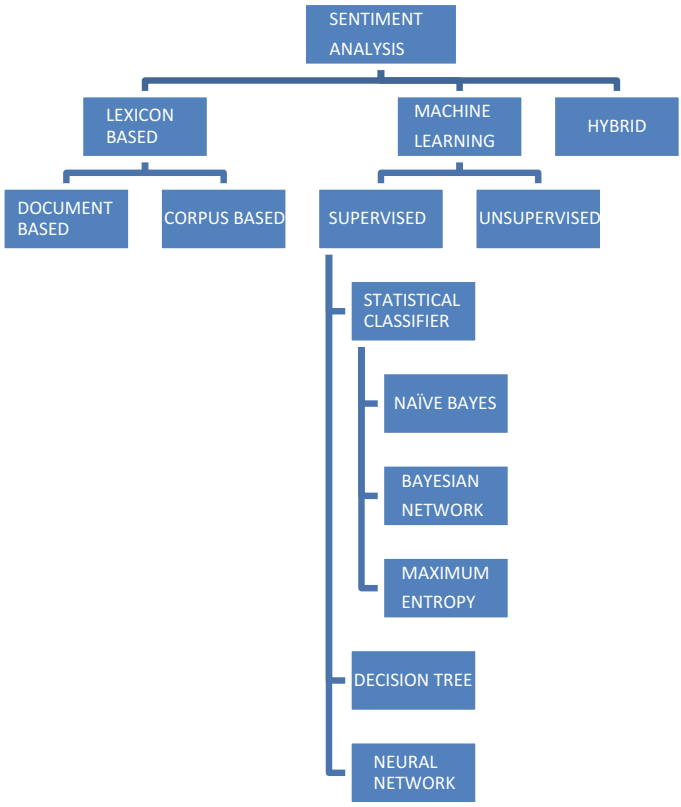

Fig. 1 Classification of sentiment analysis

2.2 Machine Learning Approach

Due to the adaptability and accuracy of this technique, it is one of the most prominent methods which is gaining attention in today's world. This classification of text can be done with the help of many supervised and unsupervised machine algorithms. Supervised learning makes use of the training documents or corpus, e.g., Naive Bayes, SVM, maximum entropy. Unsupervised learning is used when no such labeled training documents are available, e.g., neighbor classification. Artificial neural network-based algorithms are also used for the classification.

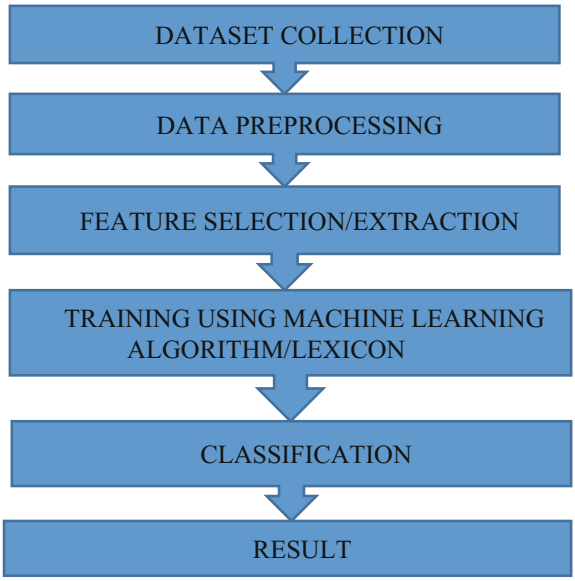

Fig. 2 Steps involved in sentiment analysis

2.3 Hybrid Approach

With the advancement in the field of sentiment analysis, researchers were allured to find the hybrid approach which could exhibit the speed of lexical approach and the accuracy of a machine learning approach [1]. In a study, the authors collected the Chinese and English text from twitter [2]. Translate the Chinese text into English using Yahoo Babelfish and Google translator. They used lexicon approach (tokenization for creating Bag of Words) and unlabeled data for dividing the corpus of text into two categories—positive and negative. The training dataset obtained from it was further used for training purpose in Naive Bayes classifier for sentiment classification.

3 Methodology

In this section, the steps involved in the process of sentiment analysis have been described. Figure 2 depicts these steps as described below.

3.1 Dataset Collection

The first step is to collect the data from the various sources. The researchers have widely used the data from social networking and micro-blogging sites like Twitter as

the dataset could be collected from it with the minimal supervision effort [2]. The data from IMDB (for movie reviews) and amazon.com (for customer's opinion/feedback) have been a great source of attraction for the researchers as these Web sites provide real-life datasets.

3.2 Data Preprocessing

This step is taken to clean the data and eliminate the noise. This step requires the elimination of stop words. Stop words are those words that do not carry any sentiments, e.g., of, have, been, to. Stemming is also done at this step in which the word is reduced to its root. For example, checking, checkers are the stemming words, and their root word is check. In this step, the data transformation can also be done if required.

3.3 Feature Selection/Extraction

This step in sentiment classification is used to extract features, i.e., text features. This could be finding adjectives, phrases which contribute to the importance of sentiments. Few features have been described in this section.

Text presence and frequency. These features could separate words or n-gram word and the frequency of occurrence of the words [3]. The selection of correct featured word is very important in sentiment analysis as the twisted words play a vital role in changing the essence of the statement. 'Not bad' holds a positive sentiment; if only bad is taken, then the sentence will be misinterpreted as the negative one. The use of *n*-gram helps in increasing the accuracy of the classification. The frequency of the words contributes to the importance of the feature [3].

Opinion words. These are the words that emphasize on individual opinions, whether bad or good, like or dislike. Some phrases implicitly express the opinion as well, for example cost a pretty penny.

Negation. The presence of negative words may change the essence of the statement. For example, it was not a pleasant day, holds a negative sentiment, although pleasant holds a positive sentiment.

Feature selection is a crucial step in sentiment analysis. Few algorithms used by researchers for feature selection are described below.

Information Gain. Information gain is used as an attribute selection method [4]. The motive is to select the attribute that is most useful in classification of examples. It measures how well a given tuple separates the training example according to its target classification. The attribute with the highest information gain, also known as entropy, is selected for the classification. It is used in decision making to select an attribute which is used in the classification of node of the tree. It assures the purity of the partition.

The information needed to classify the attribute is given by

$$\inf(D) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

where inf(D) is the actual information, pi is the probability of the attribute.

$$\inf{}_a(D) = h \sum_{j=1}^{v} \frac{|D_j|}{|D|} * \inf o\left(D_j\right) \tag{2}$$

$$\text{Gain}(a) = \inf o(D) - \inf o_a(D) \tag{3}$$

$\inf_a(D)$ is the expected information needed to classify a tuple, and Gain(a) is the information gain.

Chi Square (λ^2). Chi square finds the deviation between the observed count and expected count.

$$\lambda^2(a, b) = \frac{N(wz - yx)^2}{(w + y)(x + z)(w + x)(y + z)} \tag{4}$$

where w, x, y, z represent the frequencies that tell the absence or presence of the feature in a sample dataset [5]. W represents count of sample in which feature a and class b occurred together. It actually represents 'a' as the feature and 'b' as the class.

Other methods used for the feature selection are Gini index that measures the impurity of a set of data tuple, hidden Markov model (HMM), PMI, LDA, etc.

After the feature extraction, sentiment analysis is implemented through lexicon-based approach or machine learning approach which classifies the text into positive, negative or neutral category. In the next section, various machine learning algorithms have been discussed.

4 Classification Methods

In this section, the various supervised and unsupervised machine learning algorithms used for the purpose of classification of text in sentiment analysis have been discussed.

4.1 Naive Bayes

Bayesian classifiers are statistical classifiers. They can predict the class membership probability that the tuple belongs to the specific class. It assumes that the effect of values of an attribute of given class is independent of the values of other attributes.

This is called conditional independence. This works on the Bag of Word feature extraction method and makes independent assumption about the occurrence of words. For a given feature 'f' and class category 'c,' the conditional probability that 'f' belongs to 'c' can be formulated as:

$$P(c/f) = \frac{p(f/c) * p(c)}{p(f)} \tag{5}$$

$P(f)$ is the prior probability of the feature that it has occurred [3]. $P(f/c)$ is the prior probability that the given feature is being classified as the class label. $P(c/f)$ is the posterior probability that the given feature belongs to a particular class. Given Naive assumption, the features will be considered as independent of each other. The equation can be rewritten as

$$P(c/f) = \frac{p(c) * p(f1/c) * p(f2/c) * \cdots * p(fn/c)}{p(f)} \tag{6}$$

The main advantage of the Naive Bayes classifier is that it requires small amount of training data that calculates the parameters which is used for prediction [6].

4.2 Support Vector Machine (SVM)

Support vector machine is a non-probabilistic binary linear classifier that works on both linear and nonlinear data. Each review is represented in vectorized form which shows a data point in space [6]. It finds the best hyperplane, $\underset{w}{\rightarrow}$, to separate the textual data vector having maximum marginal distance. Once the training of the model is done, the testing data are mapped into some space, which is used to predict on which side of the hyperplane the data fall [6].

Let $c_i \in \{1, -1\}$ represent the positive and negative classes, respectively, for a document $\underset{d_i}{\rightarrow}$, and the equation for $\underset{w}{\rightarrow}$ is given by

$$\vec{w} = \sum_i \alpha_i c_i d_i, \quad \alpha_i \geq 0 \tag{7}$$

All the $\underset{d_i}{\rightarrow}$ such that $\alpha_i > 0$ are said to be support vectors [7].

Tripathi et al. used labeled polarity movie dataset having 1000 negative and 1000 positive reviews [6]. Data cleaning is done to extract the relevant features. The vectorization technique was used to convert the word data into numeric format. A matrix was created using vectorization, where each row and column represented an individual review and a feature, respectively. K-fold cross-validation technique was applied on the matrix to select the training and testing dataset for each fold [6]. Table 1 represents the comparison between the existing literatures.

Table 1 Comparison based on the existing literature

	Pang and Lee	Read	Tripathi et al.
Naive Bayes	0.864	0.789	0.895
SVM	0.8615	0.815	0.940

Pang and Lee as well as Tripathi et al. in their paper used ten fold cross-validation to perform classification, while Read used only three folds [6]. It showed that the higher number of folds results in more generalized result and SVM gives more accurate result than Naive Bayes.

4.3 Maximum Entropy Classifier

Maximum entropy is a common technique for the estimation of the probability distributions from data [8]. The principal behind maximum entropy says that the correct distribution is when the constraints set by the 'evidence' are met even when the entropy/uncertainty is maximized. Constraints are expected values of the features.

Maximum entropy classifier is a machine learning technique which is based on empirical data. Nigam et al. and Berger et al. showed that in many cases it outperforms Naive Bayes classification [9]. Raychaudhari et al. also found that maximum entropy worked better than Naive Bayes and nearest neighbor classification for their classification [10]. Unlike Naive Bayes, it does not make independent assumption about the occurrence of words. The mathematical formula for entropy is given by

$$H = \sum p(c, d) \log p(c, d) \tag{8}$$

where $p(c, d)$ is the probability that the document 'd' belongs to the specific category 'c.'

Kamal et al. showed in their research that maximum entropy outperforms the regular Naive Bayes as the classification error is reduced by more than 40% [8]. They experimented on webKB dataset. But in comparison with scaled Naive Bayes, the results were mixed. In scaled Naive Bayes, each document is scaled such that it contains constant number of word count.

In [8], the author implemented maximum entropy technique in three datasets out of which two performed better than Naive Bayes. Basic maximum entropy suffers from over-fitting. They used Gaussian prior to reduce the over-fitting problem resulted in a better performance. They showed that the feature selection is an essential factor for maximum entropy.

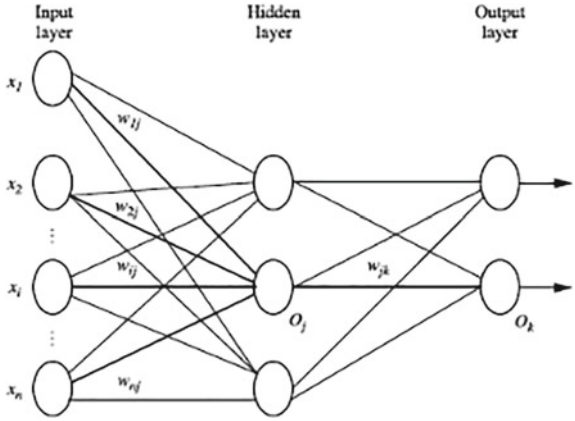

Fig. 3 Multilayer feed-forward neural network [4]

4.4 Artificial Neural Network

Artificial neural network is composed of artificial neurons, which are its basic unit. It consists of input layer, hidden layer and output layer [4]. This network consists of connected input/output units. There is a weight associated with every neuron. The units associated with an input layer are called input units. The weighted sums of the given inputs are added to the activation function that gives an output which acts as an input to the hidden layer. This is a feed-forward network as the output to one layer acts as an input to the other. This network has poor interpretability (Fig. 3).

4.5 Backpropagation Neural Network (BNN)

Backpropagation neural network is a feed-forward neural network in which the error rate is reduced through backpropagation as the actual output is matched with the expected output and then error is propagated backward by updating the weights and the bias (which helps in varying the activity) [4].

The word frequencies to the *j*th document act as an input to the neuron. The weight W is associated with each and every neuron. The linear function associated with it is given as $I = W.X_j$. For the binary classification, the sign predicts the class label. For the nonlinear classification, multilayer neural network is used. The complexity of the training dataset increases as the size of the middle layer increases because the error is backpropagated over the different layers. Smaller hidden layer tends to produce better classification of the class label.

There is a comparison between SVM and neural network [11]. They showed that ANN outperformed SVM when the experiment is conducted on the movie reviews. There is a limitation associated with both the techniques like the computational cost of ANN at the training time and SVM at the running time.

Steps for the backpropagation algorithm:

S1: Take the input from the input node.

S2: First input from the input node is equal to the output for this node which acts as the input to this network.

$$O_i = I_i \tag{9}$$

S3: Input to the hidden layer or output layer is the weighted sum from the previous nodes added by bias (where B_j is the bias), i.e.,

$$I_j = \sum_i W_{ij} O_i + B_j \tag{10}$$

S4: Output of the hidden layer or output layer can be found using activation function that could be sigmoid or logistic function.

$$O_j = \frac{1}{1 + e^{-I_j}} \tag{11}$$

S5: The errors are backpropagated for all the unit j in the output layer.

$$E_k = O_j \left(1 - O_j\right) \left(Y - O_j\right) \tag{12}$$

where Y is the expected output

S6: The errors are backpropagated from last hidden layer to the first hidden layer for all the unit j.

$$E_k = O_j \left(1 - O_j\right) \sum_k E_k w_{jk} \tag{13}$$

S7: weights are updated.
$\Delta w_{ij} = E_j O_i(l)$, l is the learning rate, $0 < l < 1$, and l is used to avoid local minimum and encourages to reach global maxima in a decision surface.

$$W_{ij} = \Delta W_{ij} + W_{ij} \tag{14}$$

S8: For updating bias (Fig. 4)

$$\Delta B_j = E_j(l) \tag{15}$$

$$B_j = \Delta B_j + B_j \tag{16}$$

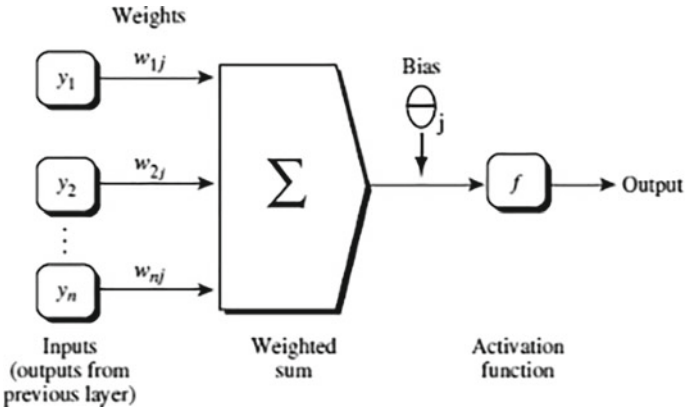

Fig. 4 Weighted sum of the output of the previous layer is added with a bias which is followed by an activation function that gives the output [4]

4.6 Semi- and Unsupervised Learning

The idea behind the classification of text is to divide it into categories. This requires labeled training data or document in supervised leaning. However, it is often difficult to get the labeled data, but the unlabeled data or document can be obtained easily. Here, the unsupervised learning can be used that does not require labeled training data. Zhou et al. worked on a strategy at the level of the features instead of instances by providing weak supervision [12].

The prior information was incorporated into sentiment classifier model, obtained as an initial classifier, extracted from an existing sentiment lexicon. They worked on the data obtained from IMDB for movie reviews and amazon.com. Their work identified that the polarity of the words may vary from domain to domain. They showed for any text classification, the approach used by them obtained better performance than weakly supervised learning techniques if the little bit relevant prior knowledge is available.

4.7 PCA

Principal component analysis (PCA) is an unsupervised learning technique. Its motive is to find principal component (PC). The data are expressed in a manner that highlights their differences and similarities. This technique emphasizes on the variations and develops strong patterns in dataset. It helps to eliminate dimensions. PCA takes a dataset with a lot of dimensionality and flattens it to 2D or 3D so we can look at it. It tries to find a meaningful way to flatten the data by focusing on

things that are different from words or sentences in a document. PCA finds direction in which it is orthogonal and with maximum variance.

4.8 K-*Nearest Neighbor (KNN)*

K-nearest neighbor classifiers are the lazy learners. When an unknown tuple is given, then KNN searches the pattern space which defines the closeness to the unknown attribute. The closeness can be found using Euclidean distance in terms of distance metrics. *K* number of tuples are chosen, and the other attributes are computed in terms of the distance. *K* defines the number of separate clusters. The unknown tuple is assumed to belong to the cluster which has the minimum distance with that cluster.

5 Applications

Sentiment analysis can be implemented in various fields that have attracted many researchers. Some of these fields have been discussed in this section.

5.1 Emotion Detection

Natural language processing can be used to implement sentiment analysis which helps in obtaining opinion about an entity. Opinion defines an attitude towards an object. Sentiments define the feelings. Emotion reflects attitude. Emotions can be joy, agony, sadness, frustration, disgust, anticipation and surprise. Sentiment analysis defines positive, negative or neutral opinion, whereas emotion detection defines various emotions. According to Plutchik, emotion detection is considered as sentiment analysis task [13].

Lu and Lin proposed an approach for detecting emotions on different events based on Web-based text mining [3]. This approach was formed on the probability distribution to find common mutual actions between the object and subject of an event.

Balahur et al. used lexicon-based approach and ML both [3]. They used the approach of common sense. They said that emotions cannot always be detected explicitly, i.e., by defining a sad word to express the emotion of grief or a happy word to express joy. SVM and SVM-SO approaches were used to achieve the goal. They showed that the use of EmotiNet (corpus storing the knowledge based on common sense) gives better results than supervised learning techniques and corpus-based approach.

5.2 Building Resource

Building resource focuses on creating dictionary, corpora and lexica in which expressions giving opinions are annotated in accordance with their polarity.

Robaldo et al. introduced building corpus in which they used opinion mining—ML- and XML-based tagging of textual expressions that convey the relevant opinion [14]. A standard methodology was used by them that annotated relevant statements in the text which was independent of any application domain. Then, domain-specific adaptation was considered that depended on the ontology used which is domain dependent. They used dataset on restaurant reviews, and query-oriented extraction process was used. The result showed that the annotation scheme was able to cover the large complexity along with the preservation of good agreement between different people.

5.3 Transfer Learning

Transfer learning obtains the knowledge from the auxiliary domain which is used to improve the learning phase in the target domain, for instance a search in Hindi to English. It is a new cross-domain learning method as it shows the several aspects of domain differences.

In SA, it can be used to build bridge between two domains as it transfers classification of sentiments from one domain to another.

6 Conclusion and Future Scope

After analyzing the various articles, it is observed that the feature selection method is a crucial step in sentiment analysis. Machine learning is the most popular approach used for the classification of sentiment analysis as the lexical approach suffers in terms of performance when the size of the dictionary increases. Many researchers have proved that SVM works better than Naive Bayes or other approaches for the sentiment classification. Yet every technique has few drawbacks and cannot resolve all the challenges.

Artificial neural network is a significant method for the classification which is gaining popularity due to its adaptive nature of adjusting themselves according to the data without explicitly specifying the functional form of the underlying project. The related fields, emotion detection, transfer learning and building resources are the emerging fields of research.

One of the biggest challenges related to the sentiment analysis is language. Interest in other languages than English is increasing and it requires the ability to analyze the

emotions independent of languages. Google translator is the best tool for translating a language into English.

For the future, we would like to work on artificial neural network for the sentiment classification and analysis on dataset in other languages like Hindi or Punjabi along with the English language. Work in the field of emotion detection is also an open area of research as the study suggests that a positive word can be a sarcasm which could lead to unsuccessful classification.

References

1. Thakkar H, Patel D (2015) Approaches for sentiment analysis on twitter: a state-of-art study, 1–8
2. Pak A, Paroubek P, Paris-sud D, Limsi-cnrs L, Cedex FO (2010) Twitter based system: using twitter for disambiguating sentiment ambiguous adjectives. Comput Linguist, 436–439
3. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5:1093–1113
4. Jiawei H, Kamber M (2001) Data mining: concepts and techniques
5. Shahana PH, Omman B (2015) Evaluation of features on sentimental analysis. Procedia Comput Sci 46:1585–1592
6. Tripathy A, Agrawal A, Rath SK (2015) Classification of sentimental reviews using machine learning techniques. Procedia Comput Sci 821–829
7. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up: sentiment classification using machine learning techniques. Proc Conf Empir Methods Nat Lang Process, 79–86
8. Nigam K, Lafferty J, McCallum A (1999) Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering, pp 61–67
9. Mehra N, Khandelwal S, Patel P (2002) Sentiment identification using maximum entropy analysis of movie reviews
10. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res 12:203–214
11. Moraes R, Valiati JF, Gavião Neto WP (2013) Document-level sentiment classification: an empirical comparison between SVM and ANN
12. Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach. Decis Support Syst 50:511–521
13. Plutchik R (1980) A general psychoevolutionary theory of emotion. Emot Theor Res Exp 1:3–33
14. Robaldo L, Di Caro L (2013) OpinionMining-ML. Comput Stand Interfaces 35:454–469