iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** III **Month of publication:** March 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

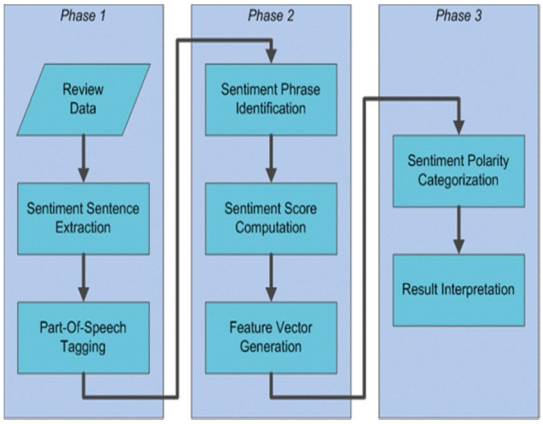
Survey- Algorithms Used For Sentiment Analysis

Mohan.I[1], Sangavi.D[2], Priyanka.K[3], Ramya.P[4]
[1]Assistant Professor, [2]Student, Information Technology
Prathyusha Engineering College, India.

Abstract--Sentiment Analysis (SA), a subfield of NLP, is the computational handling of opinions, sentiments, feedback and subjectivity of text . It is based on Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents. Algorithms used for this purpose in recent times and a variety of Sentiment Analysis applications are investigated and obtainable briefly in this survey.The main target of this survey is to give nearly full image of SA techniques and the related fields with brief details.
Keywords: Sentiment analysis; Sentiment classification; Feature selection; Emotion detection; Transfer learning; Building resources; Text Mining.

I. INTRODUCTION

Sentimental Analysis is all about getting the real voice of people towards specific product, movies, organization, news, events, services, issues and their attributes. These topics are most likely to be covered by reviews or feedback. The difference between Opinion Mining and Sentiment Analysis is that, Opinion Mining extracts and analyzes people's view about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to identify the sentiments they convey, and then classify their polarity.



There are three main classification levels in SA:
Document-Level
Sentence-Level
Aspect-Level

A. Document-Level

SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the complete article a basic information unit (talking about one topic).

B. Sentence-Level

SA aims to classify emotion expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will decide whether the sentence expresses positive or negative opinions. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents .

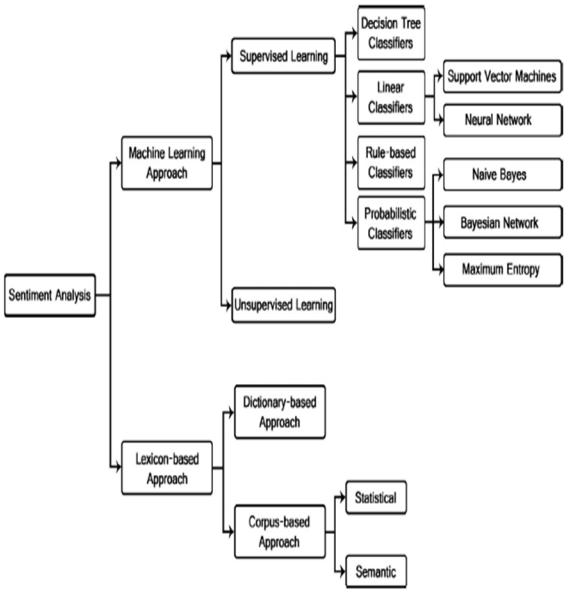
Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Classifying text at the document level or at the sentence level does not provide the essential detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level.

C. Aspect-Level

SA aims to classify the sentiment with respect to the precise aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity like this sentence ‘‘The display size of the camera is small, but the battery life is long’’. This survey tackles the first two kinds of SA. The datasets used in Sentiment Analysis are an important issue in this field for instance, product reviews. These reviews are important to the business holders as they can take business decisions according to the analysis results of customer’s opinions about their products. The reviews sources are collected from the review websites. The social network sites and micro-blogging sites are considered a very good source of information because people share and discuss their opinions about a certain topic freely. They are also used as data sources in the SA process. There are many applications and enhancements on SA algorithms that were proposed in the last few years. This survey aims to give a closer look on these enhancements and to summarize and categorize some articles presented in this field according to the various SA techniques.

This survey can be useful for new-comer researchers in this field as it covers the most famous SA techniques and applications in this paper. This survey uniquely gives a refined categorization to the various SA techniques which is not found in other surveys. It discusses also new related fields in SA which have attracted the researchers lately and their corresponding articles. These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences.



II. METHODOLOGY

The objectives of the articles are illustrated in the third column. They are divided into six categories which are (SA, ED, SC, FS, TL and BR). The BR category can be classified to lexica, Corpora or dictionaries. The authors categorized the articles that solve the Sentiment classification problem as SC. Other articles that solve a general Sentiment Analysis problem are categorized as SA. The articles that give contribution in the feature selection phase are categorized as FS. Then the authors categorized the articles that represent the SA related fields like Emotion Detection (ED), Building Resource (BR) and Transfer Learning (TL). The fourth column specifies whether the article is domain-oriented by means of Yes/No answers (Y or N). Domain-oriented means that domain-specific data are used in the SA process. The fifth column shows the algorithms used, and specifies their categories. Some articles use different algorithms other than the SC techniques which are presented in Section 4. This applies, for example, to the work presented by Steinberger. In this case, the algorithm name only is written. The sixth column specifies whether the article uses SA techniques for general Analysis of Text (G) or solves the problem of binary classification (Positive/Negative). The seventh

column illustrates the scope of the data used for evaluating the article's algorithms. The data could be reviews, news articles, web pages, micro-blogs and others. The eighth column specifies the benchmark data set or the well-known data source used if available; as some articles do not give that information. This could help the reader if he is interested in a certain scope of data. The last column specifies if any other languages other than English are analyzed in the article. The survey methodology is as follows: brief explanation to the famous FS and SC algorithms representing some related fields to SA are discussed. Then the contribution of these articles to these algorithms is presented illustrating how they use these algorithms to solve special problems in SA. The main target of this survey is to present a unique categorization for these SA related articles.

III. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment Classification techniques can be generally divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features which can be generally divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labeled training documents. The unsupervised methods are used when it is complex to find these labeled training documents. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. The hybrid Approach combines both approaches.

A. Machine Learning Approach

Machine learning approach relies on the ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features.

1) *Text Classification Problem Definition:* We have a set of training records D = {X1, X2, ..., Xn} where each record is 1098 W labeled to a class. The classification model is related to the features in the fundamental record to one of the class labels. Then, for an unknown class, the model will predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

Table 1. Article Summary:

Year	Task	Domain-Oriented	Algorithms Used	Polarity	Data Scope	Data Set/Source	Other Language
2010	SA	y	Rule-Based	G	Web Forums	Automotiveforums.com	
2010	SA	N	Semantic, LSA-based	G	Software programs users' feedback	CNETD	
2011	SA	N	2-level CRF	G	Mobile Customer Reviews	Amazon.com, opinions.com, blogs, SNS	
2011	SA	N	Multi-class SVM	G	Digital Cameras,MP3 Reviews	N/A	
2011	SA	Y	SVM, Chi-square	G	Buyer's posts web pages	ebay.com,Wikipedia.com, opinions.com	
2011	SA	y	Semantic	G	Chinese training data	MOAT NTCIR7	Chinese

Volume 4 Issue III, March 2016
ISSN: 2321-9653

2011	SA	N	Statistical	G	Book Reviews	Amazon.com	
2012	SA	Y	Context based method, NLP	G	Restaurant Reviews	N/A	
2013	SA	Y	FCA	G	Smart Phones, Tweets	Twitter	

Where, SA - Sentiment Analysis

G - General

NLP - Natural Language Processing

a) *Supervised learning:* The supervised learning methods depend on the existence of labeled training documents. There are many kinds of classifier:

i) *Probabilistic classifiers:* Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers.

Naive Bayes Classifier (NB): The Naive Bayes classifier is the simplest and most frequently used classifier. NaiveBayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the Bag of words feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(h/D) = [P(D/h)\ P(h)]/\ P(D)$$

Where,

P(h) : Prior probability of hypothesis
h (predicted value)

P(D) : Prior probability of training
data D (value given)
P(h/D) : Probability of h given D (expected o/p)
P(D/h) : Probability of D given h

(mismatched o/p)

An improved NB classifier was proposed by Kang and Yo to solve the problem of the affinity for the positive classification accuracy to appear up to approximately 10% higher than the negative classification accuracy. This creates a problem of decreasing the average accuracy when the accuracies of the two classes are expressed as an average value. They showed that using this algorithm with restaurant reviews narrowed the gap between the positive accuracy and the negative accuracy compared to NB and

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

SVM. The accuracy is improved in recall and precision compared to both NB and SVM

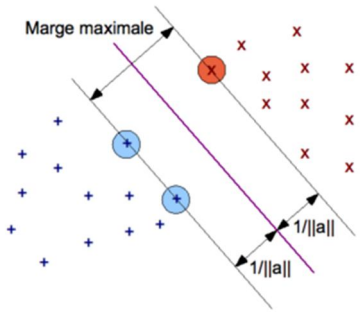
Maximum Entropy Classifier: The ME Classifier which known as a conditional exponential classifier converts labeled feature sets to vectors using encoding. This encoded vector is used to calculate weights for each feature that can then be pooled to find out the most likely label for a feature set. This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C{(feature-set, label)} pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)}$$

The other tools that were developed to automatically extract parallel data from non-parallel corpora use language specific techniques or require large amounts

b) Linear classifiers: There are many kinds of linear classifiers; among them is Support Vector Machines (SVM) which is a form of classifiers that try to decide good linear separators between different classes. Two of the most famous linear classifiers are discussed in the following subsections.

ii) Support Vector Machines Classifiers (SVM): The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. There are 2 classes x, o and there are 3 hyperplanes X, Y and Z. Hyperplane X provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane . SVMs are used in many applications, such as classifying reviews according to their quality.



Neural Network (NN): Neural Network consists of several neurons which is its basic unit. The inputs to the neurons are denoted by the vector overline Xi which is the word frequencies in the i^{th} document. There are a set of weights A which are associated with each neuron used in order to compute a function of its inputs f(). The linear function of the neural network is: pi ¼ A Xi. In a binary classification problem, it is assumed that the class label of Xi is denoted by yi and the sign of the predicted function pi yields the class label. Multilayer neural networks are used for non-linear boundaries. These multiple layers are used to induce multiple piecewise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers.

iii) Decision tree classifiers: Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to separate the data. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification. There are other kinds of predicates which depend on the similarity of documents to correlate sets of terms which may be used to further partitioning of documents. The different kinds of splits are Single Attribute split which use the presence or absence of particular words or phrases at a particular node in the tree in order to perform the split. Similarity-based multi-attribute split uses documents or frequent words clusters and the similarity of the documents to these words clusters in order to perform the split. Discriminat-based multi-attribute split uses discriminants such as the Fisher discriminate for performing the split. The decision tree implementations in text classification tend

to be small variations on standard packages such as ID3 and C4.5. Li and Jain have used the C5 algorithm which is a successor to the C4.5 algorithm. Depending on the concept of a tree; an approach was proposed by Hu and Li in order to mine the content structures of topical terms in sentence-level contexts by using the Maximum Spanning Tree (MST) structure to discover the links among the topical term ‘‘t’’ and its context words. Accordingly, they developed the so-called Topical Term Description Model for sentiment classification. In their definition, ‘‘topical terms’’ are those specified entities or certain aspects of entities in a particular domain. They introduced an automatic extraction of topical terms from text based on their domain term-hood. Then, they used these extracted terms to differentiate document topics. This structure conveys sentiment information. Their approach is different from the regular machine learning tree algorithms but is able to learn the positive and negative contextual knowledge effectively. A graph-based Approach was presented by Yan and Bing. They have presented a propagation approach to incorporate the inside and outside sentence features. These two sentence features are intra-document evidence and inter-document evidence. They said that determining the sentiment orientation of a review sentence requires more than the features inside the sentence itself. They have worked on camera domain and compared their method to both unsupervised approach and supervised approaches (NB, SVM). Their results showed that their proposed approach performs better than both approaches without using outside sentence features and outperforms other representational previous approaches.

iv) Rule-based classifiers: In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data. There are numbers of criteria in order to generate rules, the training phase construct all the rules depending on these criteria. The most two common criteria are support and confidence. The support is the absolute number of instances in the training data set which are relevant to the rule. The Confidence refers to the conditional probability that the right hand side of the rule is satisfied if the left-hand side is satisfied. Some combined rule algorithms were proposed. Both decision trees and decision rules tend to encode rules on the feature space, but the decision tree tends to achieve this goal with a hierarchical approach. Quinla has studied the decision tree and decision rule problems within a single framework; as a certain path in the decision tree can be considered a rule for classification of the text instance. The main difference between the decision trees and the decision rules is that DT is a strict hierarchical partitioning of the data space, while rule-based classifiers allow for overlaps in the decision space.

2) *Weakly, semi and unsupervised learning:* The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, large number of labeled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties. Many research works were presented in this field including the work presented by Ko and Seo. They proposed a method that divides the documents into sentences, and categorized each sentence using keyword lists of each category and sentence similarity measure. The concept of weak and semi-supervision is used in many applications. Youlan and Zho have proposed a strategy that works by providing weak supervision at the level of features rather than instances. They obtained an initial classifier by incorporating prior information extracted from an existing sentiment lexicon into sentiment classifier model learning. They refer to prior information as labeled features and use them directly to constrain model’s predictions on unlabeled instances using generalized expectation criteria.
3) *Meta classifiers:* In many cases, the researchers use one kind or more of classi- fiers to test their work. One of these articles is the work proposed by Lane and Clarke. They presented a ML approach to solve the problem of locating documents carrying positive or negative favorability within media analysis. The imbalance in the distribution of positive and negative samples, changes in the documents over time, and effective training and evaluation procedures for the models are the challenges they faced to reach their goal. They worked on three data sets Sentiment analysis algorithms and applications: A survey 1101 generated by a media-analysis company. They classified documents in two ways: detecting the presence of favorability, and assessing negative vs. positive favorability. They have used five different types of features to create the data sets from the raw text. They tested many classifiers to find the best one which are (SVM, K-nearest neighbor, NB, BN, DT, a Rule learner and other). They showed that balancing the class distribution in training data can be beneficial in improving performance, but NB can be adversely affected. Applying ML algorithms on streaming data from Twitter was investigated by Rui and Liu.

B. Lexicon-Based Approach

Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together are called opinion lexicon. There are three main approaches in order to compile or collect the opinion word list. Manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are presented in the following subsections.

1) *Dictionary-based approach:* It presented the main strategy of the dictionary-based approach. A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors. The dictionary based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations. Qiu and He used dictionary-based approach to identify sentiment sentences in contextual advertising. They proposed an advertising strategy to improve ad relevance and user experience. They used syntactic parsing and sentiment dictionary and proposed a rule based approach to tackle topic word extraction and consumers' attitude identification in advertising keyword extraction. They worked on web forums from automotvieforums.com. Their results demonstrated the effectiveness of the proposed approach on advertising keyword extraction and ad selection.

2) *Corpus-based approach:* The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur 1102 W. Medhat et al. together along with a seed list of opinion words to find other opinion words in a large corpus. One of these methods were represented by Hatzivassiloglou and McKeown. They started with a list of seed opinion adjectives, and used them along with a set of linguistic constraints to identify additional adjective opinion words and their orientations. The constraints are for connectives like AND, OR, BUT, EITHER-OR. . .. . .; the conjunction AND for example says that conjoined adjectives usually have the same orientation. This idea is called sentiment consistency, which is not always consistent practically. There are also adversative expressions such as but, however which are indicated as opinion changes. In order to determine if two conjoined adjectives are of the same or different orientations, learning is applied to a large corpus. Then, the links between adjectives form a graph and clustering is performed on the graph to produce two sets of words: positive and negative. The Conditional Random Fields (CRFs) method was used as a sequence learning technique for extracting opinion expressions. It was used too by Jiaoa and Zhou in order to discriminate sentiment polarity by multi-string pattern matching algorithm. Their algorithm was applied on Chinese online reviews. They established many emotional dictionaries. They worked on car, hotel and computer online reviews. Their results showed that their method has achieved high performance. Xu and Liao have used two-level CRF model with unfixed interdependencies to extract the comparative relations. This was done by utilizing the complicated dependencies between relations, entities and words, and the unfixed interdependencies among relations. Their purpose was to make a graphical model to extract and visualize comparative relations between products from customer reviews. They displayed the results as comparative relation maps for decision support in enterprise risk management. They worked on mobile customer reviews from amazon.com, epinions.com, blogs, SNS and emails. Their results showed that their method can extract comparative relations more accurately than other methods, and their comparative relation map is potentially a very effective tool to support enterprise risk management and decision making.

a) *Statistical approach:* Finding co-occurrence patterns or seed opinion words can be done using statistical techniques. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus, as proposed by Fahrni and Klenner. It is possible to use the entire set of indexed documents on the web as the corpus for the dictionary construction. This overcomes the problem of the unavailability of some words if the used corpus is not large enough. The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of text. If the word occurs more frequently among positive texts, then its polarity is positive. If it occurs more frequently among negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word. The similar opinion words frequently appear together in a corpus. This is the main observation that the state of the art methods are based on. Therefore, if two words appear together frequently within the same context, they are likely to have the same polarity. Therefore, the polarity of an unknown word can be determined by

calculating the relative frequency of co-occurrence with another word. This could be done using PMI. Statistical methods are used in many applications related to SA. One of them is detecting the reviews manipulation by conducting a statistical test of randomness called Runs test. Hu and Boss expected that the writing style of the reviews would be random due to the various backgrounds of the customers, if the reviews were written actually by customers. They worked on Book reviews from amazon.com and discovered that around 10.3% of the products are subject to online reviews manipulation. Latent Semantic Analysis (LSA) is a statistical approach which is used to analyze the relationships between a set of documents and the terms mentioned in these documents in order to produce a set of meaningful patterns related to the documents and terms.

b) *Semantic approach:* The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. This principle gives similar sentiment values to semantically close words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word. The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA as the work presented by Maks and Vossen. Their model described the detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor. These subjectivity relations are labeled with information concerning both the identity of the attitude holder and the orientation (positive vs. negative) of the attitude. Their model included a categorization into semantic categories relevant to SA. It provided means for the identification of the attitude holder, the polarity of the attitude and also the description of the emotions and sentiments of the different actors involved in the text. They used Dutch WordNet in their work. Their results showed that the speaker's subjectivity and sometimes the actor's subjectivity can be reliably identified. Semantics of electronic WOM (eWOM) content is used to examine eWOM content analysis as proposed by Pai and Chu. They extracted both positive and negative appraisals, and helped consumers in their decision making. Their method can be utilized as a tool to assist companies in better understanding product or service appraisals, and accordingly translating these opinions into business intelligence to be used as the basis for product/service improvements. They worked on Taiwanese Fast food reviews. Their results showed that their approach is effective in providing eWOM appraisals related to services and products. Semantic methods can be mixed with the statistical methods to perform SA task as the work presented by Zhang and Xu who used both methods to find product weakness from online reviews. Their weakness finder extracted the features and group explicit features by using morpheme-based method to identify feature words from the reviews. They used Hownet-based similarity measure to find the frequent and infrequent explicit features which describe the same aspect. They identified the implicit features with collocation statistics-based selection method PMI.

3) *Lexicon-based and natural language processing techniques:* Natural Language Processing (NLP) techniques are sometimes used with the lexicon-based approach to find the syntactical structure and help in finding the semantic relations. Moreo and Romero have used NLP techniques as preprocessing stage before they used their proposed lexicon-based SA algorithm. Their proposed system consists of an automatic focus detection module and a sentiment analysis module capable of assessing user opinions of topics in news items which use a taxonomy-lexicon that is specifically designed for news analysis. Their results were promising in scenarios where colloquial language predominates. The approach for SA presented by Caro and Grella was based on a deep NLP analysis of the sentences, using a dependency parsing as a pre-processing step. Their SA algorithm relied on the concept of Sentiment Propagation, which assumed that each linguistic element like a noun, a verb, etc. can have an intrinsic value of sentiment that is propagated through the syntactic structure of the parsed sentence. They presented a set of syntactic-based rules that aimed to cover a significant part of the sentiment salience expressed by a text. They proposed a data visualization system in which they needed to filter out some data objects or to contextualize the data so that only the information relevant to a user query is shown to the user.

In order to accomplish that, they presented a context-based method to visualize opinions by measuring the distance, in the textual appraisals, between the query and the polarity of the words contained in the texts themselves. They extended their algorithm by computing the context-based polarity scores. Their approach approved high efficiency after applying it on a manual corpus of 100 restaurants reviews. Min and Park [39] have used NLP from a different perspective. They used NLP techniques to identify tense and time expressions along with mining techniques and a ranking algorithm. Their proposed metric has two parameters that capture time

expressions related to the use of products and product entities over different purchasing time periods. They identified important linguistic clues for the parameters through an experiment with crawled review data, with the aid of NLP techniques. They worked on product reviews from amazon.com. Their results showed that their metric was helpful and free from undesirable biases.

C. Other Techniques

There are techniques that cannot be roughly categorized as ML approach or lexicon-based Approach. Formal Concept Analysis (FCA) is one of those techniques. FCA was proposed by Wille as a mathematical approach used for structuring, analyzing and visualizing data, based on a notion of duality called Galois connection. The data consists of a set of entities and its features are structured into formal abstractions called formal concepts. Together they form a concept lattice ordered by a partial order relation. The concept lattices are constructed by identifying the objects and their corresponding attributes for a specific domain, called conceptual structures, and then the relationships among them are displayed. Fuzzy Formal Concept Analysis (FFCA) was developed in order to deal with uncertainty and unclear information. It has been successfully applied in various information domain applications. FCA and FFCA were used in many SA applications as presented by Li and Tsai. In their work they proposed a classification framework based on FFCA to conceptualize documents into a more abstract form of concepts.

IV. RELATED FIELDS TO SENTIMENT ANALYSIS

There are some topics that work under the umbrella of SA and have attracted the researchers recently. In the next subsection, three of these topics are presented in some details with related articles.

A. Emotion Detection

Sentiment analysis is sometimes considered as an NLP task for discovering opinions about an entity; and because there is some haziness about the difference between opinion, sentiment and emotion, they defined opinion as a transitional concept that reflects attitude towards an entity. The sentiment reflects feeling or emotion while emotion reflects attitude . Sentiment analysis algorithms and applications: It was argued by Plutchik that there are eight basic and prototypical emotions which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Emotions Detection (ED) can be considered a SA task. SA is concerned mainly in specifying positive or negative opinions, but ED is concerned with detecting various emotions from text. As a Sentiment Analysis task, ED can be implemented using ML approach or Lexicon-based approach, but Lexicon-based approach is more frequently used. They proposed a web-based text mining approach for detecting emotion of an individual event embedded in English sentences. Their approach was based on the probability distribution of common mutual actions between the subject and the object of an event. They integrated web-based text mining and semantic role labeling techniques, together with a number of reference entity pairs and hand-crafted emotion generation rules to recognize an event emotion detection system. They did not use any large-scale lexical sources or knowledge base. They showed that their approach revealed a satisfactory result for detecting the positive, negative and neutral emotions. They proved that the emotion sensing problem is context-sensitive. Using both ML and Lexicon-based. They proposed a method based on commonsense knowledge stored in the emotion corpus (EmotiNet) knowledge base. They said that emotions are not always expressed by using words with an affective meaning i.e. happy, but by describing real-life situations, which readers detect as being related to a specific emotion. They used SVM and SVM-SO algorithms to achieve their goal. They showed that the approach based on EmotiNet is the most appropriate for the detection of emotions from contexts where no affect related words were present. They proved that the task of emotion detection from texts such as the ones in the emotion corpus ISEAR (where little or no lexical clues of affect are present) can be best tackled using approaches based on commonsense knowledge. They showed that by using EmotiNet, they obtained better results compared to the methods that employ supervised learning on a much greater training set or lexical knowledge. Affect Analysis (AA) is a task of recognizing emotions elicited by a certain semiotic modality. Neviarouskaya et al. have suggested an Affect Analysis Model (AAM). Their AAM consists of five stages: symbolic cue, syntactical structure, word-level, phrase-level and sentence-level analysis. This AAM was used in many applications presented in Neviarouskaya work . Classifying sentences using fine-grained attitude types is another work presented by Neviarouskaya et al. They developed a system that relied on the compositionality principle and a novel approach dealing with the semantics of verbs in attitude analysis. They worked on 1000 sentences from http:// www.experienceproject.com. This is a site where people share personal experiences, thoughts, opinions, feelings, passions, and confessions through the network of personal stories. Their evaluation showed that their system achieved reliable results in the task of textual attitude analysi. In their work, they introduced a bootstrapping algorithm based on contextual and lexical features for identifying paraphrases and to extract them for emotion terms, from nonparallel corpora. They started with a small number of seeds (WordNet Affect emotion words). Their approach learned

extraction patterns for six classes of emotions. They used annotated blogs and other data sets as texts to extract paraphrases from them. They worked on data from live journals blogs, text affect, fairy tales and annotated blogs. They showed that their algorithm achieved good performance results on their data set have worked on text-based affect analysis (AA) of Japanese narratives from Aozora Bunko. In their research, they addressed the problem of person/character related affect recognition in narratives. They extracted emotion subject from a sentence based on analysis of anaphoric expressions at first, then the affect analysis procedure estimated what kind of emotional state each character was in for each part of the narrative. Studying AA in mails and books was introduced by Mohammad . He has analyzed the Enron email corpus and proved that there were marked differences across genders in how they use emotion words in work-place email. He created lexicon which has manual annotations of a word's associations with positive/negative polarity, and the eight basic emotions by crowd-sourcing. He used it to analyze and track the distribution of emotion words in books and mails. He introduced the concept of emotion word density by studying novels and fairy tales. He proved that pixie tales had a much wider distribution of emotional word densities than novels.

B. Building Resources

Building Resources (BR) aims at creating lexica, dictionaries and corpora in which opinion expressions are annotated according to their polarity. Building resources is not a SA task, but it could help to improve SA and ED as well. The main challenges that confronted the work in this category are ambiguity of words, multilinguality, granularity and the differences in opinion expression among textual genres .In their work, they proposed a random walk algorithm to construct domain-oriented sentiment lexicon by simultaneously utilizing sentiment words and documents from both old domain and target domain. They conducted their experiments on three domain-specific sentiment data sets. Their experimental results indicated that their proposed algorithm improved the performance of automatic construction of domain-oriented sentiment lexico. They proposed Opinion Mining-ML, a new XML-based formalism for tagging textual expressions conveying opinions on objects that are considered relevant in the state of affairs. It is a new standard beside Emotion-ML and WordNet. Their work consisted of two parts. First, they presented a standard methodology for the annotation of affective statements in the text that was strictly independent from any application domain. Second, they considered the domain-specific adaptation that relied on the use of ontology of support which is domain dependent. They started with data set of restaurant reviews applying query-oriented extraction process. They evaluated their proposal by means of fine-grained analysis of the disagreement between different annotators. Their results indicated that their proposal represented an effective annotation scheme that 1106 W. Medhat et al. was able to cover high complexity while preserving good agreement among different people. They focused on the annotation at different levels: document, sentence and element. They also presented the EmotiBlog corpus; a collection of blog posts composed by 270,000 token about three topics in three languages: Spanish, English and Italian. They checked the robustness of the model and its applicability to NLP tasks. They tested their model on many corpora i.e. ISEAR. Their experiments provided satisfactory results. They applied EmotiBlog to sentiment polarity classification and emotion detection. They proved that their resource improved the performance of systems built for this task. In their work they proposed a semi-automatic approach to creating sentiment dictionaries in many languages. They first produced high-level gold-standard sentiment dictionaries for two languages and then translated them automatically into a third language. Those words that can be found in both target language word lists are likely to be useful because their word senses are likely to be similar to that of the two source languages. They addressed two issues during their work; the morphological inflection and the subjectivity involved in the human annotation and evaluation effort. They worked on news data. They compared their triangulated lists with the non-triangulated machine-translated word lists and verified their approach.

C. Transfer Learning

Transfer learning extracts knowledge from auxiliary domain to improve the learning process in a target domain. For example, it transfers knowledge from Wikipedia documents to tweets or a search in English to Arabic. Transfer learning is considered a new cross domain learning technique as it addresses the various aspects of domain differences. It is used to enhance many Text mining tasks like text classification , sentiment analysis , Named Entity recognition , part-of-speech tagging , ... etc.

V. DISCUSSION AND ANALYSIS

In this section, we analyze the trend of researchers in using the various algorithms, data or accomplishing one of the SA tasks. Sentiment analysis algorithms and applications: the number of the articles that give contribution to the six categories of SA tasks among years and the overall count. This figure shows that still SA and SC attract researchers more frequently. It can be noticed that they have almost equal number of contributions among years and the biggest amount in the overall count. The related fields ED, TL

Volume 4 Issue III, March 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

and BR have attracted researchers more recently as they are emerging fields of search. ML algorithms are usually used to solve the SC problem for its simplicity and the ability to use the training data which gives it the privilege of domain adaptability. Lexicon-based algorithms are frequently used to solve general SA problems because of their scalability. They are also simple and computationally efficient. Fig. 5 shows the algorithms used. As shown the number and percentage of articles that use ML and the Lexicon-based algorithms are changing among years.

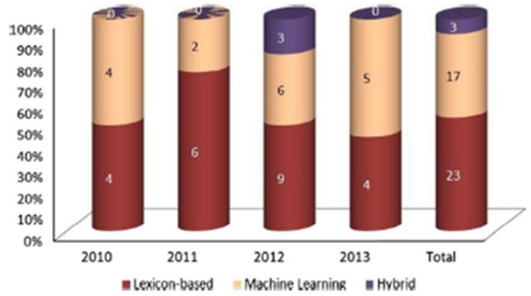
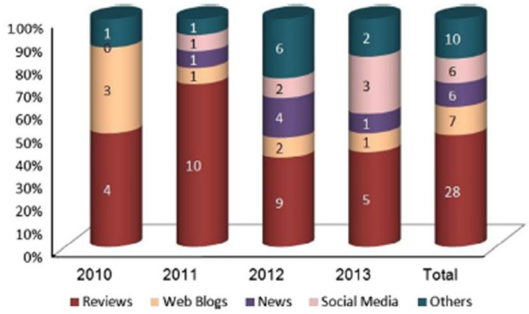


Figure 5 Number and percentage of articles according to the algorithmic approach over years.

The overall work for the recent few years shows that the researchers are using lexicon-based approach more frequently. This is because it solves many SA tasks despite its high complexity. ML approaches are still an open field of search. In the past, the binary classification problem has been a nice first step, as it involves distinguishing between the two extremes of the polarity spectrum. Therefore, binary polarity classification is a comparably easy problem to tackle, due to its inherently crisp nature, as well as the availability of (lots of) data that can easily be used for this purpose. Identifying a general mood is little bit difficult .



Number and percentage of articles targeting different text domains over years. 1108 W. Medhat et al. research that and has an increasing trend with time. The SA field is expanding to absorb other related fields rather than binary classification (pos/neg classification). We can notice that in the year 2012, most of the articles were targeting the related fields of SA other than the normal SC problem. This explains why the use of the Lexicon-based approaches is more often used recently; as the general classification is not frequently used with ML algorithms. The data used in SA are mostly on Product Reviews in the overall count . The other kinds of data are used more frequently over recent years specially the social media. The other kinds of data are news articles or news feeds; web Blogs, social media, and others. We are interested too in seeing if the data used in the articles are domain dependent or not. Many articles have proved that using domain dependent data gives more accurate results than domain-independent data as in [35,60]. it is shown that the researchers usually work in a domain-independent for its simplicity. This makes the domain-dependent a problem or as so-called a context-based SA; an ongoing field of search. SA using non-English languages has attracted researchers recently as shown in Fig. 9. The non-English languages include the other Latin languages (Spanish, Italian); Germanic languages (German, Dutch); Far East languages (Chinese, Japanese, Taiwanese); Middle East languages (Arabic). In that, still, the English language is the most frequently used language due to the availability of its resources including lexica, corpora and dictionaries. This opens a new challenge to researchers in order to build lexica, corpora and dictionaries resources for other languages.

A. Open Problems

The analysis illustrated above gives a closer look at the recent and future trend of research. While studying the recent articles, we

have discovered some points that could be considered open problems in research. The Data Problem: It has been noticed that there is lack of benchmark data sets in this field. It was stated in that few of the most famous data sets are in the field of SA. These tasks do not use the famous customer reviews as its data source. They may use novels, narratives or mails in their study which are not used in other SA tasks. IMDB and Amazon.com are very famous data sources of review data. IMDB is a source of movie reviews while amazon.com is a source of many product reviews. These data sources are used in SA and SC tasks. It is noticed that twitter was used frequently in the last year. Twitter is a very famous social network site where its tweets express people's opinions and its length is maximum 140 characters. The debate site called convinceme.net is considered also a good data set which was used in SC task. The other sources are illustrated in the rest of the table. The Language problem: It was noticed in the articles presented in this survey that the Far East languages especially the Chinese language has been used more often recently. Accordingly, many sources of data are built for these languages. The researchers are now in the phase of building resources of other Latin (European) languages. There is still a lack of resources for the Middle East languages including the Arabic language. The resources built for the Arabic language are not yet complete and not found easily as an open source. This makes it a very good trend of research now. NLP: The natural language processing tools can be used to facilitate the SA process. It gives better natural language understanding and thus can help produce more accurate results of SA. These tools were used to help in BR, ED and also SA task in the last two years. This opens a new trend of research of using the NLP as a preprocessing stage before sentiment analysis . Although mentioned the problems of opinion aggregation and contradiction analysis, they were not found in the recent articles presented by this survey. This means that they do not attract researchers recently; despite the fact that they are still opening fields of research. It is noticed that working on domain-specific corpus gives better results than working on the domain-independent corpus. There is still lack of research in the field of domain-specific SA which is sometimes called context-based SA. This is because building the domain-specific corpus is more complicated than using the domain-independent one. It is noticed that the ED and BR task work usually on domain-independent sources, while TL always uses domain-dependent sources.

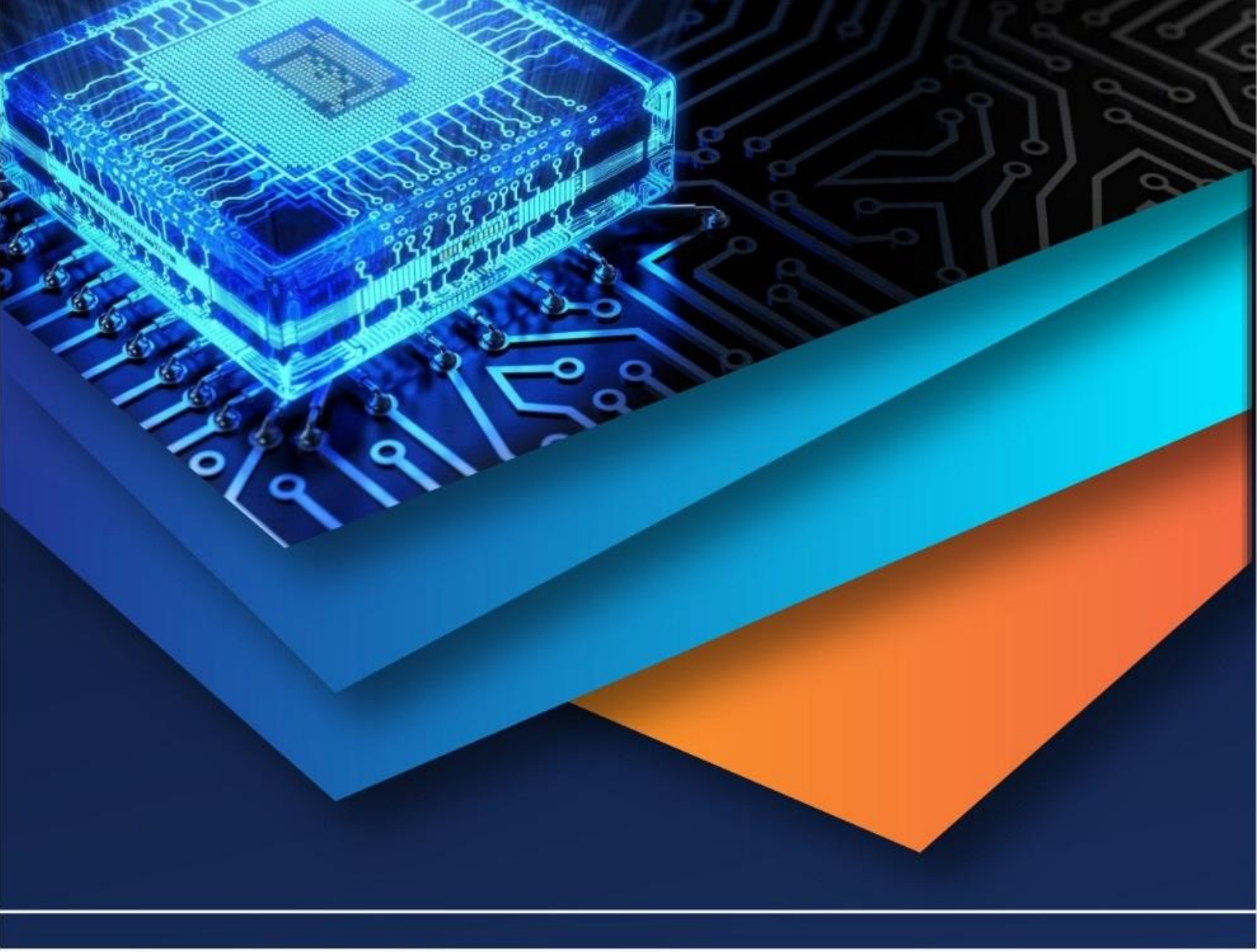
VI. CONCLUSION AND FUTURE WORK

This survey paper presented an overview on the recent updates in SA algorithms and applications. Fifty-four of the recently published and cited articles were categorized and summarized. These articles give contributions to many SA related fields that use SA techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research. NaiveBayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem. They are considered a reference model where many proposed algorithms are compared to. The interest in languages other than English in this field is growing as there is still a lack of resources and researches concerning these languages. The most common lexicon source used is WordNet which exists in languages other than English. Building resources, used in SA tasks, is still needed for many natural languages.Information from micro-blogs, blogs and forums as well as news source, is widely used in SA recently. This media information plays a great role in expressing people's feelings, or opinions about a certain topic or product. Using social network sites and micro-blogging sites as a source of data still needs deeper analysis. There are some benchmark data sets especially in reviews like IMDB which are used for algorithms evaluation. In many applications, it is important to consider the contextof the text and the user preferences. That is why we need to make more research on context-based SA. Using TL techniques, we can use related data to the domain in question as a training data.Using NLP tools to reinforce the SA process has attracted researchers recently and still needs some enhancements.

REFERENCES

[1] Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Min Knowl Discov 2012;24:478–514.
[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/ EMNLP; 2005.
[3] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.
[4] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl-Based Syst 2013;41:89–97.
[5] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.
[6] Tao Xu, Peng Qinke, Cheng Yinzhao. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. KnowlBased Syst 2012;35:279–89.
[7] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 2012;53:680–8.
[8] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.
[9] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21.
[10] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:829.
[11] Montoyo Andre´s, Martı´nez-Barco Patricio, Balahur Alexandra. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged

developments. Decis Support Syst 2012;53:675–9.
[12] Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun. DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Syst Appl 2010;37:6182–91.
[13] Lu Cheng-Yu, Lin Shian-Hua, Liu Jen-Chang, Cruz-Lara Samuel, Hong Jen-Shin. Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Syst Appl 2010;37:1643–53.
[14] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Recognition of Affect, Judgment, and Appreciation in Text. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Beijing; 2010. p. 806–14.
[15] Bai X. Predicting consumer sentiments from online text. Decis Support Syst 2011;50:732–42.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY