Ain Shams Engineering Journal (2014) 5, 1093–1113

Ain Shams University

Ain Shams Engineering Journal

www.elsevier.com/locate/asej
www.sciencedirect.com

ELECTRICAL ENGINEERING

Sentiment analysis algorithms and applications: A survey

Walaa Medhat [a,*], Ahmed Hassan [b], Hoda Korashy [b]

[a] School of Electronic Engineering, Canadian International College, Cairo Campus of CBU, Egypt
[b] Ain Shams University, Faculty of Engineering, Computers & Systems Department, Egypt

Received 8 September 2013; revised 8 April 2014; accepted 19 April 2014
Available online 27 May 2014

KEYWORDS

Sentiment analysis;
Sentiment classification;
Feature selection;
Emotion detection;
Transfer learning;
Building resources

Abstract Sentiment Analysis (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text. This survey paper tackles a comprehensive overview of the last update in this field. Many recently proposed algorithms' enhancements and various SA applications are investigated and presented briefly in this survey. These articles are categorized according to their contributions in the various SA techniques. The related fields to SA (transfer learning, emotion detection, and building resources) that attracted researchers recently are discussed. The main target of this survey is to give nearly full image of SA techniques and the related fields with brief details. The main contributions of this paper include the sophisticated categorizations of a large number of recent articles and the illustration of the recent trend of research in the sentiment analysis and its related areas.

1. Introduction

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two expressions SA or OM are interchangeable. They express a mutual meaning. However, some researchers stated that OM and SA have slightly different notions [1]. Opinion Mining extracts and analyzes people's opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Fig. 1.

Sentiment Analysis can be considered a classification process as illustrated in Fig. 1. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment 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 determine whether the sentence expresses positive or negative opinions. Wilson et al. [2] have pointed out that sentiment expressions

* Corresponding author. Address: School of Electronic Engineering, Canadian International College, 22 Emarate Khalf El-Obour, Masr Elgedida, Cairo, Egypt. Tel.: +20 24049568.
E-mail address: walaamedhat@gmail.com (W. Medhat).
Peer review under responsibility of Ain Shams University.

Production and hosting by Elsevier

2090-4479 © 2014 Production and hosting by Elsevier B.V. on behalf of Ain Shams University.
http://dx.doi.org/10.1016/j.asej.2014.04.011

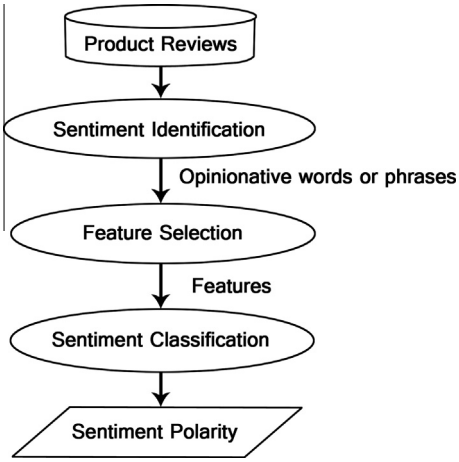


Figure 1 Sentiment analysis process on product reviews.

are not necessarily subjective in nature. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents [3]. Classifying text at the document level or at the sentence level does not provide the necessary 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. Aspect-level SA aims to classify the sentiment with respect to the specific 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 voice quality of this phone is not good, but the battery life is long*". This survey tackles the first two kinds of SA.

The data sets used in SA are an important issue in this field. The main sources of data are from the product reviews. These reviews are important to the business holders as they can take business decisions according to the analysis results of users' opinions about their products. The reviews sources are mainly review sites. SA is not only applied on product reviews but can also be applied on stock markets [4,5], news articles, [6] or political debates [7]. In political debates for example, we could figure out people's opinions on a certain election candidates or political parties. The election results can also be predicted from political posts. 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. The authors have collected fifty-four articles which presented important enhancements to the SA field lately. These articles cover a wide variety of SA fields. They were all published in the last few years. They are categorized according to the target of the article illustrating the algorithms and data used in their work. According to Fig. 1, the authors have discussed the Feature Selection (FS) techniques in details along with their related articles referring to some originating references. The Sentiment Classification (SC) techniques, as shown in Fig. 2, are discussed with more details illustrating related articles and originating references as well.

This survey can be useful for new comer researchers in this field as it covers the most famous SA techniques and applications in one research 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. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries. In this paper, the authors give a closer look on these fields.

There are numerous number of articles presented every year in the SA fields. The number of articles is increasing through years. This creates a need to have survey papers that summarize the recent research trends and directions of SA. The reader can find some sophisticated and detailed surveys including [1,3,8–11]. Those surveys have discussed the problem of SA from the applications point of view not from the SA techniques point of view.

Two long and detailed surveys were presented by Pang and Lee [8] and Liu [3]. They focused on the applications and challenges in SA. They mentioned the techniques used to solve each problem in SA. Cambria and Schuller et al. [9], Feldman [10] and Montoyo and Martínez-Barco [11] have given short surveys illustrating the new trends in SA. Tsytsarau and Palpanas [1] have presented a survey which discussed the main topics of SA in details. For each topic they have illustrated its definition, problems and development and categorized the articles with the aid of tables and graphs. The analysis of the articles presented in this survey is similar to what was given by [1] but with another perspective and different categorization of the articles.

The contribution of this survey is significant for many reasons. First, this survey provides sophisticated categorization of a large number of recent articles according to the techniques used. This angle could help the researchers who are familiar with certain techniques to use them in the SA field and choose the appropriate technique for a certain application. Second, the various techniques of SA are categorized with brief details of the algorithms and their originating references. This can help new comers to the SA field to have a panoramic view on the entire field. Third, the available benchmarks data sets are discussed and categorized according to their use in certain applications. Finally, the survey is enhanced by discussing the related fields to SA including emotion detection, building resources and transfer learning.

This paper is organized as follows: Section 2 includes the survey methodology and a summary of the articles. Section 3 tackles the FS techniques and their related articles, and Section 4 discusses the various SC techniques and the corresponding articles. In Section 5, the related fields to SA and their corresponding articles are presented. Section 6 presents the results and discussions, and finally the conclusion and future trend in research are tackled in Section 7.

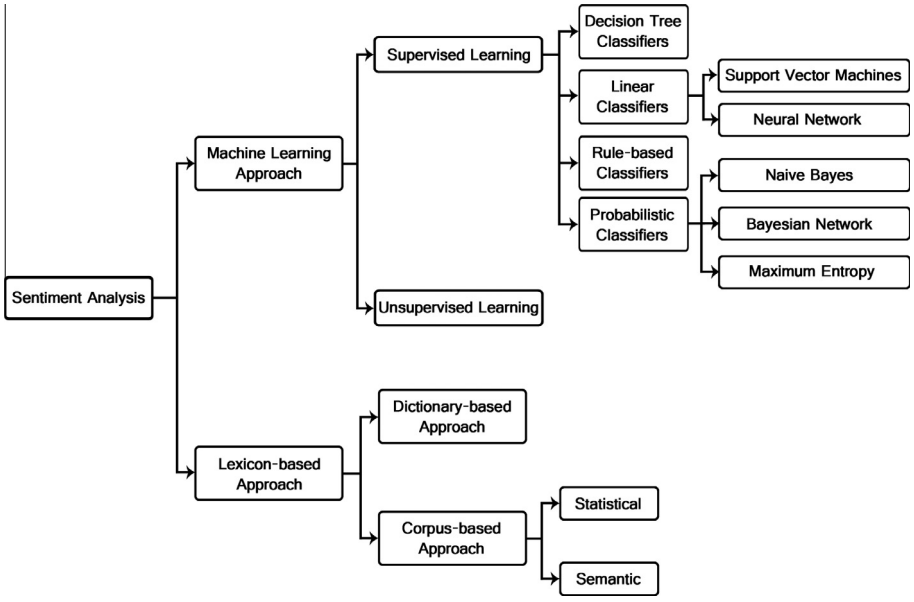


Figure 2 Sentiment classification techniques.

2. Methodology

The fifty-four articles presented in this survey are summarized in Table 1. Table 1 contains the articles reference [4–7] and [12–61]. 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 as shown in Fig. 2. 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 [43]. 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.

3. Feature selection in sentiment classification

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are [62]:

Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features [63].

Parts of speech (POS): finding adjectives, as they are important indicators of opinions.

Opinion words and phrases: these are words commonly used to express opinions including *good or bad*, *like or hate*. On the other hand, some phrases express opinions without using opinion words. For example: *cost me an arm and a leg*.

Negations: the appearance of negative words may change the opinion orientation like *not good* is equivalent to *bad*.

3.1. Feature selection methods

Feature Selection methods can be divided into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used. Lexicon-based approaches usually begin with a small set of 'seed' words. Then they bootstrap this set through synonym detection or on-line resources to obtain a larger lexicon. This proved to have many difficulties as reported by Whitelaw et al. [64]. Statistical approaches, on the other hand, are fully automatic.

The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string

Table 1 Articles Summary.

References	Year	Task	Domain-oriented	Algorithms used	Polarity	Data scope	Data set/source	Other language
[12]	2010	SA	Y	Rule-Based	G	Web Forums	automotvieforums.com	
[13]	2010	ED	N	Web-based, semantic labeling and rule-based	Pos/Neg	Web pages	N/A	
[14]	2010	ED	N	Lexicon-based, semantic	G	Personal stories	experienceproject.com	
[15]	2010	SC	N	Markov Blanket, SVM, NB, ME	Pos/Neg	Movie Reviews, News Articles	IMDB	
[16]	2010	SC	N	Graph-Based approach, NB, SVM	Pos/Neg	Camera Reviews	Chinese Opinion Analysis Domain (COAE)	Chinese
[17]	2010	SC	Y	Graph-Based approach	Pos/Neg	Movie, Product Reviews	N/A	Chinese
[18]	2010	SA	N	Semantic, LSA-based	G	Software programs users' feedback	CNETD	
[19]	2010	SC	Y	Weakly and semi supervised classification	Pos/Neg	Movie Reviews, Multi-domain sentiment data set	IMDB, Amazon.com	
[20]	2011	BR	Y	Random walk algorithm	G	Electronics, Stock, Hotel Reviews	Domain-specific chinese corpus	Chinese
[21]	2011	TL	Y	Entropy-based algorithm	G	Education, Stock, Computer Reviews	Domain-specific chinese data set	Chinese
[22]	2011	TL	Y	Ranking algorithm	G	Book, Hotel, Notebook Reviews	Domain-specific chinese data set	Chinese
[23]	2011	SC	N	CRFs	Pos/Neg	Car, Hotel, Computer Reviews	N/A	Chinese
[24]	2011	TL	Y	SAR	G	Movie Reviews, QA	MPQA, RIMDB, CHES	
[25]	2011	SA	N	2-level CRF	G	Mobile Customer Reviews	amazon.com, epinions.com, blogs, SNS and emails in CRM	
[26]	2011	SA	N	Multi-class SVM	G	Digital Cameras, MP3 Reviews	N/A	
[27]	2011	SA	Y	SVM, Chi-square	G	Buyers' posts web pages	ebay.com, wikipedia.com, epinions.com	
[28]	2011	SA	N	Semantic	G	Chinese training data	NTCIR7 MOAT	Chinese
[29]	2011	SC	N	Lexicon-based, semantic	Pos/Neg	Movie Reviews	IMDB	
[30]	2011	SC	N	Statistical (MM), semantic	Pos/Neg	Product Reviews	amazon.com	
[31]	2011	SA	N	Statistical	G	Book Reviews	amazon.com	
[32]	2011	TL	Y	Shared learning approach	G	Social Media, News data	Blogspot, Flicker, youtube, CNN-BBC	
[33]	2012	FS	N	Statistical (HMM - LDA), ME	Pos/Neg	Movie Reviews	N/A	
[34]	2012	BR	Y	Semantic	G	Restaurant Reviews	N/A	Spanish
[35]	2012	SA	Y	Context-based method, NLP	G	Restaurant Reviews	N/A	
[36]	2012	SC	N	NB, SVM	Pos/Neg	Restaurant Reviews	N/A	
[37]	2012	SA	N	Lexicon-Based, NLP	G	News	N/A	
[38]	2012	SA	N	PMI, semantic	G	Product Reviews	N/A	Chinese
[39]	2012	SA	N	NLP	G	Product Reviews	amazon.com	
[40]	2012	SC	N	Semi supervised, BN	G	Artificial data sets	N/A	
[41]	2012	BR	Y	NLP	G	blogs	ISEAR	Spanish, italian
[42]	2012	ED	Y	Corpus-based	G	Blogs data	Live Journals Blogs, Text Affect, Fairy tales, Annotated Blogs	
[6]	2012	SA	N	S-HAL, SO-PMI	G	News Pages	Sogou CS corpus	Chinese
[43]	2012	BR	N	Triangulation	G	News Articles	sentiment Dictionaries	Other Latin, Arabic
[44]	2012	SC	Y	NB, SVM, rule-based	G	2 sided debates	convinceme.net	
[45]	2012	ED	N	Lexicon-Based, SVM	G	Emotions corpus	ISEAR, Emotinet	
[7]	2012	SA	N	Semantic	Pos/Neg	Lexicons	Dutch wordnet	Dutch
[46]	2012	SA	N	SVM, K-nearest neighbor, NB, BN, DT, a Rule learner	Pos/Neg	Media	media-analysis company	

Table 1 *(continued)*

References	Year	Task	Domain-oriented	Algorithms used	Polarity	Data scope	Data set/source	Other language
[47]	2012	SC	N	SVM, 1-NN	Pos/Neg	Relationships Biography	BWSA	Dutch
[48]	2012	FS	N	Semantic, NB, SVM, DT	G	News, Satiric Articles, Customer Reviews	amazon.com	
[49]	2012	ED	N	Lexicon-Based	G	Emails, Books, Novels, fairy tales	Enron Email corpus	
[50]	2012	SC	N	Unsupervised, LDA	G	Social Reviews	2000-SINA blog data set, 300-SINA Hownet lexicon	Chinese
[51]	2012	SC	N	FCA, FFCA	Pos/Neg	Movie, eBook Reviews	Reuters 21578	
[52]	2012	BR	N	ME	G	corpora	N/A	Dutch, Chinese
[53]	2013	SC	N	SVM, ANN	Pos/Neg	Movie, GPS, Camera, Books Reviews	amazon.com	
[54]	2013	SC	N	SVM, NB, C4.5	Pos/Neg	Film Reviews	MC, MCE corpus	Spanish
[55]	2013	SA	Y	FCA	G	Smartphones, Tweets	Twitter	
[56]	2013	SC	N	NB, SVM	Pos/Neg	Movie Reviews, Tweets	Twitter	
[57]	2013	SC	N	SVM	G	Tweets	Twitter	
[5]	2013	FS	Y	Chi-square, BNS, SVM	G	Stock Market	DGAP, EuroAhoc	
[58]	2013	ED	Y	NLP	G	Narratives	Aozora Bunko	Japanese
[59]	2013	SA	N	Semantic	Pos/Neg	Fast Food Reviews	N/A	Taiwanese
[60]	2013	SC	Y	Taxonomy-based, corpus-based	Pos/Neg	Headphones, Car, Hotel Reviews	epinions.com	
[61]	2013	SC	N	Semantic	Pos/Neg	Blog Posts	TREC 2006, TREC 2007, and TREC 2008	
[4]	2013	FS	Y	PMI-Based	G	Stock News	N/A	

which retains the sequence of words in the document. BOW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word to its stem or root i.e. flies → fly).

In the next subsections, we present three of the most frequently used statistical methods in FS and their related articles. There are other methods used in FS like information gain and Gini index [62].

3.1.1. Point-wise Mutual Information (PMI)

The mutual information measure provides a formal way to model the mutual information between the features and the classes. This measure was derived from the information theory [65]. The point-wise mutual information (PMI) $M_i(w)$ between the word w and the class i is defined on the basis of the level of co-occurrence between the class i and word w. The expected co-occurrence of class i and word w, on the basis of mutual independence, is given by $P_i \bullet F(w)$, and the true co-occurrence is given by $F(w) \bullet p_i(w)$.

The mutual information is defined in terms of the ratio between these two values and is given by the following equation:

$$M_i(w) = \log\left(\frac{F(w)\cdot p_i(w)}{F(w)\cdot P_i}\right) = \log\left(\frac{p_i(w)}{P_i}\right) \tag{1}$$

The word w is positively correlated to the class i, when $M_i(w)$ is greater than 0. The word w is negatively correlated to the class i when $M_i(w)$ is less than 0.

PMI is used in many applications, and there are some enhancements applied to it. PMI considers only the co-occurrence strength. Yu and Wu [4] have extended the basic PMI by developing a contextual entropy model to expand a set of seed words generated from a small corpus of stock market news articles. Their contextual entropy model measures the similarity between two words by comparing their contextual distributions using an entropy measure, allowing for the discovery of words similar to the seed words. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Their results showed that their method can discover more useful emotion words, and their corresponding intensity improves their classification performance. Their method outperformed the (PMI)-based expansion methods as they consider both co-occurrence strength and contextual distribution, thus acquiring more useful emotion words and fewer noisy words.

3.1.2. Chi-square (χ^2)

Let n be the total number of documents in the collection, $p_i(w)$ be the conditional probability of class i for documents which contain w, P_i be the global fraction of documents containing the class i, and $F(w)$ be the global fraction of documents which contain the word w. Therefore, the χ^2-statistic of the word between word w and class i is defined as [62]:

$$\chi_i^2 = \frac{n\cdot F(w)^2\cdot (p_i(w)-P_i)^2}{F(w)\cdot(1-F(w))\cdot P_i\cdot(1-P_i)} \tag{2}$$

χ^2 and PMI are two different ways of measuring the correlation between terms and categories. χ^2 is better than PMI as it is a normalized value; therefore, these values are more comparable across terms in the same category [62].

χ^2 is used in many applications; one of them is the contextual advertising as presented by Fan and Chang [27]. They discovered bloggers' immediate personal interests in order to improve online contextual advertising. They worked on real ads and actual blog pages from ebay.com, wikipedia.com and epinions.com. They used SVM (illustrated with details in the next section) for classification and χ^2 for FS. Their results showed that their method could effectively identify those ads that are positively-correlated with a blogger's personal interests.

Hagenau and Liebmann [5] used feedback features by employing market feedback as part of their feature selection process regarding stock market data. Then, they used them with χ^2 and Bi-Normal Separation (BNS). They showed that a robust feature selection allows lifting classification accuracies significantly when combined with complex feature types. Their approach allows selecting semantically relevant features and reduces the problem of over-fitting when applying a machine learning approach. They used SVM as a classifier. Their results showed that the combination of advanced feature extraction methods and their feedback-based feature selection increases classification accuracy and allows improved sentiment analytics. This is because their approach allows reducing the number of less-explanatory features, i.e. *noise*, and limits negative effects of over-fitting when applying machine learning approaches to classify text messages.

3.1.3. Latent Semantic Indexing (LSI)

Feature selection methods attempt to reduce the dimensionality of the data by picking from the original set of attributes. Feature transformation methods create a smaller set of features as a function of the original set of features. LSI is one of the famous feature transformation methods [66]. LSI method transforms the text space to a new axis system which is a linear combination of the original word features. Principal Component Analysis techniques (PCA) are used to achieve this goal [67]. It determines the axis-system which retains the greatest level of information about the variations in the underlying attribute values. The main disadvantage of LSI is that it is an unsupervised technique which is blind to the underlying class-distribution. Therefore, the features found by LSI are not necessarily the directions along which the class-distribution of the underlying documents can be best separated [62].

There are other statistical approaches which could be used in FS like Hidden Markov Model (HMM) and Latent Dirichlet Allocation (LDA). They were used by Duric and Song [33] to separate the entities in a review document from the subjective expressions that describe those entities in terms of polarities. This was their proposed new feature selection schemes. LDA are generative models that allow documents to be explained by unobserved (latent) topics. HMM-LDA is a topic model that simultaneously models topics and syntactic structures in a collection of documents [68]. The feature selection schemes proposed by Duric and Song [33] achieved competitive results for document polarity classification specially when using only the syntactic classes and reducing the overlaps with the semantic words in their final feature sets. They worked on movie reviews and used Maximum Entropy (ME) classifier (illustrated with details in the next section).

3.2. Challenging tasks in FS

A very challenging task in extracting features is irony detection. The objective of this task is to identify irony reviews. This work was proposed by Reyes and Rosso [48]. They aimed to define a feature model in order to represent part of the subjective knowledge which underlies such reviews and attempts to describe salient characteristics of irony. They have established a model to represent verbal irony in terms of six categories of features: n-grams, POS-grams, funny profiling, positive/negative profiling, affective profiling, and pleasantness profiling. They built a freely available data set with ironic reviews from news articles, satiric articles and customer reviews, collected from amazon.com. They were posted on the basis of an online viral effect, i.e. contents that trigger a chain reaction in people. They used NB, SVM, and DT for classification purpose (illustrated with details in the next section). Their results with the three classifiers are satisfactory, both in terms of accuracy, as well as precision, recall, and F-measure.

4. Sentiment classification techniques

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach [69]. The *Machine Learning Approach* (*ML*) applies the famous ML algorithms and uses linguistic features. 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 *hybrid Approach* combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The various approaches and the most popular algorithms of SC are illustrated in Fig. 2 as mentioned before.

The text classification methods using ML approach can be roughly 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 difficult to find these labeled training documents.

The lexicon-based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. 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. This could be done by using statistical or semantic methods. There is a brief explanation of both approaches' algorithms and related articles in the next subsections.

4.1. Machine learning approach

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

Text Classification Problem Definition: We have a set of training records $D = \{X_1, X_2, \ldots, X_n\}$ where each record is

labeled to a class. The classification model is related to the features in the underlying record to one of the class labels. Then for a given instance of unknown class, the model is used to 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.

4.1.1. Supervised learning

The supervised learning methods depend on the existence of labeled training documents. There are many kinds of supervised classifiers in literature. In the next subsections, we present in brief details some of the most frequently used classifiers in SA.

4.1.1.1. 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*. Three of the most famous probabilistic classifiers are discussed in the next subsections.

4.1.1.1.1. Naïve Bayes Classifier (NB). The Naïve Bayes classifier is the simplest and most commonly used classifier. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs 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.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \quad (3)$$

$P(label)$ is the prior probability of a label or the likelihood that a random feature set the label. $P(features|label)$ is the prior probability that a given feature set is being classified as a label. $P(features)$ is the prior probability that a given feature set is occurred. Given the Naïve assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots.. * P(fn|label)}{P(features)} \quad (4)$$

An improved NB classifier was proposed by Kang and Yoo [36] to solve the problem of the tendency 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 SVM. The accuracy is improved in recall and precision compared to both NB and SVM.

4.1.1.1.2. Bayesian Network (BN). The main assumption of the NB classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. BN is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables, is specified for a model. In Text mining, the computation complexity of BN is very expensive; that is why, it is not frequently used [62].

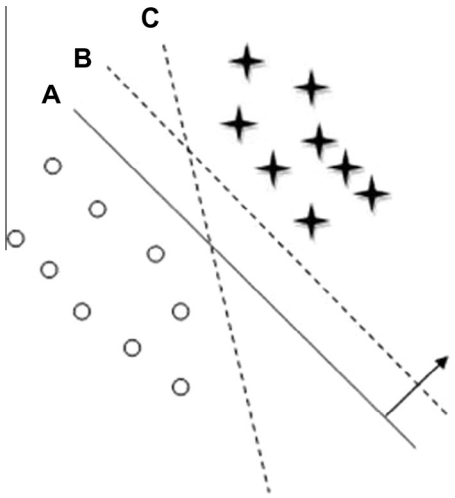


Figure 3 Using support vector machine on a classification problem.

BN was used by Hernández and Rodríguez [40] to consider a real-world problem in which the attitude of the author is characterized by three different (but related) target variables. They proposed the use of multi-dimensional Bayesian network classifiers. It joined the different target variables in the same classification task in order to exploit the potential relationships between them. They extended the multi-dimensional classification framework to the semi-supervised domain in order to take advantage of the huge amount of unlabeled information available in this context. They showed that their semi-supervised multi-dimensional approach outperforms the most common SA approaches, and that their classifier is the best solution in a semi-supervised framework because it matches the actual underlying domain structure.

4.1.1.1.3. Maximum Entropy Classifier (ME). The Maxent Classifier (known as a conditional exponential classifier) converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine 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 $C\{(featureset, label)\}$ pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))forlinlabels)} \quad (5)$$

ME classifier was used by Kaufmann [52] to detect parallel sentences between any language pairs with small amounts of training data. The other tools that were developed to automatically extract parallel data from non-parallel corpora use language specific techniques or require large amounts of

training data. Their results showed that ME classifiers can produce useful results for almost any language pair. This can allow the creation of parallel corpora for many new languages.

4.1.1.2. Linear classifiers. Given $\overline{X} = \{x_1 \ldots\ldots x_n\}$ is the normalized document word frequency, vector $\overline{A} = \{a_1 \ldots\ldots a_n\}$ is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as $p = \overline{A}.\overline{X} + b$, which is the output of the *linear classifier*. The predictor p is a separating hyperplane between different classes. There are many kinds of linear classifiers; among them is *Support Vector Machines* (*SVM*) [70,71] which is a form of classifiers that attempt to determine *good* linear separators between different classes. Two of the most famous linear classifiers are discussed in the following subsections.

4.1.1.2.1. 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. In Fig. 3 there are 2 classes x, o and there are 3 hyperplanes A, B and C. Hyperplane A 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.

Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories [72]. 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 [73].

SVMs are used in many applications, among these applications are classifying reviews according to their quality. Chen and Tseng [26] have used two multiclass SVM-based approaches: One-versus-All SVM and Single-Machine Multiclass SVM to categorize reviews. They proposed a method for evaluating the quality of information in product reviews considering it as a classification problem. They also adopted an information quality (IQ) framework to find information-oriented feature set. They worked on digital cameras and MP3 reviews. Their results showed that their method can accurately classify reviews in terms of their quality. It significantly outperforms state-of-the-art methods.

SVMs were used by Li and Li [57] as a sentiment polarity classifier. Unlike the binary classification problem, they argued that opinion subjectivity and expresser credibility should also be taken into consideration. They proposed a framework that provides a compact numeric summarization of opinions on micro-blogs platforms. They identified and extracted the topics mentioned in the opinions associated with the queries of users, and then classified the opinions using SVM. They worked on twitter posts for their experiment. They found out that the consideration of user credibility and opinion subjectivity is essential for aggregating micro-blog opinions. They proved that their mechanism can effectively discover market intelligence (MI) for supporting decision-makers by establishing a monitoring system to track external opinions on different aspects of a business in real time.

4.1.1.2.2. Neural Network (NN). Neural Network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector $overline X_{\mathrm{i}}$ which is the word frequencies in the ith 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: $p_i = A \cdot \overline{X}_i$. In a binary classification problem, it is assumed that the class label of $\overline{X}_i$ is denoted by y_i and the sign of the predicted function p_i 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. The training process is more complex because the errors need to be back-propagated over different layers. There are implementations of NNs for text data which are found in [74,75].

There is an empirical comparison between SVM and Artificial neural networks ANNs presented by Moraes and Valiati [53] regarding document-level sentiment analysis. They made this comparison because SVM has been widely and successfully used in SA while ANNs have attracted little attention as an approach for sentiment learning. They have discussed the requirements, resulting models and contexts in which both approaches achieve better levels of classification accuracy. They have also adopted a standard evaluation context with popular supervised methods for feature selection and weighting in a traditional BOWs model. Their experiments indicated that ANN produced superior results to SVM except for some unbalanced data contexts. They have tested three benchmark data sets on Movie, GPS, Camera and Books Reviews from amazon.com. They proved that the experiments on movie reviews ANN outperformed SVM by a statistically significant difference. They confirmed some potential limitations of both models, which have been rarely discussed in the SA literature, like the computational cost of SVM at the running time and ANN at the training time. They proved that using Information gain (a computationally cheap feature selection Method) can reduce the computational effort of both ANN and SVM without significantly affecting the resulting classification accuracy.

SVM and NN can be used also for the classification of personal relationships in biographical texts as presented by van de Camp and van den Bosch [47]. They marked relations between two persons (one being the topic of a biography, the other being mentioned in this biography) as positive, neutral, or unknown. Their case study was based on historical biographical information describing people in a particular domain, region and time frame. They showed that their classifiers were able to label these relations above a majority class baseline score. They found that a training set containing relations, surrounding multiple persons, produces more desirable results than a set that focuses on one specific entity. They proved that SVM and one layer NN (1-NN) algorithm achieve the highest scores.

4.1.1.3. 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 divide the data [76]. The condition or predicate is the presence or absence of one or more words. 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 [77]. *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 [78].

The decision tree implementations in text classification tend to be small variations on standard packages such as ID3 and C4.5. Li and Jain [79] 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 [17] 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 [16]. 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.

4.1.1.4. 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 [80]. 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 in [113].

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. Quinlan [76] 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.

4.1.2. 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 [81]. 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 Zhou [19] 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. In their work, they were able to identify domain-specific polarity words clarifying the idea that the polarity of a word may be different from a domain to the other. They worked on movie reviews and multi-domain sentiment data set from IMDB and amazon.com. They showed that their approach attained better performance than other weakly supervised sentiment classification methods and it is applicable to any text classification task where some relevant prior knowledge is available.

The unsupervised approach was used too by Xianghua and Guo [50] to automatically discover the aspects discussed in Chinese social reviews and also the sentiments expressed in different aspects. They used LDA model to discover multi-aspect global topics of social reviews, then they extracted the local topic and associated sentiment based on a sliding window context over the review text. They worked on social reviews that were extracted from a blog data set (2000-SINA) and a lexicon (300-SINA Hownet). They showed that their approach obtained good topic partitioning results and helped to improve SA accuracy. It helped too to discover multi-aspect fine-grained topics and associated sentiment.

There are other unsupervised approaches that depend on semantic orientation using PMI [82] or lexical association using PMI, semantic spaces, and distributional similarity to measure the similarity between words and polarity prototypes [83].

4.1.3. Meta classifiers

In many cases, the researchers use one kind or more of classifiers to test their work. One of these articles is the work proposed by Lane and Clarke [46]. 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

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 [56]. In their work, they were investigating whether and how twitter word of mouth (WOM) affects movie sales by estimating a dynamic panel data model. They used NB and SVM for classification purposes. Their main contribution was classifying the tweets putting into consideration the unique characteristics of tweets. They distinguished between the pre-consumer opinion (those have not bought the product yet) and post-consumer opinion (those bought the product). They worked on the benchmark movie reviews and twitter data. They have collected Twitter WOM data using Twitter API and movie sales data from BoxOfficeMojo.com. Their results suggest that the effect of WOM on product sales from Twitter users with more followers is significantly larger than that from Twitter users with fewer followers. They found that the effect of pre-consumption WOM on movie sales is larger than that of post-consumption WOM.

Another article compared many classifiers after applying a statistically Markov Models based classifier. This was to capture the dependencies among words and provide a vocabulary that enhanced the predictive performance of several popular classifiers. This was presented by Bai [15] who has presented a two-stage prediction algorithm. In the first stage, his classifier learned conditional dependencies among the words and encoded them into a Markov Blanket Directed Acyclic Graph for the sentiment variable. In the second stage, he used a meta-heuristic strategy to fine-tune their algorithm to yield a higher cross-validated accuracy. He has worked on two collections of online movie reviews from IMDB and three collections of online news then compared his algorithm with SVM, NB, ME and others. He illustrated that his method was able to identify a parsimonious set of predictive features and obtained better prediction results about sentiment orientations, compared to other methods. His results suggested that sentiments are captured by conditional dependencies among words as well as by keywords or high-frequency words. The complexity of his model is linear in the number of samples.

Supervised and unsupervised approaches can be combined together. This was done by Valdivia and Cámara [54]. They proposed the use of meta-classifiers in order to develop a polarity classification system. They worked on a Spanish corpus of film reviews along with its parallel corpus translated into English (MCE). First, they generated two individual models using these two corpora then applying machine learning algorithms (SVM, NB, C4.5 and other). Second, they integrated SentiWordNet sentiment corpus into the English corpus generating a new unsupervised model using semantic orientation approach. Third, they combine the three systems using a meta-classifier. Their results outperformed the results of using individual corpus and showed that their approach could be considered a good strategy for polarity classification when parallel corpora are available.

ML classifiers are used by Walker and Anand [44] to classify stance. Stance is defined as an overall position held by a person towards an object, idea or position [84]. Stance is similar to a point of view or perspective, it can be seen as identifying the "side" that a speaker is on, e.g. for or against political decisions. Walker and Anand [44] have classified stance that people hold and applied this on political debates. They utilized 104 two-sided debates from convinceme.net for 14 different debate topics and tried to identify the stance or attitude of the speakers. Their main target was to determine the potential contribution to debate side classification performance of contextual dialogue features. The main effect for context is when comparing their results with no context to those with context, where only 5 feature-topic pairs show a decrease from no context to context. They used SVM, NB and a rule-based classifier for classification purpose. They achieved debate-side classification accuracies, on a per topic basis, higher than the unigram baselines when using sentiment, subjectivity, dependency and dialogic features.

4.2. 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.

4.2.1. Dictionary-based approach

[85,86] 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 [87] or thesaurus [88] 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 [12] 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.

4.2.2. 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

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 [89]. 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 [90] was used as a sequence learning technique for extracting opinion expressions. It was used too by Jiaoa and Zhoua [23] 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 [25] 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 taxonomy-based approach for extracting feature-level opinions and map them into feature taxonomy was proposed by Cruz and Troyano [60]. This taxonomy is a semantic representation of the opinionated parts and attributes of an object. Their main target was a domain-oriented OM. They defined a set of domain-specific resources which capture valuable knowledge about how people express opinions on a given domain. They used resources which were automatically induced from a set of annotated documents. They worked on three different domains (headphones, hotels and cars reviews) from epinions.com. They compared their approach to other domain-independent techniques. Their results proved the importance of the domain in order to build accurate opinion extraction systems, as they led to an improvement of accuracy, with respect to the domain-independent approaches.

Using the corpus-based approach alone is not as effective as the dictionary-based approach because it is hard to prepare a huge corpus to cover all English words, but this approach has a major advantage that can help to find domain and context specific opinion words and their orientations using a domain corpus. The corpus-based approach is performed using statistical approach or semantic approach as illustrated in the following subsections:

4.2.2.1. 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 [91]. 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 [82].

The polarity of a word can be identified by studying the occurrence frequency of the word in a large annotated corpus of texts [83]. 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 [82].

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 Bose [31] 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 [66]. Cao and Duan [18] have used LSA to find the semantic characteristics from review texts to examine the impact of the various features. The objective of their work is to understand why some reviews receive many helpfulness votes, while others receive few or no votes at all. Therefore, instead of predicting a helpful level for reviews that have no votes, they investigated the factors that determine the number of helpfulness votes which a particular review receives (include both "yes" and "no" votes). They worked on software programs users' feedback from CNET Download.com. They showed that the semantic characteristics are more influential than other characteristics in affecting how many helpfulness vote reviews receive.

Semantic orientation of a word is a statistical approach used along with the PMI method. There is also an implementation of semantic space called *Hyperspace Analogue to Language* (*HAL*) which was proposed by Lund and Burgess [93]. Semantic space is the space in which words are represented by points; the position of each point along with each axis is somehow related to the meaning of the word. Xu and Peng [6] have developed an approach based on HAL called *Sentiment Hyperspace Analogue to Language* (*S-HAL*). In their model, the semantic orientation information of words is characterized by a specific vector space, and then a classifier was trained to identify the semantic orientation of terms (words or phrases). The hypothesis was verified by the method of semantic orientation inference from PMI (SO-PMI). Their approach produced a set of

weighted features based on surrounding words. They worked on news pages and used a Chinese corpus. Their results showed that they outperformed the SO-PMI and showed advantages in modeling semantic orientation characteristics when compared with the original HAL model.

4.2.2.2. 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 [86].

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 [7]. 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 [59]. 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 [38] 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. They have grouped products feature words into corresponding aspects by applying semantic methods. They have utilized sentence-based SA method to determine the polarity of each aspect in sentences taking into consideration the impact of adverbs of degree. They could find the weaknesses of the product, as it was probably the most unsatisfied aspect in customers' reviews, or the aspect which is more unsatisfied when compared with their competitor's product reviews. Their results expressed the good performance of the weakness finder.

4.2.3. 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 [94]. Moreo and Romero [37] 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 [35] 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.

4.2.3.1. Discourse information. The importance of discourse in SA has been increasing recently. Discourse information can be found either among sentences or among clauses in the same sentence. Sentiment annotation at the discourse level was studied in [95,96]. Asher et al. [95] have used five types of rhetorical relations: *Contrast*, *Correction*, *Support*, *Result*, and *Continuation* with attached sentiment information for annotation. Somasundaran et al. [96] have proposed a concept called opinion frame. The components of opinion frames are opinions and are the relationships between their targets [3]. They have enhanced their work and investigated design choices in modeling a discourse scheme for improving sentiment classification [97].

Rhetorical Structure Theory (RST) [98] describes how to split a text into spans, each representing a meaningful part of the text. Heerschop et al. [29] have proposed a framework that performed document SA (partly) based on a document's discourse structure which was obtained by applying RST on sentence level. They hypothesized that they can improve the

performance of a sentiment classifier by splitting a text into important and less important text spans. They used lexicon-based for classification of movie reviews. Their results showed improvement in SC accuracy compared to a baseline that does not take discourse structure into account.

A novel unsupervised approach for discovering intra-sentence level discourse relations for eliminating polarity ambiguities was presented by Zhou et al. [28]. First, they defined a discourse scheme with discourse constraints on polarity based on RST. Then, they utilized a small set of cue phrase-based patterns to collect a large number of discourse instances which were converted to semantic sequential representations (SSRs). Finally, they adopted an unsupervised method to generate, weigh and filter new SSRs without cue phrases for recognizing discourse relations. They worked on Chinese training data. Their results showed that the proposed methods effectively recognized the defined discourse relations and achieved significant improvement.

Zirn et al. [30] have presented a fully automatic framework for fine-grained SA on the sub-sentence level, combining multiple sentiment lexicons and neighborhood as well as discourse relations. They use Markov logic to integrate polarity scores from different sentiment lexicons using information about relations between neighboring segments. They worked on product reviews. Their results showed that the use of structural features improved the accuracy of polarity predictions achieving accuracy scores up to 69%.

The usefulness of RST in large scale polarity ranking of blog posts was explored by Chenlo et al. [61]. They applied sentence-level methods to select the key sentences that conveyed the overall on-topic sentiment of a blog post. Then, they applied RST analysis to these core sentences to guide the classification of their polarity and thus to generate an overall estimation of the document polarity with respect to a specific topic. They discovered that Bloggers tend to express their sentiment in a more apparent fashion in elaborating and attributing text segments rather than in the core of the text itself. Their results showed that RST provided valuable information about the discourse structure of the texts that can be used to make a more accurate ranking of documents in terms of their estimated sentiment in multi-topic blogs.

4.3. 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 [99] as a mathematical approach used for structuring, analyzing and visualizing data, based on a notion of duality called Galois connection [100]. 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 [101].

FCA and FFCA were used in many SA applications as presented by Li and Tsai [51]. In their work they proposed a classification framework based on FFCA to conceptualize documents into a more abstract form of concepts. They used training examples to improve the arbitrary outcomes caused by ambiguous terms. They used FFCA to train a classifier using concepts instead of documents in order to reduce the inherent ambiguities. They worked on a benchmark test bed (Reuters 21578) and two opinion polarity data sets on movie and eBook reviews. Their results indicated superior performance in all data sets and proved its ability to decrease the sensitivity to noise, as well as its adaptability in cross domain applications.

Kontopoulos et al. [55] have used FCA also to build an ontology domain model. In their work, they proposed the use of ontology-based techniques toward a more efficient sentiment analysis of twitter posts by breaking down each tweet into a set of aspects relevant to the subject. They worked on the domain of smart phones. Their architecture gives more detailed analysis of post opinions regarding a specific topic as it distinguishes the features of the domain and assigns respective scores to it.

Other concept-level sentiment analysis systems have been developed recently. Mudinas et al. [114] have presented the anatomy of pSenti. pSenti is a concept-level sentiment analysis system that is integrated into opinion mining lexicon-based and learning-based approaches. Their system achieved higher accuracy in sentiment polarity classification as well as sentiment strength detection compared with pure lexicon-based systems. They worked on two real-world data sets (CNET software reviews and IMDB movie reviews). They outperformed the proposed hybrid approach over state-of-the-art systems like SentiStrength.

Cambria and Havasi have introduced SenticNet 2 in [115]. They developed SenticNet 2; a publicly available semantic and affective resource for opinion mining and sentiment analysis; in order to bridge the cognitive and affective gap between word-level natural language data and the concept-level sentiments conveyed by them. Their system was built by means of sentic computing which is a new paradigm that exploits both Artificial Intelligence and SemanticWeb. They showed that their system can easily be embedded in real-world applications in order to effectively combine and compare structured and unstructured information.

Concept-level sentiment analysis systems have been used in other applications like e-health. This includes patients' opinion analysis [116] and crowd validation [117].

5. 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.

5.1. Emotion detection

Sentiment analysis is sometimes considered as an NLP task for discovering opinions about an entity; and because there is some ambiguity 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 [1].

It was argued by Plutchik [102] 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.

ED on a sentence level was proposed by Lu and Lin [13]. 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 approach was presented by Balahur et al. [45]. 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. [103] 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 [104–106].

Classifying sentences using fine-grained attitude types is another work presented by Neviarouskaya et al. [14]. 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 analysis.

Affect emotion words could be used as presented by Keshtkar and Inkpen [42] using a corpus-based technique. 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.

Ptaszynski et al. [50] 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 [49]. 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 fairy tales had a much wider distribution of emotional word densities than novels.

5.2. 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* [11].

Building Lexicon was presented by Tan and Wu [20]. 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 lexicon.

Building corpus was introduced by Robaldo and Di Caro [34]. 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

was able to cover high complexity while preserving good agreement among different people.

Boldrini et al. [41] have focused on the creation of EmotiBlog, a fine-grained annotation scheme for labeling subjectivity in nontraditional textual genres. 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.

Building Dictionary was presented by Steinberger et al. [43]. 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.

5.3. 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 [107], sentiment analysis [108], Named Entity recognition [109], part-of-speech tagging [110], ... etc.

In Sentiment Analysis; transfer learning can be applied to transfer sentiment classification from one domain to another [21] or building a bridge between two domains [22]. Tan and Wang [21] proposed an Entropy-based algorithm to pick out high-frequency domain-specific (HFDS) features as well as a weighting model which weighted the features as well as the instances. They assigned a smaller weight to HFDS features and a larger weight to instances with the same label as the involved pivot feature. They worked on education, stock and computer reviews that come from a domain-specific Chinese data set. They proved that their proposed model could overcome the adverse influence of HFDS features. They also showed that their model is a better choice for SA applications that require high-precision classification which have hardly any labeled training data.

Wu and Tan [22] have proposed a two-stage framework for cross-domain sentiment classification. In the first stage they built a bridge between the source domain and the target domain to get some most confidently labeled documents in the target domain. In the second stage they exploited the intrinsic structure, revealed by these labeled documents to label the target-domain data. They worked on books, hotels, and notebook reviews that came from a domain-specific Chinese data set. They proved that their proposed approach could improve the performance of cross-domain sentiment classification.

The Stochastic Agreement Regularization algorithm deals with cross-domain polarity classification [111]. It is a probabilistic agreement framework based on minimizing the Bhattacharyya distance between models trained using two different views. It regularizes the models from each view by constraining the amount by which it allows them to disagree on unlabeled instances from a theoretical model. The Stochastic Agreement Regularization algorithm was used as a base for the work presented by Lambova et al. [24] which discussed the problem of cross-domain text subjectivity classification. They proposed three new algorithms based on multi-view learning and the co-training algorithm strategy constrained by agreement [112]. They worked on movie reviews and question answering data that came from three famous data sets. They showed that their proposed work give improved results compared to the Stochastic Agreement Regularization algorithm.

Diversity among various data sources is a problem for the joint modeling of multiple data sources. Joint modeling is important to transfer learning; that is why Gupta et al. [32] have tried to solve this problem. In their work, they proposed a regularized shared subspace learning framework, which can exploit the mutual strengths of related data sources while being unaffected by the effects of the changeability of each source. They worked on social media news data that come from famous social media sites as Blogspot, Flicker and Youtube and also from news sites as CNN, BBC. They proved that their approach achieved better performance compared to others.

6. 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. The following graphs illustrate the number of the articles

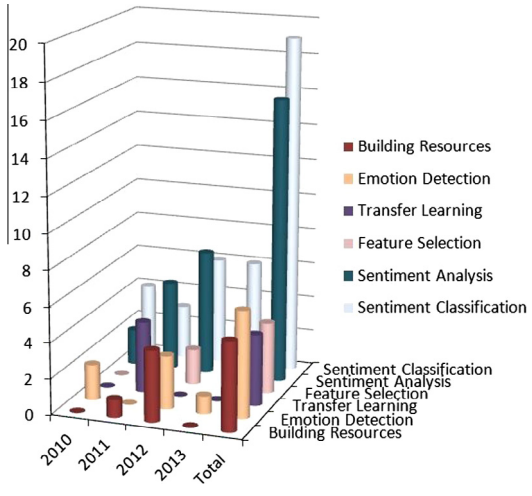


Figure 4 Number of articles tardifferent sentiment analysis tasks over years.

(which were presented in Table 1) through years according to their contributions in many criteria.

Fig. 4 illustrates 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 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. 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. Tsytsarau and Palpanas [1] have found that most of the work they presented was using ML approaches which means that, in the last few years, the researchers are heading toward the general analysis of texts. The uses of hybrid methods are not yet frequent because its computational complexity is higher.

Fig. 6 illustrates that the trend of researches has recently been to make a general categorization of sentiments rather

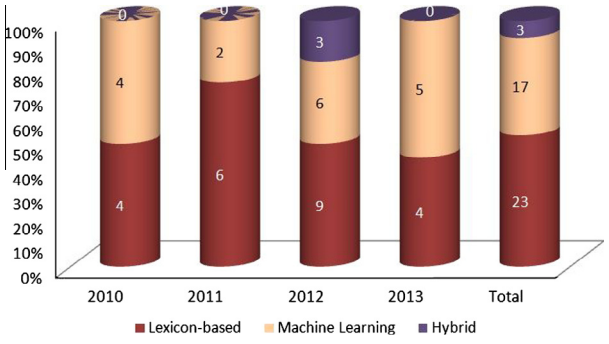


Figure 5 Number and percentage of articles according to the algorithmic approach over years.

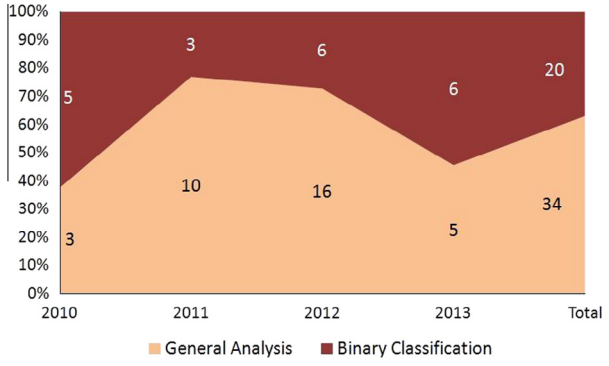


Figure 6 Number and percentage of articles according to the sentiment representation over years.

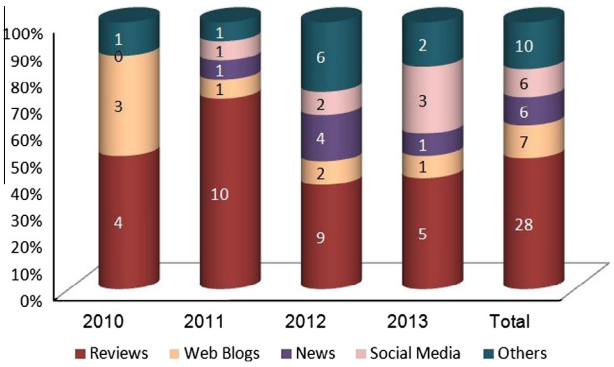


Figure 7 Number and percentage of articles targeting different text domains over years.

than making pos/neg classification in the overall count. It shows that the number and percentage of articles, in the last four years that make general classification, is greater than those who make pos/neg classification. In the last year, the number of articles is almost the same which means that the interest in pos/neg classification is still ongoing. However, this increase in percentage of general classification implies that the field of SA analysis is maturing. 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

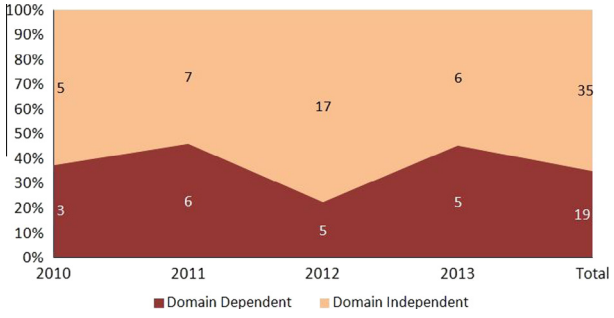


Figure 8 Number and percentage of articles targeting domain dependent and independent text over years.

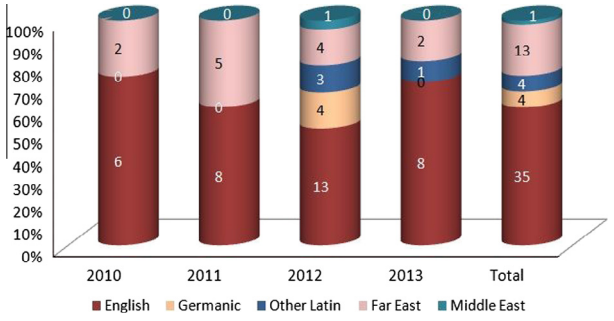


Figure 9 Number and percentage of articles using different natural languages over years.

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 as shown in Fig. 7. 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]. In Fig. 8, 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). Fig. 9 shows 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.

6.1. 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 [1] that few of the most famous data sets are in the field of SA. Table 2 illustrates some famous data sources and data sets which were used to accomplish the different tasks of SA. It can be noticed that ISEAR and Emotinet are used in the ED and BR articles. 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.

Table 2 Data sets.

References	Task	Data Set/Source
[41]	BR	ISEAR
[43]	BR	Sentiment dictionaries
[42]	ED	Live Journals Blogs, Text Affect, Fairy tales, Annotated Blogs
[45]	ED	ISEAR, Emotinet
[49]	ED	Enron Email corpus
[24]	TL	MPQA, RIMDB, CHES
[32]	TL	Blogspot, Flicker, youtube, CNN-BBC
[48]	FS	amazon.com
[12]	SA	automotvieforums.com
[18]	SA	CNETD
[25]	SA	amazon.com, epinions.com, blogs, SNS
[27]	SA	ebay.com, wikipedia.com, epinions.com
[31]	SA	amazon.com
[39]	SA	amazon.com
[55]	SA	Twitter
[15]	SC	IMDB
[19]	SC	IMDB, Amazon.com
[44]	SC	convinceme.net
[50]	SC	2000-SINA blog data set, 300-SINA Hownet lexicon
[51]	SC	Reuters 21578
[53]	SC	amazon.com
[56]	SC	Twitter
[57]	SC	Twitter
[60]	SC	epinions.com

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 [1] 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.

7. 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. Naïve Bayes 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 context of 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 Hsu-an-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. Knowl-Based 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;56:82–9.

[11] Montoyo Andrés, Martínez-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.

[16] Zhao Yan-Yan, Qin Bing, Liu Ting. Integrating intra- and inter-document evidences for improving sentence sentiment classification. Acta Automatica Sinica 2010;36(October'10).

[17] Yi Hu, Li Wenjie. Document sentiment classification by exploring description model of topical terms. Comput Speech Lang 2011;25:386–403.

[18] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach. Decis Support Syst 2011;50:511–21.

[19] He Yulan, Zhou Deyu. Self-training from labeled features for sentiment analysis. Inf Process Manage 2011;47:606–16.

[20] Tan Songbo, Wu Qiong. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Syst Appl 2011:12094–100.

[21] Tan Songbo, Wang Yuefen. Weighted SCL model for adaptation of sentiment classification. Expert Syst Appl 2011;38:10524–31.

[22] Qiong Wu, Tan Songbo. A two-stage framework for cross-domain sentiment classification. Expert Syst Appl 2011;38:14269–75.

[23] Jiao Jian, Zhou Yanquan. Sentiment Polarity Analysis based multi-dictionary. In: Presented at the 2011 International Conference on Physics Science and Technology (ICPST'11); 2011.

[24] Lambov Dinko, Pais Sebastião, Dias Gãel. Merged agreement algorithms for domain independent sentiment analysis. In: Presented at the Pacific Association for, Computational Linguistics (PACLING'11); 2011.

[25] Xu Kaiquan, Liao Stephen Shaoyi, Li Jiexun, Song Yuxia. Mining comparative opinions from customer reviews for competitive intelligence. Decis Support Syst 2011;50:743–54.

[26] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis Support Syst 2011;50:755–68.

[27] Fan Teng-Kai, Chang Chia-Hui. Blogger-centric contextual advertising. Expert Syst Appl 2011;38:1777–88.

[28] Zhou L, Li B, Gao W, Wei Z, Wong K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2001 conference on Empirical Methods in Natural Language Processing (EMNLP'11); 2011.

[29] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM'11); 2011.

[30] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP'11); 2011.

[31] Hu Nan, Bose Indranil, Koh Noi Sian, Liu Ling. "Manipulation of online reviews: an analysis of ratings, readability, and sentiments". Decis Support Syst 2012;52:674–84.

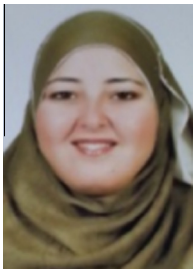
[32] Gupta Sunil Kumar, Phung Dinh, Adams Brett, Venkatesh Svetha. Regularized nonnegative shared subspace learning. Data Min Knowl Discov 2012;26:57–97.

[33] Duric Adnan, Song Fei. Feature selection for sentiment analysis based on content and syntax models. Decis Support Syst 2012;53:704–11.

[34] Robaldo Livio, Di Caro Luigi. OpinionMining-ML. Comput Stand Interfaces 2012.
[35] Caro Luigi Di, Grella Matteo. Sentiment analysis via dependency parsing. Comput Stand Interfaces 2012.
[36] Kang Hanhoon, Yoo Seong Joon, Han Dongil. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst Appl 2012;39:6000–10.
[37] Moreo A, Romero M, Castro JL, Zurita JM. Lexicon-based comments-oriented news sentiment analyzer system. Expert Syst Appl 2012;39:9166–80.
[38] Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Syst Appl 2012;39:10283–91.
[39] Min Hye-Jin, Park Jong C. Identifying helpful reviews based on customer's mentions about experiences. Expert Syst Appl 2012;39:11830–8.
[40] Ortigosa-Hernández Jonathan, Rodríguez Juan Diego, Alzate Leandro, Lucania Manuel, Inza Iñaki, Lozano Jose A. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. Neurocomputing 2012;92:98–115.
[41] Boldrini Ester, Balahur Alexandra, Martínez-Barco Patricio, Montoyo Andrés. Using EmotiBlog to annotate and analyse subjectivity in the new textual genres. Data Min Knowl Discov 2012;25:603–34.
[42] Keshtkar Fazel, Inkpen Diana. A bootstraping method for extracting paraphrases of emotion expressions from texts. Comput Intell 2012;vol. 0.
[43] Steinberger Josef, Ebrahim Mohamed, Ehrmann Maud, Hurriyetoglu Ali, Kabadjov Mijail, Lenkova Polina, Steinberger Ralf, Tanev Hristo, Vázquez Silvia, Zavarella Vanni. Creating sentiment dictionaries via triangulation. Decis Support Syst 2012;53:689–94.
[44] Walker Marilyn A, Anand Pranav, Abbott Rob, Fox Tree Jean E, Martell Craig, King Joseph. That is your evidence?: Classifying stance in online political debate. Decis Support Syst 2012;53:719–29.
[45] Balahur Alexandra, Hermida Jesús M, Montoyo Andrés. Detecting implicit expressions of emotion in text: a comparative analysis. Decis Support Syst 2012;53:742–53.
[46] Lane Peter CR, Clarke Daoud, Hender Paul. On developing robust models for favourability analysis: model choice, feature sets and imbalanced data. Decis Support Syst 2012;53:712–8.
[47] van de Camp Matje, van den Bosch Antal. The socialist network. Decis Support Syst 2012;53:761–9.
[48] Reyes Antonio, Rosso Paolo. Making objective decisions from subjective data: detecting irony in customer reviews. Decis Support Syst 2012;53:754–60.
[49] Mohammad SM. From once upon a time to happily ever after: tracking emotions in mail and books. Decis Support Syst 2012;53:730–41.
[50] Xianghua Fu, Guo Liu, Yanyan Guo, Zhiqiang Wang. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowl-Based Syst 2013;37:186–95.
[51] Li Sheng-Tun, Tsai Fu-Ching. A fuzzy conceptualization model for text mining with application in opinion polarity classification. Knowl-Based Syst 2013;39:23–33.
[52] Kaufmann JM. JMaxAlign: A Maximum Entropy Parallel Sentence Alignment Tool. In: Proceedings of COLING'12: Demonstration Papers, Mumbai; 2012. p. 277–88.
[53] Moraes Rodrigo, Valiati João Francisco, Gavião Neto Wilson P. Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert Syst Appl 2013;40:621–33.
[54] Martín-Valdivia María-Teresa, Martínez-Cámara Eugenio, Perea-Ortega Jose-M, Alfonso Ureña-López L. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Syst Appl 2013.
[55] Kontopoulos Efstratios, Berberidis Christos, Dergiades Theologos, Bassiliades Nick. Ontology-based sentiment analysis of twitter posts. Expert Syst Appl 2013.
[56] Rui Huaxia, Liu Yizao, Whinston Andrew. Whose and what chatter matters? The effect of tweets on movie sales. Decis Support Syst 2013.
[57] Li Yung-Ming, Li Tsung-Ying. Deriving market intelligence from microblogs. Decis Support Syst 2013.
[58] Ptaszynski Michal, Dokoshi Hiroaki, Oyama Satoshi, Rzepka Rafal, Kurihara Masahito, Araki Kenji, Momouchi Yoshio. Affect analysis in context of characters in narratives. Expert Syst Appl 2013;40:168–76.
[59] Pai Mao-Yuan, Chu Hui-Chuan, Wang Su-Chen, Chen Yuh-Min. Electronic word of mouth analysis for service experience. Expert Syst Appl 2013;40:1993–2006.
[60] Cruz Fermín L, Troyano José A, Enríquez Fernando, Javier Ortega F, Vallejo Carlos G. Long autonomy or long delay?' The importance of domain in opinion mining. Expert Syst Appl 2013;40:3174–84.
[61] Chenlo J, Hogenboom A, Losada D. Sentiment-based ranking of blog posts using rhetorical structure theory. In: Presented at the 18th international conference on applications of Natural Language to Information Systems (NLDB'13); 2013.
[62] Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data. Springer New York Dordrecht Heidelberg London: © Springer Science + Business Media, LLC'12; 2012.
[63] Yelena Mejova, Padmini Srinivasan. Exploring feature definition and selection for sentiment classifiers. In: Proceedings of the fifth international AAAI conference on weblogs and social media; 2011.
[64] Whitelaw Casey, Garg Navendu, Argamon Shlomo. Using appraisal groups for sentiment analysis. In: Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM); 2005. p. 625–31.
[65] Cover TM, Thomas JA. Elements of information theory. New York: John Wiley and Sons; 1991.
[66] Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R. Indexing by latent semantic analysis. JASIS 1990;41:391–407.
[67] Jolliffee IT. Principal component analysis. Springer; 2002.
[68] Griffiths Thomas L, Steyvers Mark, Blei David M, Tenenbaum Joshua B. Integrating topics and syntax. Adv Neural Inform Process Syst 2005:537–44.
[69] Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC'11; 2011. p. 88–99.
[70] Cortes C, Vapnik V. Support-vector networks, presented at the Machine Learning; 1995.
[71] Vapnik V. The nature of statistical learning theory, New York; 1995.
[72] Joachims T. Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In: Presented at the ICML conference; 1997.
[73] Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. Autom Rem Cont 1964:821–37.
[74] Ruiz M, Srinivasan P. Hierarchical neural networks for text categorization. In: Presented at the ACM SIGIR conference; 1999.
[75] Ng Hwee Tou, Goh Wei, Low Kok. Feature selection, perceptron learning, and a usability case study for text categorization. In: Presented at the ACM SIGIR conference; 1997.

[76] Quinlan JR. Induction of decision trees. Machine Learn 1986;1:81–106.
[77] Lewis David D, Ringuette Marc. A comparison of two learning algorithms for text categorization. SDAIR 1994.
[78] Chakrabarti Soumen, Roy Shourya, Soundalgekar Mahesh V. Fast and accurate text classification via multiple linear discriminant projections. VLDB J 2003;2:172–85.
[79] Li Y, Jain A. Classification of text documents. Comput J 1998;41:537–46.
[80] Liu Bing, Hsu Wynne, Ma Yiming. Integrating classification and association rule mining. In: Presented at the ACM KDD conference; 1998.
[81] Ko Youngjoong, Seo Jungyun. Automatic text categorization by unsupervised learning. In: Proceedings of COLING-00, the 18th international conference on computational linguistics; 2000.
[82] Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'02); 2002.
[83] Read J, Carroll J. Weakly supervised techniques for domain-independent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; 2009. p. 45–52.
[84] Somasundaran S, Wiebe J. Recognizing stances in online debates. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP; 2009. p. 226–34.
[85] Hu Minging, Liu Bing. Mining and summarizing customer reviews. In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04); 2004.
[86] Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING'04); 2004.
[87] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K. WordNet: an on-line lexical database. Oxford Univ. Press; 1990.
[88] Mohammad S, Dunne C, Dorr B. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
[89] Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'97); 1997.
[90] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning (ICML'01); 2001.
[91] Fahrni A, Klenner M. Old wine or warm beer: target-specific sentiment analysis of adjectives. In: Proceedings of the symposium on affective language in human and machine, AISB; 2008. p. 60–3.
[93] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav Res Methods 1996;28:203–8.
[94] Bolshakov Igor A, Gelbukh Alexander. Comput Linguis (Models, Resources, Applications) 2004.
[95] Asher N, Benamara F, Mathieu Y. Distilling opinion in discourse: a preliminary study, presented at the COLING'08; 2008.
[96] Somasundaran S, Wiebe J, Ruppenhofer J. Discourse level opinion interpretation, presented at the Coling'08; 2008.
[97] Somasundaran S, Namata G, Wiebe J, Getoor L. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In: Presented at the 2009 conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
[98] Mann W, Thompson S. Rhetorical structure theory: toward a functional theory of text organization. Text 1988;8, 243–28.
[99] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival, Reidel, Dordrecht-Boston; 1982, p. 445–70.
[100] Priss U. Formal concept analysis in information science. In: Presented at the annual review of information science and technology; 2006.
[101] Li S, Tsai F. Noise control in document classification based on fuzzy formal concept analysis. In: Presented at the IEEE International Conference on Fuzzy Systems (FUZZ); 2011.
[102] Plutchik R. A general psychoevolutionary theory of emotion. Emotion: Theory Res Exp 1980;1:3–33.
[103] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Recognition of affect conveyed by text messaging in online communication, presented at the Online Communities and Social Comput., HCII'07; 2007.
[104] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Compositionality principle in recognition of fine-grained emotions from text. In: Proceedings of the third international ICWSM conference; 2009.
[105] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. EmoHeart: automation of expressive communication of emotions in second life. Online Communities, LNCS 2009;5621:584–92.
[106] Neviarouskaya Alena, Tsetserukou Dzmitry, Prendinger Helmut, Kawakami Naoki, Tachi Susumu, Ishizuka Mitsuru. Emerging system for affectively charged interpersonal communication. In; Presented at the ICROS-SICE international joint conference, Fukuoka International Congress Center, Japan; 2009.
[107] Joachims T. Learning to classify text using support vector machines: methods, theory and algorithms. MA, USA: Norwell; 2002.
[108] Pang Bo, Lee Lillian. Opinion mining and sentiment analysis. Found Trends Inform Retriev; 2008.
[109] Zhang Tong, Johnson David. A robust risk minimization based named entity recognition system. In: Presented at the seventh conference on Natural language learning at HLT-NAACL; 2003.
[110] Ratnaparkhi Adwait. A maximum entropy model for part-of speech tagging. In: Proceedings of the conference on empirical methods in natural language processing, April 1996.
[111] Ganchev K, Graca J, Blitzer J, Taskar B. Multi-view learning over structured and non-identical outputs. In: Proceedings of the 24th conference on Uncertainty in Artificial Intelligence (UAI'08); 2008. p. 204–11.
[112] Wan X. Co-training for cross-lingual sentiment classification. in: Proceedings of the joint conference of the 47th annual meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL/IJCNLP'09); 2009. p. 235–43.
[113] Medhat W, Hassan A, Korashy H. Combined algorithm for data mining using association rules. Ain Shams J Electric Eng 2008;1(1).
[114] Mudinas Andrius, Zhang Dell, Levene Mark. Combining lexicon and learning based approaches for concept-level sentiment analysis. Presented at the WISDOM'12, Beijing, China; 2012.

[115] Cambria Erik, Havasi Catherine, Hussain Amir. SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the twenty-fifth international florida artificial intelligence research society conference; 2012.
[116] Cambria Erik, Benson Tim, Eckl Chris, Hussain Amir. Sentic PROMs: application of sentic computing to the development of a novel unified framework for measuring health-care quality. Expert Syst Appl 2012;39:10533–43.
[117] Cambria Erik, Hussain Amir, Havasi Catherine. Towards crowd Validation of the UK National Health Service. Presented at the Web Science Conf, Raleigh, NC, USA; 2010.



Walaa Medhat, is an Engineering Lecturer in School of Electronic Engineering, Canadian International College, Cairo campus of CBU. She got her M.Sc. and B.Sc. from Computers and Systems Engineering Department, Ain Shams University in 2008, 2002 respectively. Fields of interest: Text mining, Data mining, Software engineering, Programming Languages and Artificial Intelligence.

Ahmed Hassan, is an associate professor in the Computers and Systems Engineering Department, Ain Shams University since 2009. He is the executive Director of the Information and Communication Technology Project (ICTP), Ministry of Higher Education, Egypt. He got his Ph.D., M.Sc. and B.Sc. from Ain Shams University in 2004, 2000, 1995 respectively. He works also as the secretary of the IEEE, Egypt section since 2012. His research interests include Data Mining, Software Engineering, Programming Languages, Artificial Intelligence and Automatic Control.

Hoda Korashy, is a Prof. at Department of Computers & Systems, Faculty of Engineering, Ain Shams University, Cairo, Egypt. Major interests are in database systems, data mining, web mining, semantic web and intelligent systems.