Khan *et al. Complex Adapt Syst Model (2016) 4:2*
DOI 10.1186/s40294-016-0016-9

Complex Adaptive Systems Modeling
a SpringerOpen Journal

REVIEW

Open Access

Sentiment analysis and the complex natural language

Muhammad Taimoor Khan[1*], Mehr Durrani[2], Armughan Ali[2], Irum Inayat[3], Shehzad Khalid[1] and Kamran Habib Khan[4]

*Correspondence: taimoor.muhammad@gmail.com
[1] Bahria University, Shangrilla Road, Sector E-8, Islamabad, Pakistan
Full list of author information is available at the end of the article

Abstract

There is huge amount of content produced online by amateur authors, covering a large variety of topics. Sentiment analysis (SA) extracts and aggregates users' sentiments towards a target entity. Machine learning (ML) techniques are frequently used as the natural language data is in abundance and has definite patterns. ML techniques adapt to domain specific solution at high accuracy depending upon the feature set used. The lexicon-based techniques, using external dictionary, are independent of data to prevent overfitting but they miss context too in specialized domains. Corpus-based statistical techniques require large data to stabilize. Complex network based techniques are highly resourceful, preserving order, proximity, context and relationships. Recent applications developed incorporate the platform specific structural information i.e. meta-data. New sub-domains are introduced as influence analysis, bias analysis, and data leakage analysis. The nature of data is also evolving where transcribed customer-agent phone conversation are also used for sentiment analysis. This paper reviews sentiment analysis techniques and highlight the need to address natural language processing (NLP) specific open challenges. Without resolving the complex NLP challenges, ML techniques cannot make considerable advancements. The open issues and challenges in the area are discussed, stressing on the need of standard datasets and evaluation methodology. It also emphasized on the need of better language models that could capture context and proximity.

Keywords: Sentiment analysis, Machine learning, Sentiment orientation, Complex networks

Introduction

Sentiment analysis (Pang and Lillian 2008) is a type of text classification that deals with subjective statements. It is also known as opinion mining, since it processes opinions in order to learn about public perception. Sentiment analysis and opinion mining are the same, and are used interchangeably throughout the document. It uses natural language processing (NLP) to collect and examine opinion or sentiment words. SA is explained as identifying the sentiments of people about a topic and its features (Pang and Lillian 2008). The reason for the popularity of opinion mining is because people prefer to take advice from others in order to invest sensibly. Determining subjective attitudes in big social data is a hotspot in the field of data mining and NLP (Hai et al. 2014).

Manufacturers are also interested to know which features of their products are more popular in public, in order to make profitable business decisions. There is a huge repository of opinion content available at various online sources in the form of blogs, forums, social media, review websites etc. They are growing, with more opinionated content poured in continuously. It is, therefore, beyond the control of manual techniques to analyze millions of reviews and to aggregate them towards a rapid and efficient decision. Sentiment analysis techniques perform this task through automated processes with minimal or no user support. The online datasets may also contain objective statements, which do not contribute effectively in sentiment analysis. Such statements are filtered at pre-processing.

Opinion mining deals with identifying opinion patterns and presenting them in a way that is easy to understand. The outcome of sentiment analysis can be in the form of binary classification, such as categorizing opinions as recommended or not recommended. It can be considered as a multi-class classification problem on a given scale of likeness. Cambria et al. (2013) used common-sense knowledge to improve the results of sentiment analysis. The results can be presented in the form of a short summary generated from the overall analysis. Sentiment analysis has various sub streams including emotion analysis, trend analysis, and bias analysis etc. Its applications has outgrown from business to social, political and geographical domains. Sentiment analysis is applied to emails for gender identification through emotion analysis (Mohammad and Yang 2011). Emotion is applied to fairy tales to draw interesting patterns (Mohammad 2011). Considering text a complex network of words that are associated to each other with sentiments, graph based analysis techniques are used for NLP tasks.

Natural language processing

Opinion mining requires NLP, to extract semantics of opinion words and sentences. However, NLP has open challenges that are too complex to be handled accurately till date. Since sentiment analysis makes extensive use of NLP, it has this complex behavior reflected. The assumptions in NLP for text categorization do not work with opinion mining, as they are different in nature. Documents having high frequency of matching words may not necessarily possess same sentiment polarity. It is because, a fact in text categorization could be either correct or incorrect, and is well known to all. Unlike facts, a variety of opinions can be correct about the same product, due to its subjective nature. Another difference is that, opinion mining is sensitive to individual words, where a single word like NOT may change the whole context. The open challenges are negations without using NOT word, sarcastic and comparative sentences etc. The later section has a detailed discussion on NLP issues that affect sentiment analysis.

The subjective content from the online sources have simple, compound or complex sentences. Simple sentences possess single opinion about a product, while compound sentences have multiple opinions expressed together. Complex sentences have implicit meaning and are hard to evaluate. Regular opinions pertain to a single entity only, while comparative opinions have an object or some of its aspects discussed in comparison to another object. Comparative opinions can either be objective or subjective. An example of a subjective sentence having comparison is "The sound effects of game X are much better than that of game Y" whereas an example of objective sentence with comparison is

"Game X has twice as many control options as that of Game Y". Opinion mining expects a variety of sentence types, since people follow different writing styles in order to express themselves in a better way.

Sentiment analysis

The machine learning (ML) based techniques are supervised, semi supervised or unsupervised. The supervised techniques require labeled data, while the semi supervised techniques need manual tuning from domain experts. The unsupervised techniques make use of statistical analysis on large volume of data. ML techniques has a large feature set using Bag-of-words (BOW). Results are improved by pruning repetitive and low quality features. The opinion words are extracted to identify the polarity of opinion expressed for a feature. The performance of a classifier is measured through its effectiveness at the cost of efficiency. Effectiveness is calculated as precision/recall and F-measure, which are measurements of relevance.

Sentiment analysis can also be considered as a complex network. It consists of nodes and edges joining them. Many complex systems from a variety of domains are represented as network including environmental modeling (Niazi et al. 2010), business systems (Aoyama 2002), wireless sensors, and ad-hoc networks (Niazi and Hussain 2009). Networks are rich in information, having a range of local and global properties. Text corpora can be used with words as nodes and edges representing the structural or semantic association between them. The adjacent nodes sharing a link are closely associated and directly affect each other through the weight of the link they share. Representing text as complex network, various properties like centrality, degree distribution, components, communities, paths etc. can be used to explore the data thoroughly. Through multi-partite graphs, nodes can be distributed among various clusters with inter-cluster edges only. It separates different types of entities discussed in comparison. Entities are linked to their respective aspects/features and then to the sentiments associated. The sentiments can be linked with the reasons shared in support of those sentiments.

Data sources

Opinion mining has diverse subjective data sources that are available online. They cover a large number of topics and are up-to-date with current issues. Introduction of Web2.0 in the last decade has enabled people to post their thoughts and opinions on a range of topics. The data produced online is growing all the time produced by people from different backgrounds (Katz et al. 2015). Opinion mining makes use of this data generated by millions of users all over the world. According to Business Week survey in 2009, 70 % of the people consult online reviews and ratings to make a purchase. Comscore/The Kelsey group in 2007 reported that 97 % of the people who made purchases based on online reviews, found them to be honest.

The user generated subjective content is of value to be assessed and summarized for prospective customers. These online data sources are in the form of blogs, reviews and social media websites. The popularity of blogging is on the rise, where people from different walks of life express their opinions about various entities and events and get comments on them. At times, it leads to a form of discussion among the author and various users commenting on them. A detailed analysis on blogging styles of authors, as they

follow their own unique approaches for expressing their feelings is provided in (Chau and Xu 2007). Blogs contain opinions about various products, services, their features, packages and promotions. Most of the online studies on opinion extraction use blogs as datasets (Qiang and Rob 2009) to perform detailed analysis.

There are professional review websites providing customers' feedbacks, used for sentiment analysis. E-commerce websites allow customers to comment on their products. Social media is another popular medium of sharing information among like-minded people. Here, a variety of subjects are discussed where people express their opinions, based on their own experience. Social media websites have a very complex structure for extracting information having user opinions. They allow users to express their views through sharing articles and other media sources as an external link. Twitter, also referred to as microblogging, has the problem of reviews being too short and at times miss the context.

This review article is organized into the following divisions. Section 2 reviews the Sentiment analysis techniques and the NLP issues. Section 3 provides a discussion on the review studied and Sect. 4 list the application areas for sentiment analysis. Section 5 has concluded the study to important issues drawn from the study. Section 6 has distribution of the work carried out by the authors.

Review

The sentiment analysis techniques categorize reviews into positive and negative bins or multiple degrees of it. The social data can be analyzed at three different levels i.e. user data, relationship data and content (Tang et al. 2014). In survey (Guellil and Boukhalfa 2015) these categories are further elaborated. Recommender systems are extended to support textual content using knowledge (Tang et al. 2013). In our previous work (Khan and Khalid 2015) sentiment analysis is highlighted to address health care problems from the view point of a user. The issues faced in SA also depend on the data sources and nature of analysis required. An important aspect of social data analysis is the identification of sentiments and sentiment targets (Tuveri and Angioni 2014; Zhang and Liu 2014). Opinion mining also consider the additional features of opinion holder and time. Sentiment analysis techniques can be separated into three groups: supervised, semi-supervised and unsupervised techniques.

The supervised techniques are the machine learning classifiers. They are more accurate, however, need to be trained on a relevant domain. The unsupervised statistical techniques do not require training. They are efficient in dynamic environment but at the cost of accuracy. Sentiment analysis techniques analyze opinion datasets to generate a general perception that people have about a product. The classification of sentiments in a review document is performed through identifying and separating all the positive and negative opinion words. Considering the strength of these words, along with their polarity, helps in multi-class classification. Machine learning classifiers such as Naive-Bayes, k-nearest neighbor and centroid based classifier etc., are successfully used for this purpose. Semantic orientation based techniques used for opinion mining are Lexicon based and statistical analysis. Lexicon based technique works with individual words while statistical analysis incorporates words co-occurrence using point wise mutual information (PMI) and latent semantic analysis (LSA). Semi-supervised techniques start with a small

set of opinion words from the given domain, and expand on it. More opinion words are explored by querying the starting seeds. The newly found words are queried again to find more words until no new words are returned. Orientation of the opinion word form the basis for classification. Other attributes used are frequency of occurrence, location and co-occurrence with other words. The taxonomy of these approaches is shown in Fig. 1.

Sentiment classification

These are the machine learning classifiers used for sentiment analysis. They can be applied to text documents at three levels for analysis. A document level approach, which studies the whole document as a single entity is appropriate for text categorization. However, document level approach is not viable for sentiment analysis with documents having multiple opinions. Therefore, sentiment analysis is performed extensively at sentence or word level. Word level analysis is also known as sentiment level analysis. ML techniques suits sentiment analysis as the data is in abundance and there is obvious presence of patterns (Schouten and Frasincar 2015). The classifiers are trained on label dataset having samples representing all classes. A test dataset is used to evaluate the performance of the classifiers for the given task. Let the set of documents as $\{D = d_1,\ldots,d_n\}$, and set of classes labeled as $\{C = c_1,\ldots,c_n\}$, then the task is to classify document d_i in D with a label c_i in C. This task can be performed using supervised classifiers. The more frequently used classifiers for sentiment analysis are discussed below.

Naïve Bayes

Naive Bayes (NB) classifier is extensively used for text classification. It learns from a training dataset of annotated feature vectors, with labels as positive and negative (in case of binary classification). The probability of a feature vector is calculated with each label using the annotated training dataset. The feature vector is assigned a label that has highest probability for it. If this information is preserved, it can be used to show confidence in a label for a feature vector. In further modifications of NB a fuzzy region is defined in which feature vectors hold both labels with a certain level of confidence. Text data normally have high dimensional feature vectors. Therefore, the process of calculating probability is repeated for each feature vector, and then all the probabilities contribute towards the final decision. The feature set is represented as $F = f_1, f_2\ldots f_m\}$, where probability of a document belonging to a class shown as:

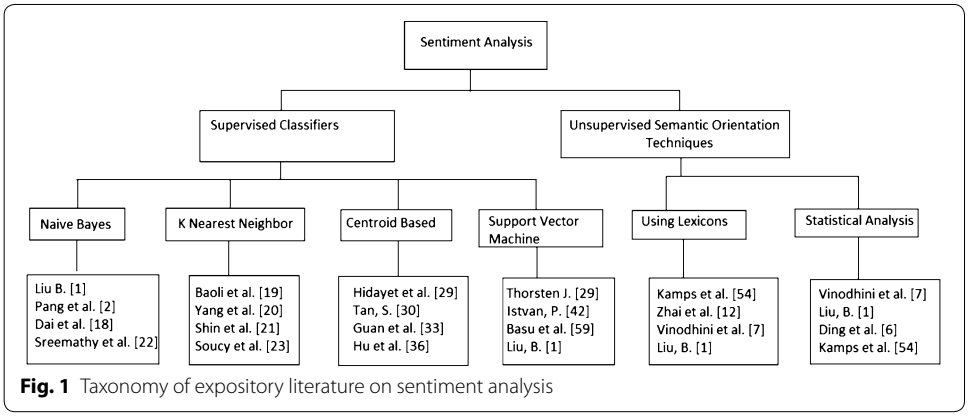

Fig. 1 Taxonomy of expository literature on sentiment analysis

$$P(c_i|d_j{}^*) = \frac{p(c_i)(\prod_{i=1}^{m} p(f_i|c_i))}{p(d_j^*)} \quad (1)$$

Shows the probability of a document d_j represented by its vector d_j^* belonging to a class *ci*. It is the product of probabilities for all the features in the feature set. The document vector d_j^* is assigned to a class ci in order to maximize $P\left(c_i \middle| d_j^*\right)$. The logarithm of probabilities are summed up to classify an opinion document. It is preferred over product of probabilities to avoid underflow. It addresses the missing value problem as well. Slack variables add smoothing effect against noisy data. Weights can also be assigned to features which define their contribution towards the classification. It is a biased approach, where prominent features are given high weights to play a major role in choose a sentiment label.

Naive Bayes works on the assumption that all the sentences of a review document are opinion sentences. It also assumes that features of a document are independent of each other. Despite of this unrealistic assumption, Naïve Bayes is very successful and is used in various practical applications. The assumption of treating features as independent of each other makes Naive Bayes highly efficient (Dai et al. 2007). Although, Naive Bayes classifier is simple, yet it is effective because of its robustness to irrelevant features. It performs well in domains with many equally important features. It is considered to be more reliable for text classification and sentiment analysis. The accuracy of the classifier improves with pre-processing noise. It also used as transfer learning when trained on a dataset similar to the target dataset.

Nearest neighbor

k-nearest neighbor classifier has been frequently used in literature for text classification. It considers the labels of k nearest neighbors to classify a test document. A special case of the k-NN problem is typically referred to as classimbalance problem identified in (Yang and Liu 1999). Classes with more training data have higher influence to predict same label for the new document. There are fewer chances of acquiring a class label if that class has fewer training examples. (Li et al. 2003) catered this problem by using variable value of *k* for each class. Thus, the class having more training data will have higher value of k as compared to the one having few samples. This solution is helpful in online classification, where there is time constraint on trying different values of *k*.

A study on performance of k-NN using pre-processed dataset is conducted in (Shin et al. 2006) claiming 10 % improvement when noise and outliers are filtered out. An optimum value is chosen as threshold to separate regular data from noise. Sentiment analysis is performed with a reduced set of feature vector in (Sreemathy and Balamurugan 2012) to avoid the curse of dimensionality. Accuracy of the model improves as irrelevant features were removed. Features are assigned weights to vary their contribution towards decision making. Weights are extracted from probability of information in documents across different categories. Tree-fast k-NN is introduced as fast kNN model (Soucy and Mineau 2001). This tree based indexing of retrieval system improves the accuracy of k-NN in distance calculation. Its effective against large feature sets. The order of features and their thresholds are identified from within the training data. k-NN has promising

results in sentiment analysis; however, it is more susceptible to noise and high dimensional feature set. Therefore, more of the work in k-NN for text classification has focused on feature selection and reduction techniques as they are the driving factors of k-NN's performance.

Centroid based

Centroid based (CB) classifier calculates centroid vector or prototype vector for each class in the training dataset. Centroid vector is the central point of the class and may not represent an actual training data. The distance of each test document is calculated with the prototype vector of the class and is classified based on similarity with it. Its performance depends on the chosen centroid vectors. It is efficient since time and space complexities are proportional to the number of classes rather than training documents. To double the training data reverse of reviews are generated in (Xia et al. 2015) by inverting the sentiment terms and their labels. Using both sets of training data with Mutual Information (MI) the results were improved when only selected reviews were inverted. External dictionary WordNet is used to generate inverse for sentiment terms, however, pseudo-antonyms can be generated internally using the corpus.

ms, however, pseudo-antonyms can be generated internally using the corpus. A variety of approaches have been used for CB classifier. Rocchio algorithm calculates centroid to represent feature space of documents (Ana and Arlindo 2007; Tan 2007a, b). Centroid is computed through average of positive examples in (Han and Karypis 2000) and sum of positive cases i.e. the related training examples (Chuang et al. 2000). Normalized sum of positive vectors used in (Lertnattee and Theeramunkong2004), cosine similarity between the test document and the Centroid of a class (Hidayet and Tunga 2012). Centroid is used with inverse of class similarity as well improving the accuracy close to 100 % on the given dataset when characters are chosen as features instead of n-grams.

Centroid evaluation is sensitive to noise in the training dataset which affects the overall performance of the classifier. This shortfall is exposed when Centroid classifier is applied to a slightly different domain. The reason for this drawback is that some opinion words are domain dependent. They have different polarity or strength of polarity when used in a different domain. Smoothing techniques have being proposed in (Tan 2007a, b; Lertnattee and Theeramunkong 2006; Guan 2009) that minimizes the effect of noise in the dataset. (Chizi et al. 2009) defined a weighting scheme giving higher weights to explicit opinion words. Characters and special characters for feature selection are used in (Ozgur and Gungor 2009). The work in (Shankar and Karypis 2000; Tan et al. 2005) is focused on adjusting the value of centroid based with feedback looping, hypothesis margin and weight-adjustment respectively. They try to rectify class Centroid, if it is not calculated accurately. Centroid based classifier performs efficiently as it doesn't consider training data each time to decide a test document.

Support vector machine

Support vector machine classifier is used for text classification in various studies. It finds a separation among the data using the annotated training dataset. The margin of separation between classes, which is known as hyperplane, is used to classify the incoming data. The hyperplane should give maximum separation between the classes. It is

applicable even in the presence of high dimensional feature set representation. It classify based o hyperplane among classes. Like centroid-based, SVM also consider the hyperplane to classify a test document. (Brown et al. 1997) has compared SVM with artificial neural networks for text classification and has found it better. Since it has promising results in text classification, it also performs well for opinion mining. They have also claimed in (Brown et al. 1997) that SVM is better than Naive Bayes and decision trees classification algorithms. However, SVM consumes more resources at the training stage. Although, it is efficient with large feature set, Feldman et al.(2011) has shown that dimensionality reduction in feature set further improves the performance of SVM. It exhibits linear complexity and can scale up to a large dataset.

SVM has a limitation of over-reliance on selection of suitable kernel function. Kernel is calculated through Linear, Polynomial, Gaussian or sigmoid methods but they tend to be domain specific. Kernel functions that perform well for one domain may not repeat it for next. Its accuracy is also sensitive to number of training samples close to hyperplane. Slack variables are introduced to limit the impact of boundary samples by generalizing the classifier, known as soft margin classification. They also help to avoid over-fitting the training data.

Unsupervised techniques

The unsupervised sentiment analysis techniques do not require training data and rather rely on semantic orientation. They make use of lexicons to identify the positive or negative semantics of opinion words. The meaning of the word, expressed by its use in a context is called lexicon. An online or off-line dictionary is consulted for this purpose. Statistical analysis techniques are also unsupervised, identifying the orientation of sentiment words through statistical evaluations. They require large volume of data for high accuracy.

Languages consists of lexicons that are the words used for a particular sense, and a grammar that connect these lexicons. Part-of-speech rules are used to extract sentiment phrases from text document. Search engines are used to identify the orientation of sentiment words that are missing in the dictionary. Its polarity is identified through the nearby words brought by search engines. They purely rely on external sources and therefore cannot address the context. Lexicon based techniques perform well for general domains while statistical techniques addresses the context and are useful in specialized domains. The two types of approaches are discussed in detail.

Dictionary (Lexicon) based techniques

Lexicon based techniques extract opinion lexicons from the document and analyzes its orientation without the support of any training data. These techniques process the opinion words separately, ignoring the relationship between them. Lexicons refer to the semantic orientation. Lexicons are independent of the source data and therefore it does not fall for over-fitting. But context not addressed either in this approach (Katz et al. 2015; Cambria 2013). Search engines are used to find the meaning of unknown opinion lexicons. They are searched and the top N results are accepted to identify its orientation. The semantics of lexicons can be categorized as positive or negative with weights representing their strength. This approach struggles with lexicons having domain specific

polarity. For example, good has positive polarity in any type of domain but "heavy weight" has positive polarity for bike domain but negative for the domain of electronic devices.

In its simplest form, sentiment words are split into positive and negative as binary distribution. A more sophisticated approach has fuzzy lexicons, introducing a grey area between the two categories. These fuzzy lexicons exist in both the classes with a score associated to it, representing the strength of each label. Various manual and semi-automatic techniques can be used for building lexicons. Princeton University's WordNet is a popular lexicon source available for sentiment analysis. Dictionaries like WordNet, extracts synonyms and antonyms for the provided opinion words. Manual cleansing is employed to rectify the lists generated for the unknown sentiment words. These opinion words are used to classify a review as positive or negative.

Fixed syntactic patterns are also used for expressing opinions which are composed of part-of-speech (POS) tags. The basic idea of this technique is to identify the patterns in which words co-occur with each other and to exploit those patterns for understanding its semantic orientation. One example of such pattern is an adverb followed by an adjective. A more sophisticated approach was proposed by (Mohammad and Yang 2011), which used a WordNet distance based method to determine the sentiment orientation. The distance $d(t_1, t_2)$ between terms t_1 and t_2 is the length of the shortest path that connects them in WordNet, as shown in Eq. 2. The semantic orientation (SO) of an adjective term t is determined by its relative distance from two reference (or seed) terms good and bad. The polarity of opinion term t is resolved through eq.

$$SO(t) = \frac{d(t, bad) - d(t, good)}{d(bad, good)} \tag{2}$$

Statistics (Corpus) based techniques

Statistical analysis of large corpus of text can also be used to determine the sentiment orientation of words. Co-occurrence of words is evaluated without consulting any external support. Two methods are used for this purpose which are point wise mutual information (PMI) and latent semantic analysis (LSA). PMI method for co-occurrence is given as:

$$p(w_1, w_2) = \frac{p(w_1, w_2)}{p(w_1)\, p(w_2)} \tag{3}$$

where w_1 and w_2 refers to two words in a given sentence. The main concept behind PMI based techniques is that the semantic orientation of a word has a tendency of being closely related to that of its neighbors. Equation 3 gives the probability of words w_1 and w_2 to co-exist, based on the measure of degree of statistical dependence between the two. This approach is, however, implemented differently in LSA based techniques. In LSA, matrix factorization technique is used with singular value decomposition to demonstrate the statistical co-occurrence of words. More formally, this process can be specified as:

$$LSA(w) = LSA(w, \{+paradigms\}) - LSA(w, \{-paradigms\}) \tag{4}$$

where a word w is passed to LSA with positive and negative paradigms. LSA based techniques develop a matrix having rows as words and columns as sentences or paragraphs. Each cell possesses a weight corresponding to the relation of the word in row with the sentence or paragraph in columns. This matrix is decomposed into three matrices using singular value decomposition (SVD).

Complex challenges

Opinion mining is a relatively new area of research and there are open challenges that need to be answered. Some of the challenges are common to opinion mining in general while others are related to their own sources and context depending upon the domain of the dataset. These issues affect the performance of machine learning techniques, but it has little control on them. Figure 2 gives NLP challenges faced in sentiment analysis, distributing them into their logical groups. The groupings are based on the parsing level, at which these issues occur. The following sub section has detailed discussion on the NLP issues.

Document level

Document level NLP challenges are the ones that are faced at the document or review level. They deal in general with the review document or the reviewer style. It is common to find reviews that have the information about an object, given in an informal manner. Capitalization is over or under used. Spelling mistakes are ignored or words being shortened. It makes the analysis very difficult for the automatic techniques to identify features and associate them. The unknown words (shortened/miss spelled) are matched with similar words to identify the aspect or opinion words. Slang specific to a certain region are also occasionally used in reviews and discussions. Reviews having sarcastic expressions are the hardest to deal with. Even though they have the opinion words explicitly mentioned, they do not serve the purpose for which they are normally used. For example, *"*What an awesome phone! It stopped responding in few hours*"*. These types of expressions are quite frequent in political reviews.

There are some document level challenges that are specific to certain domains only. The opinion words can also be domain dependent, where they have different orientation depending upon the domain in which they are used. This problem arise in specialized domains e.g. medical or astronomy etc. The opinion words in this case cannot be evaluated correctly, without domain knowledge. The general opinion words, however, have same orientation irrespective of the domain in which they are used e.g. good, bad etc. Opinion data attained from platforms like blogs and forums face the problem of dealing with discussions. They allow their users to comment on reviews, which at times gets into the shape of a debate. In discussions, users may agree with each other on some points while disagree on others. They need to be tackled differently, as they require the flow of context to be maintained from comment to comment in a sequence. Although, they are tough to process, discussions are very informative in which authors not only show their liking or disliking, but also support it through reasoning.

Spamming has been an issue faced at multiple frontiers of online data and sentiment analysis is no different. In fact it is the most highlighted area of sentiment analysis, considering its popularity and impact for industries. Spammers post false reviews about

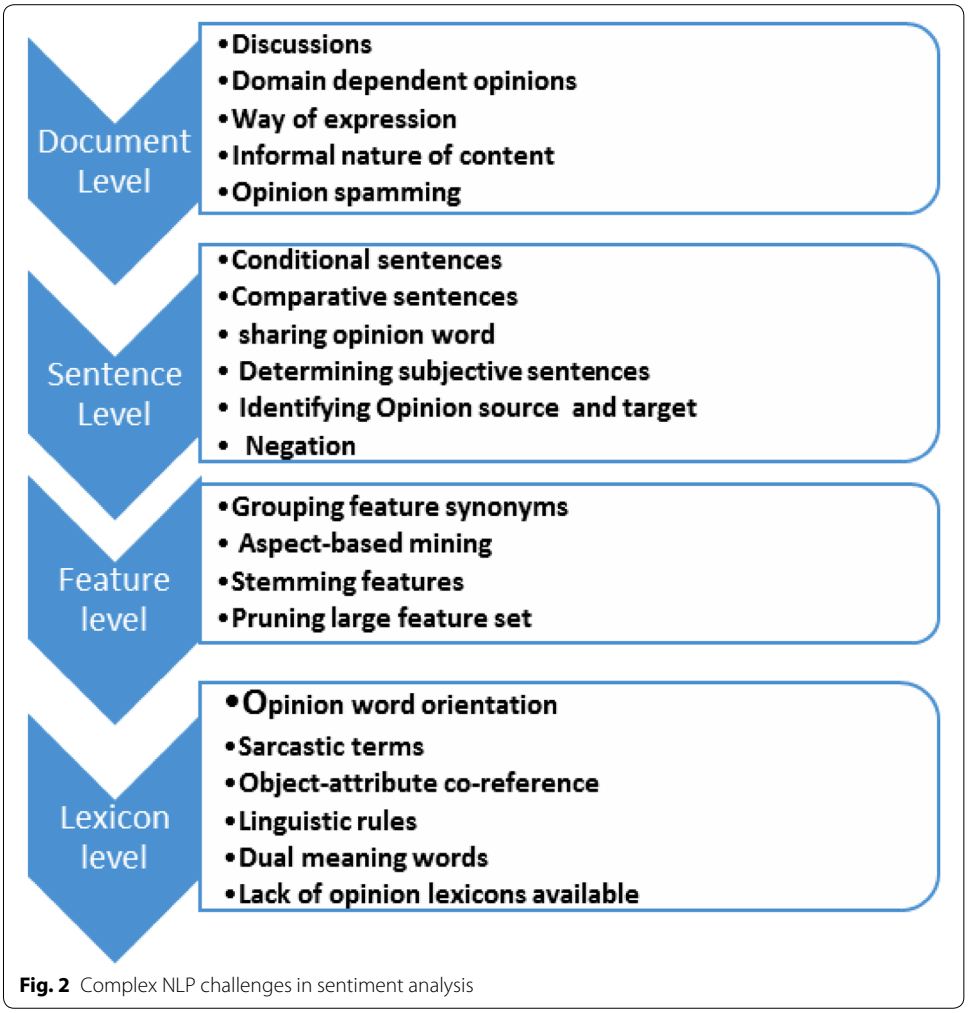

Fig. 2 Complex NLP challenges in sentiment analysis

products either to promote or demote it and if there are more numbers of such reviews it will affect the performance of all opinion mining techniques adversely. (Mukherjee et al. 2011; Mukherjee and Liu 2012) has stressed on the identification and filtering of spam reviews, prior to applying any mining techniques. Spam reviews can be written by individuals or commercial companies, dealing in such business. Spam opinions include a fake review where the author does not write his own feelings about a product. Other types of spamming include irrelevant, non-review content and advertising text etc. Psychological studies are used to identify spamming, that helps to find patterns when people lie. Meta information can also provide insight into it, as spammers normally tend to post more content in shorter time. This information can be very helpful if thoroughly studied to search for outliers. In (Lim et al. 2010; Kamps et al. 2004; Mohammad and Tony 2011) different machine learning techniques are used to detect spammers, who are then assigned a spamming behavior number to keep track of them.

Sentence level

These challenges are faced at the sentence level while parsing a review. They arise when the sentence expressing a review is not a simple sentence, that is expressing a single

product or feature with a single opinion word. Complex sentences can be comparative, conditional or having grouped opinions etc. (Narayanan et al. 2009) worked with sentences that are in the form of a question for the opinion audience. These sentences place a condition on an entity and are hard to parse and evaluate for a certain opinion. For example *"*if you are not happy with your notepad ++ code editor, try this new version of dreamweaver*"*. In this sentence author is positive about dreamweaver" whereas he/she does not say anything about the *"*notepad ++*"*. Without *"*if*"* it would be clearly a negative opinion of the *"*notepad ++*"* but now its inconclusive towards it.

Comparative sentences express a situation in which the opinion about one feature is discussed in comparison to another. In this situation, identifying the target of opinion words is very important, as there are more than one targets discussed. Secondly, aspects are to be associated to their respective products discussed in comparison to each other. For example *"*HP Laptops are stylish as compared to Dell and Sony*"*. In order to resolve this opinion, the information about the style of HP, Dell and Sony is crucial. In this case, without identifying the target features and their related opinions, the situation cannot be resolved (Jindal and Liu 2006). Must-link refer to the situation in which a single opinion word is shared by more than one features or entities. For example *"*My new office has attractive furniture, coloring and decoration*"*. In this sentence, the opinion attractive is being shared between three features of the entity office. Reviews are more opinion centered as compared to blogs and forums where the focus may deviate from the topic. In such discussions, not all the sentences can be evaluated for extracting opinions. There are certain sentences that do not bear any opinion which needs to be filtered. For example "Our team has strong batting line-up" is evaluative whereas "I am excited for our team's batting" is non-evaluative (Zhai et al. 2011). Mihalcea and Carlo 2009 has considered the problem of identifying words that make sense subjectively.

The opinion source and target identification is very important to classify opinions accurately. A target is the receiving entity of the opinion to which the opinion is entitled whereas; source is the person holding the opinion. Source identification is a concern when authors present the opinion of a third person. In the example *"*I bought a pen 2 days ago. It was such a nice pen. Its feel in hand is really cool. Its tip is soft and is very fluent. However, my mother was mad with me as I did not tell her before I bought it. She also thought the pen was too expensive, and wanted me to return it to the shop". The author himself/herself is the source of the first four lines whereas the source of the last two lines is author's mother, (Patella and Ciaccia 2009). Dealing with negation is also of high importance, since it overturns the orientation of the opinion words. For example, the sentence *"*I am not interested in this car*"* is negative. Negation words need to be dealt with a lot of care as not all occurrences of such words mean negation. For example, not in "not only but also" is not used for negation. Similarly negation can also be used without explicitly using any negation words like *"*Theoretically it takes care of the screen resolution*"*.

Feature level

These are the open issues faced at the features level in sentiment analysis. Natural languages are highly rich allowing a variety of words and phrases that could be used to express one's feelings. They consist of words that are used interchangeably for the same feature. If these synonyms are not identified, it will result in redundant features, which

at worse may have different opinion classification (Zhai et al. 2010). The features that are referred using different words, are grouped together and all their opinions are merged to get an aggregate opinion. For example *"*picture*"* and *"*photo*"* refers to the same feature of camera. Such features needs to be grouped based on synonyms otherwise they might be missed out or classified incorrectly. Feature stemming and pruning are also essential to identify similar opinion words and group them together. It reduces the set of opinion words that are used for classification. For example words like attraction, attractive, attracted, attracting are stemmed to the word attract and are considered as a single opinion word. Since all of the above words have same opinion with same orientation therefore, stemming them will enable the classifier to treat them as the same word attract. Reducing the feature set improves the performance of the classifiers. The opinion target may also be implicit, in which case the opinion is mentioned without explicitly giving the product or its feature. It normally happens when the target product or feature is already in discussion in previous sentence.

Complex products like phones, laptops cannot be recommended or not recommended as a whole. They have many features or aspects which need more in-depth study. If the opinion words are associated directly to the target domain, while by passing its feature, a lot of valuable information is missed. Aspect-based sentiment analysis (ABSA) consider aspect or features as the opinion targets. The features are associated to products to aggregate features opinions for products also. For example *"*Dell laptops have powerful batteries*"*. Such opinions need to be associated to the battery, which is a feature of *"*Dell laptop*"* and should not be referenced to it as a whole. This is also important for the reason that potential buyers are after certain strong features in the product considering their own situation and liking (Somprasertsri and Latitrojwong 2010).

Lexicon level

The problems faced at lexicon level are related to identifying the semantics of the word used. Dual meaning words and expressions depends on context of use. These words cannot be considered as positive or negative without having context knowledge. For example "the battery of this phone works for longer duration but the start-up takes longer too". Here *"*longer*"* is a positive opinion for battery backup but a negative opinion for start-up time (Ding and Liu 2007). The general opinion lexicon refers to opinion words like good, excellent, bad and poor etc. Most of the sentence and document based opinion mining use them as core of their techniques. There is only a small set of opinion lexicons publicly available. A universal opinion lexicon is required that would provide information on all such words (Qiu et al. 2011). A semi-automatic technique of dealing with this problem is to find synonyms and antonyms of initially given lexicon seeds passed to search engine. The process is repeated several times to explore as many opinion words as possible.

Ding and Liu (2010) refers to the problem of product-aspect co-reference. It is required in scenarios where products and their aspects along with the associated opinions, are not expressed in the same sentence. This is called opinion passage on aspect, expressing opinion as a group of consecutive sentences. For example, *"*I bought a Honda bike yesterday. It looks beautiful. I took it out for a ride yesterday. That was a great feeling*"*. In this example, It refers to bike whereas, that refers to ride which is a feature of

bike. This co-relation among products and their aspects need to be identified for associating products to their aspects in multiple sentences. It is called co-reference resolution problem was identified by Ott et al. (2011). Partially supervised clustering techniques are favored for this problem. Linguistic rules are very important to get sense of the opinions, rather than trusting the sentiment word only. Although, they are hard to apply, but are helpful in exploring the implicit meaning of words, for the sense in which they are used. For example "This car has good interior and not only that, the price is affordable too". In spite of having the word not, the meaning of the sentence is not inverted as is normally the case with the negation word.

Discussion

Sentiment analysis is an area of diversified research fields including machine learning, natural language processing, language identification and text summarization. Most of its issues are related to NLP which are quite complex and under research focus. The text obtained from reviews need to be classified into different languages, while working with a multi-lingual system. For each language, evaluative and subjective sentences are identified while others are discarded. Trimming is applied on the subject data for reducing the feature set which are further classified into either positive and negative (binary classification) or greater number of classes. Regression techniques are preferred for using multi-class problems. Table 1 provides a comparative analysis of the techniques used for sentiment anlaysis. Opinion mining has certain common grounds with text classification using techniques from Information retrieval. The NLP issues discussed affect all Sentiment analysis techniques, however, supervised techniques are more vulnerable to it. Opinion orientation has a context inclined towards psychology and linguistics. Complex networks can

Table 1 Comparison of the techniques and approaches included in the study

SNo.	Naïve Bayes	k-Nearest neighbor	Centroid
1	Yes	Yes	Yes
2	Yes	Yes	No
3	Yes	No	No
4	Word probability	Value of *k*	Centroid vector
5	Probability weights	Distance similarity	Vector distance
6	Simple and fast	Handle co-related features	Classify on vector distance
7	Assume feature independence	Sensitive to irrelevant features	Sensitive to noise
8	Yes	Too expensive	No
SNo.	**Support vector machine**	**Lexicon (dictionary) based**	**Statistical (corpus) based**
1	Yes	No	No
2	No	NA	NA
3	No	No	Yes
4	Kernel function	Word polarity	Feature matrix
5	Hyperplane	Word polarity	Word distance
6	Classify on hyperplane	Can identify new lexicons	Handle online data
7	Require more resources	Struggle with domain context	Conceptual document size
8	No	Yes	Yes

1 require training, *2* use training data to classify, *3* probabilistic approach, *4* driving factor, *5* similarity metric, *6* strength, *7* weakness, *8* support for streaming data

help in resolving context through preserving sequence and by associating words in a sentence, sentences in a paragraph and even paragraphs in an article. Sentiment analysis has its roads crossed with many different research areas and therefore, its problems are to be addressed with solutions coming from areas other than machine learning.

Sentiment analysis and opinion mining is out of its earlier stages and there is a strong need to standardize the datasets and evaluation methodology (Schouten and Frasincar 2015). Accuracy, area under the curve, precision/recall and F-measure are frequently used for evaluation. The bagof-words approach do not contain information about context and proximity and therefore, needs to be replaced with concept-centric approach. In survey (Cambria and White 2014) the need for bag-of-concept is emphasized and even bag-of-narratives was suggested. Sentiments also need to be contextualized and conceptualized (Gangemi et al. 2014). The work of Weichselbraun et al. (2014) is a stepping stone towards contextualizing sentiment analysis by integrating different semantic repositories i.e. WordNet, SentiWordNet, WordNetAffect etc. It also help to distinguish among specific aspects that were previously studied in isolation. SentiWordNet has low strength sentiments that are not contributing positively (Tsai et al. 2013).

The techniques developed for sentiment analysis needs to focus on the type of supporting application as they have different content style. Microblogging (twitter) and transcribed text is unstructured having more noise and therefore, lexicon-based techniques do not perform well (Katz et al. 2015). Similarly depending upon the nature of platform structural information can also be incorporated e.g. likes, share, retweets, hashtags etc. Ofek (2014) showed drop in accuracy when twitter data was used instead of Wall street journal content, even after including emoticons and hashtags (Ofek 2014). Machine learning techniques are more supportive to accommodate structural information e.g. meta-data as non-textual features (Katz et al. 2015). ML techniques depend on the feature set to which proximity and context based features can also be added. Transcribed text is used in Takeuchi and Yamaguchi (2014) and Cailliau and Cavet (2013) introducing new type of textual content. It also contain terms like "Emm" and "Aah" etc. that doesn't have any meaning but are used while speaking. Similarly sentences are left incomplete and grammar is ignored. This opens new avenues to these techniques to deal with this type of content.

Application areas

Previously if customers want to know about something, they would ask their friends and family while businesses would conduct surveys and polls. Sentiment analysis applications have spread to almost every possible domain, from consumer products, services, health care, and financial services to social events and political elections etc. Customers may analyze the feedback of various features of the product given by other customers in a way that would help in decision making. The sentiment analysis outcome of products and its features can be compared for competing products. Jaafar et al. (2015) consider big social data analysis as a concern for search engines and industries.

An application having opinion reason mining could be more helpful for both customers and companies towards making a sound decision. In opinion reason mining, not only the opinions about aspects are extracted but reason of the opinion is also extracted. It further helps manufacturers as their problems are highlighted. In (Zhang and Skiena

2010; Sakunkoo and Nathan 2009) expert investors use twitter moods to predict stock market. Blog and news sentiment were used to study trading strategies (Groh and Jan 2011). In (Stoyanov and Claire 2006) social influences in online book reviews were studied. Sentiment analysis is used to characterize social relations (Akkaya et al. 2009). (Mohammad and Tony 2011; Mohammad 2011) worked with emotion analysis on various sources. This is a very interesting study that can help to find what male and female customers look for, in a certain product that can be focused on.

Opinion mining on social media can have applications to rank celebrities, sportsmen and championships based on their popularity in public. It can be used to find the popularity of politicians prior to election etc. Influence analysis is performed in Nguyen et al. (2015) while Rabade et al. (2014) has discussed different influence indicators. There are very useful applications of opinion mining available that may be used online for finding the orientation of a text. Some of the notable ones are online message sentiment filtering, e-mail sentiment classification, web blog author's attitude analysis etc. Data leakage analysis is an emerging area that can be focused on in security systems (Katz et al. 2014).

Conclusion

Opinion mining has its boundaries extended from computer science to management sciences. Sentiment analysis, though recently introduced as in research focus for commercial and social content. A detailed analysis of the problem through ML based techniques has made it clear that SA and NLP has many open issues that are beyond the control of the methods in practice. Having close relevance to NLP, sentiment analysis faces NLP issues like co-reference resolution, negation handling, and word sense disambiguation etc., which add more difficulties due to their variation. However, it is also useful to realize that sentiment analysis is a highly restricted NLP problem because the system does not need to fully understand the semantics of each and every word. Complex network analysis has been popularly used for various problems and can produce useful patterns in subjective text. Knowledge-bases systems incorporate domain specific guidance from a knowledge source to improve results in specialized domains. More ML based solutions proposed, however, there is a strong need for considering solutions coming from different research domains. Machines generated data needs to be considered as meta-data along the content dimension for many useful purposes.

Authors' contributions

MTK performed critical analysis on the data, drafted the manuscript and concluded the concepts presented. Il helped with data acquisition about machine learning sentiment analysis techniques and drawing important interpretations from it. MD identified the key challenges in Natural Language processing and categorized them based on the level where they tend to occur. AA carried out a study to investigate applications of Sentiment analysis and how they can affect the future of business intelligence. SK supervised the whole activity, helped in drafting and revised the whole document critically for proof of concepts. KHK helped in improving the quality of content with the revised version of the manuscript. All authors read and approved the final manuscript.

Author details

[1] Bahria University, Shangrilla Road, Sector E-8, Islamabad, Pakistan. [2] COMSATS Institute of Information Technology, Kamra Road, Attock, Pakistan. [3] University of Malaya, Kuala Lumpur, Malaysia. [4] University of Haripur, Hattar Road, Haripur, Pakistan.

Acknowledgments

We would like to thank Bahria University, Islambad for providing the necessary environment and support to carry out this work.

Competing interests

The authors declare that they have no competing interests.

Received: 1 November 2015 Accepted: 18 January 2016
Published online: 03 February 2016

References

Akkaya Cem, Janyce Wiebe, Rada Mihalcea (2009) Subjectivity word sense disambiguation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing EMNLP
Ana C, Arlindo LO (2007) Semi-supervised single-label text categorization using centroid-based classifiers. ACM 844–851
Aoyama M (2002) A business-driven web service creation methodology, saint-w. IEEE
Bar-Haim R, Dinur E, Feldman R, Fresko M, Goldstein G (2011) Identifying and following expert investors in stock micro-blogs, In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)
Basu A et al (2003) Support vector machines for text categorization. In: Proceedings of the IEEE Hawaii International conference on system sciences
Brown MPS et al (1997) Support vector machine classification of microarray gene expression data. In: Proceedings of the National academy of science, p 262–267
Cailliau F, Cavet A (2013) Mining automatic speech transcripts for the retrieval of problematic calls. Comput Linguist Intell Text Process 83–95
Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2:15–21
Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research [review article]. Comput Intell Mag IEEE 9(2):48–57
Chau M, Xu J (2007) Mining communities and their relationships in blogs: a study of online hate groups. Int J Hum Comput Stud 65(1):57–70
Chizi B, Rokach L, Maimon O (2009) A survey of feature selection techniques. Encyclopedia of data warehousing and mining. 1888–1895
Chuang W, Tiyyagura A, Yang J, Giuffrida G (2000) A fast algorithm for hierarchical text classification. In: 2nd International Conference on Data Warehousing and Knowledge Discovery, p 409–418
Dai Qingliang Miao Qiudan Li Ruwei (2009) AMAZING: a sentiment mining and retrieval system. Expert Syst Appl 36:7192–7198
Dai W et al (2007) Transferring Naive Bayes Classifiers for Text Classification. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence 2007
Ding X, Liu B (2007) The utility of linguistic rules in opinion mining. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands
Ding X, Liu B (2010) Resolving object and attribute conference in opinion mining, COLING 2010. In: 23rd International Conference on Computational Linguistics, Proceedings of the Conference 2010
Feldman R et al (2011) The stock sonar-sentiment analysis of stocks based on a hybrid approach. In: Twenty-Third IAAI Conference
Gangemi A, Presutti V, Recupero RD (2014) Frame-based detection of opinion holders and topics: a model and a tool. Comput Intell Magaz IEEE 9(1):20–30
Groh G, Hauffa J (2011) Characterizing social relations via NLP based sentiment analysis. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media ICWSM, 2011
Guan H, Zhou J, Guo M (2009) A class-feature-centroid classifier for text categorization, In Proceedings of the 18th international conference on World wide web Madrid. 2009
Guellil I, Boukhalfa K (2015) Social big data mining: a survey focused on opinion mining and sentiments analysis. In: Programming and Systems (ISPS), 2015 12th International Symposium IEEE, p 1–10
Hai Z, Chang K, Kim JJ, Yang CC (2014) Identifying features in opinion mining via intrinsic and extrinsic domain relevance. IEEE Trans Knowl Data Eng 26(3):623–634
Han EH and Karypis G (2000) Analysis and experimental results. In: Principles of Data Mining and Knowledge Discovery, proceedings of the 4th European conference on centroid based document classification, p 424–431
Hidayet T, Tunga G (2012) A high performance centroid-based classification approach for language identification. Pattern Recognit Lett 33:2077–2084
Hu G, Jingyu Z, Minyi G (2009) A class-feature-centroid classifier for text categorization. ACM 201–210
Hull D (1994) Improving text retrieval for the routing problem using latent semantic indexing. In: Proceedings of SIGIR-94, p 282–289
Jaafar N, Al-Jadaan M, Alnutaifi R (2015) Framework for social media big data quality analysis. In New Trends in Database and Information Systems II. Springer International Publishing, p 301–314
Jindal N, Liu B (2006) Mining comparative sentences and relations. AAAI Press, Menlo Park, pp 1331–1336
Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the European conference of machine learning
Kamps J, Marx M, Mokken RJ, De Rijke M (2004) Using WordNet to measure semantic orientation of adjectives. In: Proceedings of 4th International Conference on Language Resources and Evaluation, p 1115-1118
Katz G, Elovici Y, Shapira B (2014) CoBAn: a context based model for data leakage prevention. Inform Sci 262:137–158
Katz G, Ofek N, Shapira B (2015) ConSent: context-based sentiment analysis. Knowl Syst 84:162–178
Khan MT, Khalid S (2015) Sentiment Analysis for Health Care. Int J Priv Health Inform Manag 3(2):78–91. doi:10.4018/IJPHIM.2015070105
Lertnattee V, Theeramunkong T (2004) Effect of term distributions on centroid-based text categorization. Inf Sci 158(1):89–115
Lertnattee V, Theeramunkong T (2006) Class normalization in centroid-based text categorization. Inf Sci 176(12):1712–1738

Li B, Yu S, Lu Q (2003) An improved k-nearest neighbor algorithm for text categorization. CoRR 0306099
Lim EP, Nguyen VA, Jindal N, Liu B, Lauw HW (2010) Detecting product review spammers using rating behaviors. In: Proceedings of ACM International Conference on Information and Knowledge Management CIKM
Machova K, Marhefka L (2014) Opinion classification in conversational content using N-grams. In: Recent Developments in Computational Collective Intelligence, Springer International Publishing, p 177–186
Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
Mihalcea R, Strapparava C (2009) The lie detector: Explorations in the automatic recognition of deceptive language, In: Proceedings of the ACL-IJCNLP 2009
Mohammad S (2011) Once upon a time to happily ever after: tracking emotions in novels and fairy tales. In: Proceedings of the ACL 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) 2011
Mohammad SM, Yang TW (2011) Tracking sentiment in mail: how genders differ on emotional axes. In: Proceedings of the ACL Workshop on ACL 2011, Workshop on Computational Approaches to Subjectivity and Sentiment Analysis WASSA 2011
Mukherjee A et al (2011) Detecting group review spam. ACM 93-94
Mukherjee A, Liu B (2012) Spotting fake reviewer groups in consumer reviews. ACM
Mukherjee A et al (2011) Detecting group review spam. ACM, p 93–94
Narayanan R et al (2009) Sentiment analysis of conditional sentences. ACL, p180–189
Nguyen DT, Hwang D, Jung JJ (2015) Time-frequency social data analytics for understanding social big data. In Intelligent Distributed Computing VIII. Springer International Publishing 223–228
Niazi M, Hussain A (2009) Agent-based tools for modeling and simulation of selforganization in peer-to-peer, ad hoc, and other complex networks. IEEE Commun Magaz 47(3):166–173
Niazi M and Hussain A (2011) Social network analysis of trends in the consumer electronics domain, consumer electronics (ICCE) 2011. IEEE international conference, p 219–220
Niazi MA, Siddique Q, Hussain A, Kolberg M (2010) Verification and validation of an agent-based forest fire simulation model. In: Proceedings of the 2010 Spring Simulation Multiconference
Ofek N, Rokach L, Mitra P (2014) Methodology for connecting nouns to their modifying adjectives. In Comput Linguist Intell Text Process 271–284
Ott M, Choi Y, Cardie C, Hancock JT (2011) Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics ACL
Ozgur L, Gungor T (2009) Text classification with the support of pruned dependency patterns. Pattern Recognit Lett. 31:1598–1607
Pang B, Lee L (2008) Opinion mining and sentiment analysis. Fund Trends Inf Ret 2:1–2
Patella M, Ciaccia P (2009) Approximate similarity search: a multi-faceted problem. J Discrete Algorithms 7:36–48
Poria S, Gelbukh A, Hussain A, Howard N, Das D, Bandyopadhyay S (2013) Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intell Syst 2:31–38
Qiu G et al (2011) Opinion word expansion and target extraction through double propagation. Comput Linguist l37:9–27
Rabade R, Mishra N, Sharma S (2014) Survey of influential user identification techniques in online social networks. In: Recent Advances in Intelligent Informatics. Springer International Publishing, p 359–370
Sakunkoo P, Sakunkoo N (2009) Analysis of social influence in online book reviews, Proceedings of 3rd International AAAI Conference on Weblogs and Social Media ICWSM, 2009
Schouten K, Frasincar F (2015) Survey on aspect-level sentiment analysis. 99
Shankar S, Karypis G (2000) Weight adjustment schemes for a centroid based classifier, Army High Performance Computing Research Center
Shin K, Abraham A, Han S (2006) Improving kNN text categorization by removing outliers from training set. Comput Linguis Intell Text Process 3878:563–566
Somprasertsri G, Latitrojwong P (2010) Mining feature-opinion in online customer reviews for opinion summarization. 16:938–955
Soucy P, Mineau GW (2001) A simple K-NN algorithm for text categorization. In: Proceedings of ICDM-01, IEEE International Conference on Data Mining, p 647–648
Sreemathy J, Balamurugan PS (2012) An efficient text classification using K-NN and Naive Bayesian. Engg Journals Publications, London
Stoyanov V, Cardie C (2006) Partially supervised conference resolution for opinion summarization through structured rule learning. In: Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP 2006
Takeuchi H, Yamaguchi T (2014) Text mining of business-oriented conversations at a call center. In Data Mining for Service, p 111–129
Tan S (2007a) Large margin dragpushing strategy for centroid text categorization. Expert Syst Appl 33(1):215–220
Tan S (2007b) An improved centroid classifier for text categorization. Expert Syst Appl 35(1–2):279–285
Tan S, Cheng X, Ghanem M, Wang B, Xu H (2005) A novel refinement approach for text categorization. CIKM 469-476
Tang J, Chang Y, Liu H (2014) Mining social media with social theories: a survey. ACM SIGKDD Explor Newslett 15(2):20–29
Tang J, Hu X, Liu H (2013) Social recommendation: a review. Soc Netw Anal Min 3(4):1113–1133
Tsai ACR, Wu CE, Tsai RTH, Hsu JYJ (2013) Building a concept-level sentiment dictionary based on commonsense. Knowl IEEE Intell Syst 2:22–30
Tuveri F, Angioni M (2014) An opinion mining model for generic domains distributed systems and applications of information filtering and retrieval. Springer, Heidelberg, pp 51–64
Vinodhini G, Chandrasekaran RM (2012) Sentiment analysis and opinion mining: a survey. Int J Adv Res Comput Sci Technol 2(6):2165–2172
Weichselbraun A, Gindl S, Scharl A (2013) Extracting and grounding context-aware sentiment lexicons. IEEE Intell Syst 28(2):39–46

Weichselbraun A, Gindl S, Scharl A (2014) Enriching semantic knowledge bases for opinion mining in big data applications. Knowl Syst 69:78–85
Xia R, Zong C, Hu X, Cambria E (2013) Feature ensemble plus sample selection: domain adaptation for sentiment classification. IEEE Intell Syst 28(3):10–18
Xia R, Xu F, Zong C, Li Q, Qi Y, Li T (2015) Dual sentiment analysis: data expansion by creating reversed reviews
Yang Y, Liu X (1999) A Re-examination of text categorization methods. In: Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p 4–49
Zhai Z et al (2010) Grouping product features using semi-supervised learning with softconstraints, COLING 2010. In: 23rd International Conference on Computational Linguistics, Proceedings of the Conference 2010
Zhai Z et al (2011) Identifying evaluative sentences in online discussions, In: Proceedings of the 25th AAAI Conference on Artificial intelligence. AAAI 2011, San Francisco, USA
Zhang Qiang Ye Ziqiong, Law Rob (2009) Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst Appl 36:6527–6535
Zhang L, Liu B (2014) Aspect and entity extraction for opinion mining. Data mining and knowledge discovery for big data. Springer, Heidelberg, pp 1–40
Zhang W, Skiena S (2010) Trading Strategies to Exploit Blog and News Sentiment. In: Proceedings of ICWSM