ITC 1/49

Information Technology and Control
Vol. 49 / No. 1 / 2020
pp. 36-54
DOI /10.5755/j01.itc.49.1.23769

A Domain-Independent Automatic Labeling System for Large-Scale Social Data Annotation Using Lexicon and Web-Based Augmentation

Received 2019/07/04 | Accepted after revision 2019/12/31

crossref **http://dx.doi.org//10.5755/j01.itc.49.1.23769**

HOW TO CITE: Khatoon, S., Abu Romman, L., Hasan, M. (2020). A Domain-Independent Automatic Labeling System for Large-Scale Social Data Annotation Using Lexicon and Web-Based Augmentation. *Information Technology and Control*, 49(1), 36-54. https://doi.org//10.5755/j01.itc.49.1.23769

A Domain-Independent Automatic Labeling System for Large-Scale Social Data Annotation Using Lexicon and Web-Based Augmentation

Shaheen Khatoon, Lamis Abu Romman, Md Maruf Hasan

Department of Information Systems, King Faisal University, P.O. Box:400, Al-Ahsa, 31982, Saudi Arabia;
e: mails: {ssyed, laburomman, mhasan}@kfu.edu.sa

Corresponding author: ssyed@kfu.edu.sa

Recently, with the large-scale adoption of social media, people have begun to express their opinion on these sites in the form of reviews. Potential consumers are often forced to wade through a massive amount of reviews to make an informed decision. Sentiment analysis has become a fast and effective way to gauge consumers' opinions automatically. However, such analysis often requires a tedious process of manual annotation of extensive training examples or manually crafted lexicon to find Semantic Orientation (SO) of online reviews. In this paper, we present a method to automate the laborious process of labeling extensive textual data in an unsupervised, domain-independent, and scalable manner. The proposed method combines the lexicon-based and Web-based Pointwise Mutual Information (PMI) statistics to find the Semantic Orientation (SO) of opinion expressed in a review. Based on the proposed method, a system called Domain-Independent Automatic Labeling System (DIALS) has been implemented, which takes a collection of text from any domain as input and generates a fully labeled dataset without any manual intervention. The result generated can be used to track and summarize the online discussion and/or use to train any classifier in the next stage of development. The effectiveness of the system is tested by comparing its results with baseline machine learning and lexicon-based methods. Experiments on cross-domain datasets have shown that the proposed system consistently showed improved recall and accuracy as compared to baseline machine learning and lexicon-based methods.

KEYWORDS: Information retrieval; Sentiment analysis; Unsupervised learning.

1. Introduction

Recent advancements in the Internet, social media and mobile devices have changed the way how information is produced, transferred, and consumed. Business organizations are using various social sites such as Facebook and Twitter to interact with customers to provide various services. Consumers are increasingly using these sites to search for information and to make purchase decisions. As a result, a massive amount of user-generated content in the form of posts, blogs, comments, and reviews is available on various online sites and has established the connection between a product manufacturer and its customers. Business owners can use such information to gauge public opinion to improve their services. Users can use the information to tap into the wisdom of crowds for informed decisions making. However, it is often hard to wade through a massive amount of data to find valuable insight from the enormous amount of continuously changing data. This problem has raised a question of how to overcome information overload and provide a rich and coherent user experience. This question has opened a vibrant venue for mining and analyzing online reviews. Previous studies have shown that there is a need to continuously collect, monitor, analyze, summarize, and visualize relevant information from these reviews to derive actionable insights. Therefore, companies have started analyzing online discussion to perform analytics such as sentiment analysis, opinion mining [19], topic modeling, and trend analysis [9, 27] to get valuable insight to improve product and service according to customer expectations.

However, it is often hard and very challenging for the research community to come up with precise scientific and intelligent methods to analyze and find meaningful insight due to the inherent complexities of processing natural language. By considering the fact that in today's competitive business environment, one of the major concerns of each company is to understand their customer's satisfaction. To measure customer satisfaction, companies have been using different tools to automatically collect customers' feedback and categorize them into different polarities such as positive, negative, and neutral and then take appropriate action on time. Such a research problem is known as sentiment analysis, opinion mining, review mining, and opinion extraction [19]. Previously, much research has been done by industry and academia in this area [19, 20]; however, monitoring opinion sites to distill the information contained in them remains a challenging task due to the proliferation of different opinion sites and inherent complexities of Natural Language Processing (NLP).

In this research project, we have proposed a multistage generalized social media analytical framework to continuously monitor and analyze business-related activities to support the exploration of unstructured data and transforming it into an actionable business insight to facilitate various business-related applications. The framework aims to integrate appropriate natural language processing, information retrieval, advanced semantic, and machine learning technique to extract and represent valuable knowledge for automated inference and reasoning. As a part of the larger project, in this paper, we present a domain-independent automatic labeling system to find the semantic orientation of online reviews automatically.

Previously, a large amount of work has been done on sentiment analysis by mainly using two popular methods, i.e., lexicon-based and machine learning-based. Lexicon-based methods use a set of opinion words or phrases with known orientation to determine the sentiment orientation of unknown documents or sentences, however, due to the opposite orientation of the same words in different domains as well as unavailability of comprehensive opinion lexicon the task of sentiment analysis is still a challenge. More information on using the lexicon-based approach is available in Liu's book [19]. In machine learning-based techniques, the text is transformed into features to train the classifier. Many techniques used supervised learning algorithms such as decision tree, naïve based, deep learning, and Support Vector Machine (SVM) to classify the text into a positive or negative class [43]. For supervised classification algorithms, the classifier needs to train using labeled examples from the source domain and then use the classifier to label new examples from the target domain. Studies have shown that supervised learning techniques achieved better performance than lexicon-based approaches [30]. However, labeling training data is time-consuming and labor-intensive. Also, a classifier trained on one domain often performs poorly on another domain; therefore, for each domain manual labeling of thousands of examples set is required to train a classifier. Hence, supervised learning algorithms have difficulty in scaling up to a large number of applications.

On the other hand, lexicon-based approaches depend on the presence of predefined words in the lexicon with a known SO to determine the SO of the unseen sentence or a document. These approaches are often constrained by the level of richness of the underlying dictionaries, such as language constructs, rules, and usage patterns. It takes strenuous manual annotation to develop such a dictionary, which causes significant difficulties for existing lexicon-based methods. As a result, lexicon-based approaches suffer through low recall, by incorrectly classifying unseen text, if some words in a text are not found in a predefined lexicon.

In this research, we have proposed a bootstrapping approach to annotate and label the given dataset by integrating lexicon-based and Web-based PMI-statistics to compute the SO of a given review text iteratively. The integration of lexicon with the Web corpus improves the accuracy of classification by calculating the semantic orientation of words not available in the predefined dictionary using PMI measures. Hence, our iterative approach is capable of incrementally augmenting the original lexical database with unknown words along with their SO values. Furthermore, we also adapted the PMI measure in such a way that it can be used to dynamically build a phrasal database (phases with their SO values) iteratively as well. The current implementation of the Domain-Independent Automatic Labeling System (DIALS) works as follows:

In the first phase, the sentiment orientation of reviews is identified using the predefined lexicon. We have used sentiment lexicon provided by Liu Bing [19, 20] to identify document-level sentiment where each review is treated as a document. Hu and Liu [15] proposed an iterative algorithm solely using the synonyms and antonyms relationships available in the WordNet to populate their seeded lexicon. However, not every word and phrase can be resolved using such synset relationships. If such a case is encountered, Hu and Liu proposed that the particular word or phrase is simply discarded or labeled manually. Such an approach is not suitable for DIALS as it needs to find a sensible mechanism to deal with the unknown words to incrementally augment the initial lexical database continuously. To address this problem, we have used Web corpus to find the SO of unknown words using a modified measure as proposed in Turney [41]. When DIALS encounters an unknown word, it automatically constructs a query using the unknown word along with a set of positive and negative reference words using a search engine to estimate the PMI-like semantic orientation. Such queries try to find the co-occurrence of the unknown word in positive and negative context with a list of positive reference words randomly selected from predefined lexicon such as “excellent”, “awesome”, “wonderful” and negative reference word such as “poor”, “creepy”, “terrible”. We adopted the association of unknown words the positive and negative reference words by considering the fact that if unknown words frequently appear together with extreme positive or negative words, they are likely to have the same polarity as evident in Turney’s work [41]. The system then estimates the SO of unknown words based on pointwise mutual information between unknown and reference words. In the next step, the system calculates SO of each review by aggregating SO of each word in a review. The system also adds such unknown words with their SO value in the existing lexical database iteratively. In this way, when the system is deployed in a specific application domain, the initial lexicon would grow over time with domain-specific words and phrases, which would subsequently reduce the likelihood of using the Web corpus via the search engine (this would mitigate the performance overhead in labeling large datasets).

The advantage of the proposed system is that it does not need intensive linguistic analysis and manual labeling. Hence, it is a fully automated, unsupervised, and domain-independent since it does not require domain-specific labeled data to train any classifier for a new domain as it is done in a supervised learning approach. The proposed system can be deployed to any business application (without any further development), where automatic tracking of online discussion is required. The output of DIALS can be used in many NLP applications: for example, to produce a summary of opinions on underlying product or service such as, we can show how many reviews express negative opinions and how many reviews express positive opinions; how sentiment changes over time; to generate a structured summary from unstructured texts. The research community can also use the output of DIALS for quickly labeling a large collection of text from any domain to build supervised learning models in the next phase of development.

2. Related Work

Several studies have been done in the field of text mining, e.g., classifying text according to document source information, such as author and publisher [18, 36, 39]. Another related area is classifying documents according to the genre, where subjective features are used to categorize into distinct genres, such as “editorial”, “novel”, “news”, etc. [12, 16, 37]. Other explicitly attempts have been made on sentence subjectivity classification using features that indicate whether the sentence is subjective or objective [3, 7, 34]. While these techniques for genre classification and subjectivity analysis can be used to categorize documents that express an opinion, however, these techniques do not determine our classification task of finding the semantic orientation of opinion being expressed.

This section of this paper only reflects the most relevant research, which has been focused on the classification of text based on Semantic Orientation (SO). Two main approaches, i.e., lexicon-based and machine learning, have been commonly used in literature to identify SO of a document.

Lexicon-based approach [15, 38] uses a dictionary of opinion words to determines the polarity of opinion by using features of sentiment words or phrases in a document, also called unsupervised learning method. Opinion words are the words used to express positive or negative orientation, such as “excellent” and “poor”. Lexicon-based methods depend on the presence of predefined words in the lexicon with a known semantic orientation to determine the polarity of the unseen sentence or document. However, the presence of sentiment orientation of each word from the given text or sentence in the lexicon is not possible, due to the varied and changing nature of the language used in reviews from different domains. Therefore, lexicon-based approaches suffer through low recall.

Machine learning-based approaches consider sentiment analysis as a text classification problem, and any existing supervised learning method can be directly applied by using syntactic and/or linguistic features [21]. Most techniques use some form of supervised learning by applying different learning techniques such as Naïve Bayes, Decision Tree, Maximum Entropy (ME) and/or Support Vector Machine (SVM) [30, 43]. These techniques typically train sentiment classifiers using features such as term frequencies, Part of Speech (POS) tagging, semantic and syntactic features, etc. Pang et al. [30] took this approach to classify movie reviews into positive and negative classes. It has been observed that Naïve Bayes and SVM performed very well by using unigrams bag of words as features in classification. In subsequent studies, many more features and learning algorithms have been tried by the researchers. The key in subsequent research is to improve sentiment classification by effectively engineering feature sets, like in many supervised learning algorithms. For example, Gamon [13] used supervised learning to classify customer feedback, which is short and noisy as compared to reviews. They used deep linguistic features combined with feature reduction techniques to train the SVM classifier. These features included Part of Speech (POS) trigram, structural patterns in the phrase tree, POS combined with semantic relation, and logical features such as tense information and transitivity of predicates. Results showed that deep linguistic features improve the classification accuracy of noisy data as compared to using only surface-level features such as n-gram.

Similarly, Mullen and Collier [26] used enriched feature sets combined with n-gram to compute sentiment orientation of words. These additional features include computation of sentiment orientation of word and/or phrase using PMI [41], adjective value identification using three factors such as strong or weak, good or bad, active or passive and calculation of semantic orientation of the words or phrase that are within k words proximity or sentence proximity that mention the entities being reviewed. These additional features have shown some improvements over unigram but not shown a big difference in terms of accuracy.

Mejova and Srinivasan [22] explored various feature definition and selection strategies for sentiment polarity classification. In the first step, they tested term frequency versus binary weighting, negation-enriched features, n-grams, or phrases. Afterward, they moved to feature selection using frequency-based vocabulary trimming, part-of-speech, and lexicon selection. Results showed that for a lager dataset classifier trained on a small number of features outperformed the classifier trained on all features.

In the above machine learning approaches, the classifiers built on supervised learning methods achieved quite a high accuracy to correctly classify an unknown text [2, 5, 6, 8]. These approaches used a labeled dataset to train the classifier to predict the label of new unseen data, which is time-consuming and difficult. Moreover, supervised learning is highly domain-dependent, a classifier trains on one domain may perform very poorly in another domain. Therefore, supervised learning methods are not suitable for sentiment analysis on social sites, where virtually people can post about any domain. As compared to these approaches, our approach is fully automated and does not need any linguistic knowledge or human interaction to annotate or label the data. Moreover, it can be applied to any domain without the need to build training data for each domain separately due to the integration of a massive corpus in the form of a web search.

To avoid the tedious job of manual labeling, researchers have introduced various learning methods to automate or semi-automate the process of labeling example sets. A closely related work is introduced by Zhang et al. [44], where they proposed a hybrid method by using the lexicon and supervised learning to classify tweets into positive or negative. First, the lexicon-based method is used to label the initial tweet sets, and then the results are manually augmented with the sentiment indicators. In the next step, augmented results are used to tag the remaining tweets, then the next iteration starts. Afterward, labeled tweets are used to train the classifier model SVM; once the model is trained, it is used to identify the sentiment associated with new unseen tweets automatically. Another related study is reported in [42], which used a subjectivity lexicon by compiling a dictionary of subjective words and used rule-based classifiers to identify training data for sentence-level subjectivity classification. The classifier classifies a sentence if it contains two or more strong subjective clues; otherwise, it does not label the data. SO-CAL is yet another lexicon-based method proposed by Taboada et al. [38]. SO-CAL uses a dictionary of sentiment words and phrases with their associated orientation and strength to compute sentiment scores. However, constructing the lexicon covering all opinionated English words is practically infeasible. Hence, these systems suffer through low recall, if words from given texts are not available in the predefined lexicon. In the proposed method, we addressed this problem of a low recall by integrating the Web as a massive text corpus and the use of Web PMI to compute the semantic orientation of unknown words.

An integrated approach of the lexicon and self-learning method is proposed by Qiu et al. [33]. They first used a lexicon-based method to classify some reviews initially, and then more reviews are classified through negative/positive ratio control. Afterward, the supervised classifier is trained by taking some reviews classified in the first phase as training data. The advantage of this method is that it does not need any manual labeling; however, by relying only on the lexicon, the opinionated reviews with the words not available in lexicon classified as neutral. As compared to this, our proposed approach uses additional features to construct a set of queries to search co-occurrence of unknown words with positive and negative reference through Web search and iteratively add new words in the underlying predefined lexicon.

Tong [40] proposed a lexicon-based method to track online discussion about a movie and display a plot on the numbers of positive and negative sentiment over time scale. They prepared a specialized lexicon tagged with positive and negative phrases such as “wonderful visuals”, “great acting”, “uneven editing”. New reviews are classified by looking for specific phrases available in the lexicon. Each phrase must be manually added in the unique lexicon and manually tagged as positive or negative. The lexicon is domain-dependent (e.g., movies), and it must be rebuilt to apply for a new domain. In contrast, our system does not require any manual annotation and can be easily adaptable to any new domain without any further human intervention.

Hu and Liu [15] proposed a bootstrapping process to find the SO of a sentence. They created a small list of seed adjectives with positive and negative labels and used WordNet synset relationship to grow the seeded list. An iterative process is used to calculate the SO of a sentence by first mining the product features on which review is expressed and then extract adjectives words on those features. Each target adjective is then searched in Wordnet for the synset and antonym set; if the target adjective's synset/antonym set has known orientation in the seed list, then the orientation of the target adjective is set same as the synset or opposite of antonym set. The process continues until the SO of all target words is found.   Their iterative algorithm solely using the synonyms and antonyms relationships available in the WordNet to populate their seeded lexicon. For those adjectives that WordNet cannot recognize are discarded (or manually labeled). Such an approach is not suitable for DIALS as it needs to find a sensible mechanism to deal with the unknown words to incrementally augment the initial lexical database continuously.

We have found Turney's [41] used a unique way to compute the SO of a review using Web-based PMI statistics. He applied an unsupervised learning technique based on the mutual information between the document's phrases and the words "excellent" and "poor", where the mutual information is computed using statistics gathered by a search engine. The algorithm performs well, but it is designed for classifying the given review as positive or negative on the go; however, this work does not calculate the SO of words or phrases. To calculate the SO of unknown words, we used a variant of Turney's approach by randomly selecting positive and negative reference words from initial lexical database such as "excellent", "awesome", "wonderful" etc. for positive reference words and "poor", "creepy", "terrible" etc. for negative reference words. The system then issues a set of search engine queries to find the co-occurrence of an unknown these reference words. The SO of the unknown word is then calculated by using mutual information statistics gathered by the search engine and inserted into the initial lexical database. Hence the initial lexical database grows over time with more domain-specific words and phrases. Calculating the SO of words and phrases is paramount in our case since our purpose is to create a single lexical database used to classify text from large scale social sites instead of using Web-based PMI statistics to classify text as Turney did.

Furthermore, Turney's original approach requires Web search for any review with an unknown word, which admittedly creates a performance bottleneck. In contrast, DIALS adaptation of PMI-based SO estimation is to augment the lexical database not to use for text classification. Therefore, as compared to Turney's work, our adaptation of PMI measure to estimate SO for unknown words reduces the time needed to process new social datasets.

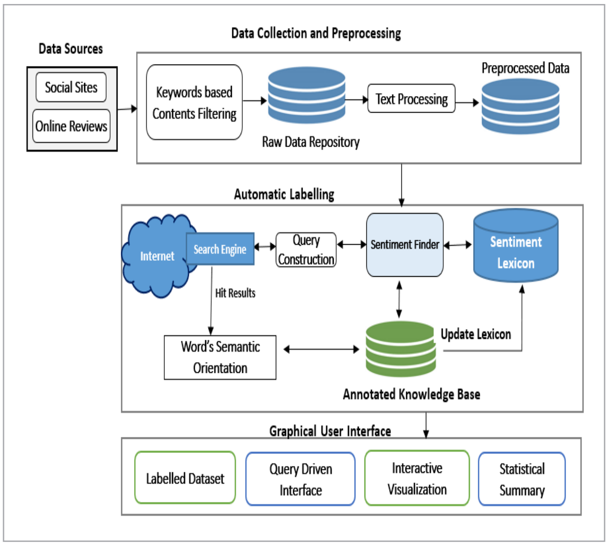
Most of the previous research explained above are fully or partially dependent on prior knowledge and/or manual augmentation. The proposed system is a hybrid approach which integrates lexicon-based and Web-based PMI statistics to iteratively augment the initial lexical database with more domain-specific word and phrases, hence eliminated the need for any domain-specific prior knowledge or manual augmentation. Additionally, the lexicon is dynamic and update incrementally with the domain-specific knowledge instead of just using a static lexicon used in most of the previously reported work. Moreover, high recall is achieved by iteratively updating the lexicon with domain-specific knowledge, unlike other lexicon-based approaches, which either incorrectly labeled, classified as neutral, or ignored those reviews whose words are not found in the predefined lexicon.

3. The Proposed Automatic Labeling System

We have implemented the proposed Domain-Independent Automatic Labeling System (DIALS), as a part of a larger project, which aims to develop an end-to-end social media analytical system to facilitate and gain intelligence for business-related applications. This paper presents part of the research project so far undertaken. We develop a tool called DIALS, which will help us to automatically label the training corpus for developing a supervised machine learning model in the next stage of the development. Overall, DIALS takes written reviews, feedback and/or posts as an input to produce a labeled dataset (positive or negative). The system first extracts sentiment words using the known initial lexical database seeded with 6800 words with known sentiment orientation [15]. First, we use this dictionary to compute the SO of each target word; a detail explanation is given below. If any target word from the given text is not found in the lexical database, we use an online search to compute the co-occurrence of target words with positive and negative contextual words to estimate the semantic orientation. Once, SO of the target word is calculated, it is added in the initial lexicon with its SO to dynamically expand the list of words and phrases in the lexicon. This iterative process continues until no new words' SO can be calculated from the target dataset. Figure 1 presents the high-level architecture of the proposed system; it comprises a graphical user interface for data collection, preprocessing for data preparation, and automatic labeling subcomponents, as explained below.

Figure 1

Functional Architecture of Domain-Independent Automatic Labeling System



3.1. Data Acquisition and Pre-processing

The data collector module uses keyword filtering to monitor and crawl raw data from various social sites continuously; in this case, Twitter and online review sites via Application Programming Interfaces (APIs). These APIs are used to harvest attributes specific to our requirements including, post's ID, text, geolocation, and date/time. The reason for selecting Spatio-temporal tagged text is to allow users to visualize and query valuable information across different dimensions in the next stage. Before analyzing these reviews, the following pre-processing steps have been applied in given order to achieve the best possible results.

Tokenization is performed in the first step, where tokens of the character streams are generated from a given text by splitting the text at every blank space. Characters without any semantic information are removed, including commas, semicolons, stop words, URLs, numbers, and other special characters. Tokens are then converted into the lower case to decrease the influence of typos and to bring the words expressed in different cases to the same words. The corpus of a retrieved post is then stored in an excel file for further processing.

3.2. Automatic Labeling

In this phase, we proposed a bootstrapping approach to iteratively calculate the SO of the retrieved posts from social media. In the first step, we have used a predefined lexicon as an initial lexicon with 2006 positive and 4794 negative words with known SO [15]. Previously, researchers have used the lexicon-based sentiment analysis by using predefined lexicons such as the LIWC dictionary [31], SentiWordNet [4, 17], the Q-WordNet [1]. In this study, we have used Hu and Liu [15] lexicon to estimate SO of the given text, because this dictionary has been used in similar applications [24, 25] and proven useful.

In the first step, the sentiment finder module iteratively scans through the text to determine its SO by extracting target words from the text and searching the target words in the initial lexicon. If all target words are found in the initial lexicon, the system calculates the SO of given text using the following equation:

$$SO(Text_i)=\sum_{1}^{n}(PosWords)-\sum_{1}^{n}(NegWords). \qquad (1)$$

In the next step, a positive label is assigned to the given text if the derived result is greater than zero; otherwise, a negative label is assigned.

$$SO(Text_i)=\begin{cases} +ve, & so>0 \\ -ve, & so<0 \end{cases}. \qquad (2)$$

The process repeats iteratively until the given dataset is fully labeled. This step is purely dependent on the availability of the target words in the initial lexicon.

One of the most significant disadvantages of the lexicon-based methods is that the predefined lexicons are often unreliable, as they are either created manually or built automatically. Hence, it is practically impossible to have all words across all domains in a single dictionary. It is quite likely that some words from the given text are not available in the lexicon; hence, they whether labeled incorrectly or not labeled at all. To address such problems, we have integrated web-search with the lexicon-based method.

To do so, we have extended the lexicon-based method by integrating the web-based PMI statistics inspired by Turney's [41] work to handle missing words in the lexicon. If the sentiment finder module fails to find the matching term in the lexicon, it automatically constructs a query ("target word" AND "reference word") and issues that query to a search engine to find the association of target words with positive and negative reference words. The positive and negative reference words are selected randomly from the initial lexicon. The SO of the target word is then calculated by computing the difference of the word with positive and negative reference words. More specifically, the target word is assigned a numeric value by taking the mutual information between the target word and related reference words. The SO of the target word is positive when it is strongly associated with positive reference words and negative when it is strongly associated with negative reference words. Besides, to assess the sentiment orientation, the numerical value also indicates the strength of the SO. The proposed approach is different from the Turney's work [41] in the sense that we extract words from the given text to find their positive or negative association with the randomly selected positive and negative words. Whereas, Turney [41], extracted the phrases that contain adjectives or adverbs to find their association with the fix positive and negative words "excellent" and "poor". His purpose is to classify a review at run time by taking a review as an input and producing a classification label as recommended or not recommended.

However, Turney's [41] work does not report the SO of each word or phrase; instead, it only gives the classification result of each review. Since the purpose of the proposed approach is not to classify the given text as positive or negative on the go but to find the SO of target words and phrases to be added in the initial lexicon to expand the underlying lexicon with more domain-specific words and phrases. Therefore, for our case identifying SO of words is paramount to incrementally update the underlying lexicon with the SO of missing words calculated from Web-based PMI statistics. Furthermore, in Turney's work, each review requires the Web search and substantial processing of returned results; therefore, it is not computationally efficient for classifying the reviews from social sites. Our approach of calculating only the SO of words and phrases using Web search at first admittedly reduce the processing time of returned results and second, it does not rely on Web search for each review since it first computes the SO of words available in the lexicon, Web search only be performed for missing words. Furthermore, when deployed in a specific business domain, over time, the lexicon grows with more domain-specific words and dependency on the Web search becomes less and less. Hence, the proposed approach is rather simple and efficient for cross-domain sentiment analysis.

By using the proposed approach, the SO of the target words is calculated by using Pointwise Mutual Information (PMI) between two words as follows [11]:

$$PMI(word_1, word_2) = \log_2 \left[\frac{p(word_1, word_2)}{p(word_1)\, p(word_2)} \right], \qquad (3)$$

where p(word$_1$, word$_2$) is the probability that word$_1$ and word$_2$ occur together. The log-ratio between p(word$_1$, word$_2$) and p(word$_1$) p(word$_2$) measures the degree of statistical independence between the two words and the amount of information gained with the presence of one word with the other.

Based on the Equation (3), the SO of the target words is calculated as follows:

$$\text{Hits}_{\text{pos}} = \left[\frac{\text{Hits}\left(\text{Targetword}_i, \text{PositiveReferenceWords}\right)}{\text{Hits}\left(\text{Targetword}_i\right)} \right], \qquad (4)$$

$$\text{Hits}_{\text{neg}} = \left[\frac{\text{Hits}\left(\text{Targetword}_i, \text{NegativeReferenceWords}\right)}{\text{Hits}\left(\text{Targetword}_i\right)} \right], \qquad (5)$$

where, Hits$_{\text{pos}}$ is the PMI score of the number of hits for a query that combines the instance word (i.e., target word) with positive reference words such as *"excellent" and "good",* divide by the hits for the instance word alone. Similarly, Hits_{neg} is the PMI score of the number of hits for a query that combines the instance word with negative reference words such as *"poor"*, "*bad*" and "*terrible*", divide by the hits for the instance word alone.

Sentiment finder calculates the PMI by issuing a query to the Yahoo search engine. Yahoo search engine is chosen because of its ability to handle logical operators AND/OR and proximity queries such as the NEAR operator. The number of documents retrieved as a result of the query considered as "*number of hits*" and used to calculate the semantic orientation of the target word as follow:

$$\text{SO(Targetword}_\text{i}\text{)=log}_2\left[\frac{\text{Hits}_{\text{pos}}}{\text{Hits}_{\text{neg}}}\right]. \quad (6)$$

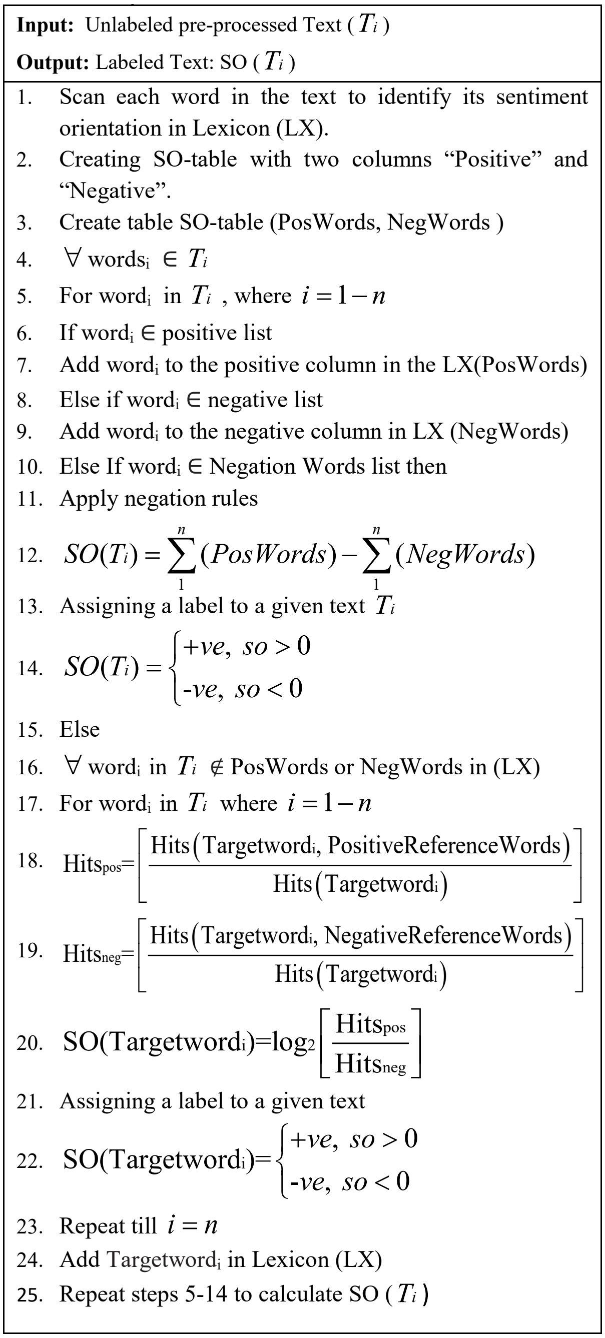
The log-ratio between Hits_{pos} and Hits_{neg} measures the amount of information gained with the presence of target words with positive and negative reference words. A positive label is assigned to a given target word if the value of log-ratio between Hits_{pos} and Hits_{neg} is greater than zero; otherwise, a negative label is assigned.

$$\text{SO(Targetword}_\text{i}\text{)=}\begin{cases} +ve, & so > 0 \\ \text{-}ve, & so < 0 \end{cases}. \quad (7)$$

Sematic information calculated using Web-based PMI statistics is then added to the initial lexicon with the target word and its SO. Hence, one of the unique features of the proposed system is its ability to expand the list of words and phrases in the initial lexicon over time. As the lexicon grows, the dependency on Web search becomes less, which makes the system faster, efficient, and scalable. The final step is to estimate the SO of a given text, which is calculated by using the Equation (1). The step by step procedure for calculating the SO of a given text is presented in Figure 2. The proposed approach is further extended by handling the negation. Whenever a negation occurs with the word, it changes the meaning of the sentence to the opposite of that without the negation. Once the SO of the word is computed, it is checked for negation words such as *"not, never, no, un, dis, im, in, ir"* which serve to reverse the meaning. We have prepared a list of such words. Our algorithm handles the negation by detecting these words and reversing SO of the given

Figure 2

Procedure to predict the semantic orientation



```
Input:  Unlabeled pre-processed Text ( $T_i$ )
Output: Labeled Text: SO ( $T_i$ )
1.  Scan each word in the text to identify its sentiment orientation in Lexicon (LX).
2.  Creating SO-table with two columns "Positive" and "Negative".
3.  Create table SO-table (PosWords, NegWords )
4.  $\forall$ words$_i$ $\in T_i$
5.  For word$_i$ in $T_i$ , where $i = 1 - n$
6.  If word$_i$ $\in$ positive list
7.  Add word$_i$ to the positive column in the LX(PosWords)
8.  Else if word$_i$ $\in$ negative list
9.  Add word$_i$ to the negative column in LX (NegWords)
10. Else If word$_i$ $\in$ Negation Words list then
11. Apply negation rules
12. $SO(T_i) = \sum_1^n (PosWords) - \sum_1^n (NegWords)$
13. Assigning a label to a given text $T_i$
14. $SO(T_i) = \begin{cases} +ve, & so > 0 \\ \text{-}ve, & so < 0 \end{cases}$
15. Else
16. $\forall$ word$_i$ in $T_i$ $\notin$ PosWords or NegWords in (LX)
17. For word$_i$ in $T_i$ where $i = 1 - n$
18. $\text{Hits}_{\text{pos}}\text{=}\left[\frac{\text{Hits}(\text{Targetword}_\text{i}\text{, PositiveReferenceWords})}{\text{Hits}(\text{Targetword}_\text{i})}\right]$
19. $\text{Hits}_{\text{neg}}\text{=}\left[\frac{\text{Hits}(\text{Targetword}_\text{i}\text{, NegativeReferenceWords})}{\text{Hits}(\text{Targetword}_\text{i})}\right]$
20. $\text{SO(Targetword}_\text{i}\text{)=log}_2\left[\frac{\text{Hits}_{\text{pos}}}{\text{Hits}_{\text{neg}}}\right]$
21. Assigning a label to a given text
22. $\text{SO(Targetword}_\text{i}\text{)=}\begin{cases} +ve, & so > 0 \\ \text{-}ve, & so < 0 \end{cases}$
23. Repeat till $i = n$
24. Add Targetword$_i$ in Lexicon (LX)
25. Repeat steps 5-14 to calculate SO ( $T_i$ )
```

sentence from positive to negative, and vice versa, by applying the following rule:

Negation Negative → Positive, e.g., "not bad"

Negation Positive → Negative, e.g., "not good"

Negation Neutral → Negative, e.g., "never worked".

A complete example of calculating the SO of a given text using Equations 4-6 is shown in Table 1. It can be noticed that the SO of the word "*accurate*", "*ugly*" and "*intolerant*" is computed accurately by the system as expected.

Table 1

Examples of estimating words' semantic orientation using Web Search

	Accurate	Ugly	Intolerant
$Hits_{pos}$	37,700,000	20,000,000	8,490,000
$Hits_{neg}$	25,700,000	24,600,000	14,500,000
$\frac{Hits_{pos}}{Hits_{neg}}$	1.4669260	0.81300813	0.5855172
$\log_2\left[\frac{Hits_{pos}}{Hits_{neg}}\right]$	0.5527961 (positive)	-0.29865832 (negative)	-0.7722164 (negative)

4. The Proposed System Prototype

Based on the algorithm above, we have developed a Domain-Independent Automatic Labeling System (DIALS), a working prototype of the labeling system depicted in Figure 1. In general, the system provides an interface that helps users to get the labeled data without requiring extensive linguistic knowledge or large manually labeled training dataset. The system is made available for fellow researchers and can be found on the following link: https://pybanner.kfu.edu.sa/pybanner2014/opinion/.

Figures 3-4 present a step by step procedure of using the system to analyze SO of a given text. The input to the system is unlabeled text crawled from the online review sites on selected keywords. The system can process and label input text in two different ways. Users can provide free text and use "*Analyze text*" functionality to compute the SO of a given text. The system also provides the functionality to analyze multiple reviews in an excel file for automatically labeling large datasets. In this case, the system analyzes reviews in bulk by considering each row of an excel as a separate review and displays SO against each row. The user can also download the labeled dataset, which can be used for training any machine learning model in the subsequent steps of their target applications.

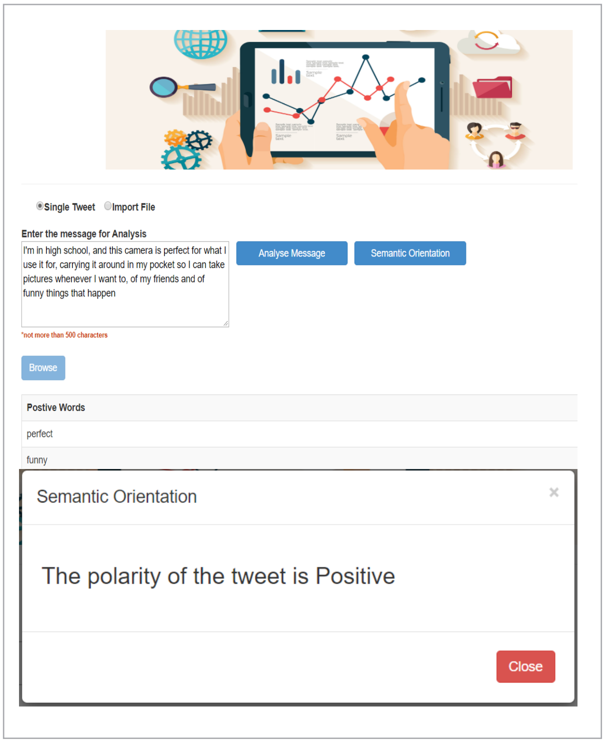
The example below illustrates the end-to-end functionality of the system. We have provided the following text to the system collected from one of the Amazon products "camera"

"I'm in high school, and this camera is perfect for what I use it for, carrying it around in my pocket so I can take pictures whenever I want to, of my friends and of funny things that happen".

As shown in Figure 3, the system has extracted positive words "*perfect" and "funny*", since the SO of these words found in the predefined lexicon, so only based

Figure 3

Polarity detection of free text using DIALS



Enter the message for Analysis

I'm in high school, and this camera is perfect for what I use it for, carrying it around in my pocket so I can take pictures whenever I want to, of my friends and of funny things that happen

*not more than 500 characters

Postive Words
perfect
funny

Semantic Orientation

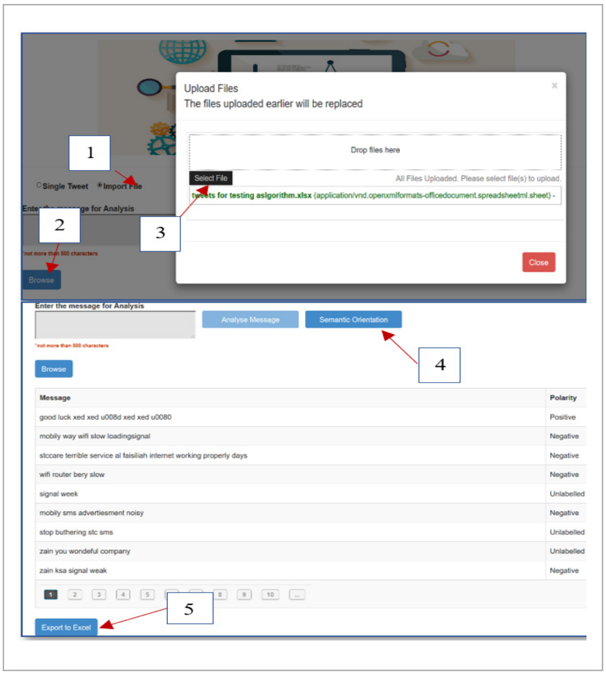
The polarity of the tweet is Positive

on the SO of these two words, the overall polarity of a given text is calculated as "positive". Hence, for the reviews whose words' SO is found in the predefined lexicon, the system is straightforward and fast.

Figure 4 shows that the user can upload the dataset from any business domain in the excel file, and the system can calculate the SO of each text in bulk. Download functionality is provided to facilitate users to export labeled datasets that can be used to train any classifier without any user involvement to further annotate the data.

Figure 4

Polarity detection of text in bulk using DIALS



Upload Files

The files uploaded earlier will be replaced

tweets for testing aslgorithm.xlsx (application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) -

All Files Uploaded. Please select file(s) to upload.

Enter the message for Analysis

*not more than 500 characters

Message	Polarity
good luck xed xed u008d xed xed u0080	Positive
mobily way wifi slow loadingsignal	Negative
stccare terrible service al faisiliah internet working properly days	Negative
wifi router bery slow	Negative
signal week	Unlabelled
mobily sms advertiesment noisy	Negative
stop buthering stc sms	Unlabelled
zain you wondeful company	Unlabelled
zain ksa signal weak	Negative

To test this functionality, we used the DIALS to run several experiments on the dataset generated through different social sites. Since we developed the system to facilitate the business-related intelligent applications, we have tested it intensively on Saudi Telecom related data collected from Twitter. The following examples show how the system computes the polarity of text using the predefined lexicon and Web search.

Tweet$_1$ = *"STC customer care, you have the worst customer service ever".*

Scan each word in the tweet and search the predefined lexicon to find the SO of each target word in the given text and then calculate the SO of the given text as follows:

$$SO(Tweet_i) = \sum_{1}^{n}(PosWords) - \sum_{1}^{n}(NegWords)$$

$$SO(Tweet_i) = \begin{cases} +ve, & so > 0 \\ -ve, & so < 0 \end{cases}.$$

Output: SO (Tweet$_1$) = negative

Tweet$_2$ =*"what an intolerant customer representative"*

For this tweet word *"intolerant"* was not found in the lexicon; therefore, the system has used the Web search to first calculate SO of *"intolerant"* as positive or negative as follow:

word$_i$ → (intolerant)

Issue a query to the Yahoo search engine to find the number of retrieved documents Hits$_{pos}$.

Hits$_{pos}$ = hits ("intolerant" and "excellent")

Issue a query to Yahoo search engine to find the number of retrieved documents Hits$_{neg}$

Hits$_{neg}$ = hits ("intolerant" and "poor")

Calculate the SO "intolerant" with positive and negative reference words using Equations (4)-(6):

SO ("intolerant") = 8,490,000 / 14,500,000 = 0.5855172

$\log_2$ (SO) = $\log_2$ (0.5855172) = -0.7722164

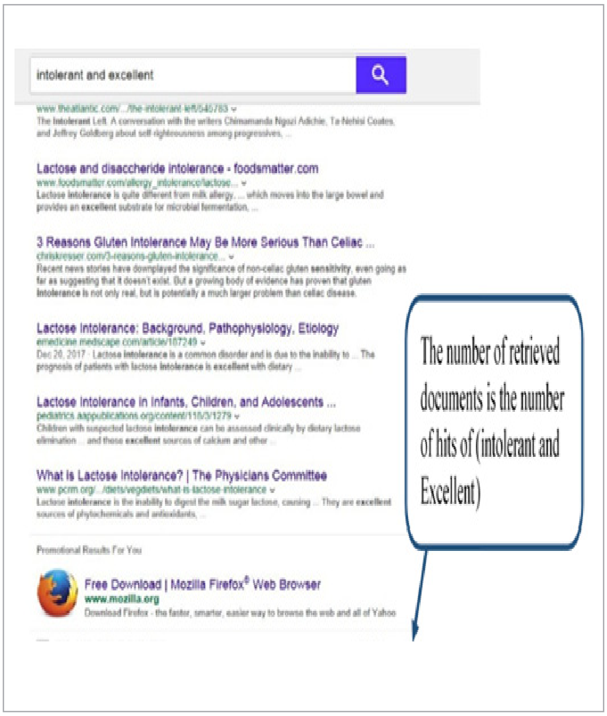
using Equation (7), the SO of "intolerant" is set as negative.

The system automatically created the entry of the word intolerant along with its SO in the initial lexicon. Over time, lexicon will become enriched with more domain-specific words, and dependency on the Web search will become less. Hence, we will be able to avoid Web search and substantial processing time of results.

Figures 5-6 show the retrieved results for the query constructed by the system for the word *"intolerant"* with reference words *"excellent"* and *"poor"* using the Yahoo search engine.

Figure 5

Query results with a search term and positively associated words



intolerant and excellent

www.theatlantic.com/.../the-intolerant-left/545783

The Intolerant Left. A conversation with the writers Chimamanda Ngozi Adichie, Ta-Nehisi Coates, and Jeffrey Goldberg about self-righteousness among progressives, ...

Lactose and disaccheride intolerance - foodsmatter.com

www.foodsmatter.com/allergy_intolerance/lactose...

Lactose intolerance is quite different from milk allergy. ... which moves into the large bowel and provides an excellent substrate for microbial fermentation, ...

3 Reasons Gluten Intolerance May Be More Serious Than Celiac ...

chriskresser.com/3-reasons-gluten-intolerance...

Recent news stories have downplayed the significance of non-celiac gluten sensitivity, even going as far as suggesting that it doesn't exist. But a growing body of evidence has proven that gluten intolerance is not only real, but is potentially a much larger problem than celiac disease.

Lactose Intolerance: Background, Pathophysiology, Etiology

emedicine.medscape.com/article/187249

Dec 20, 2017 · Lactose intolerance is a common disorder and is due to the inability to ... The prognosis of patients with lactose intolerance is excellent with dietary ...

Lactose Intolerance in Infants, Children, and Adolescents ...

pediatrics.aappublications.org/content/118/3/1279

Children with suspected lactose intolerance can be assessed clinically by dietary lactose elimination ... and these excellent sources of calcium and other ...

What Is Lactose Intolerance? | The Physicians Committee

www.pcrm.org/.../diets/vegdiets/what-is-lactose-intolerance

Lactose intolerance is the inability to digest the milk sugar lactose, causing ... They are excellent sources of phytochemicals and antioxidants, ...

Promotional Results For You

Free Download | Mozilla Firefox® Web Browser

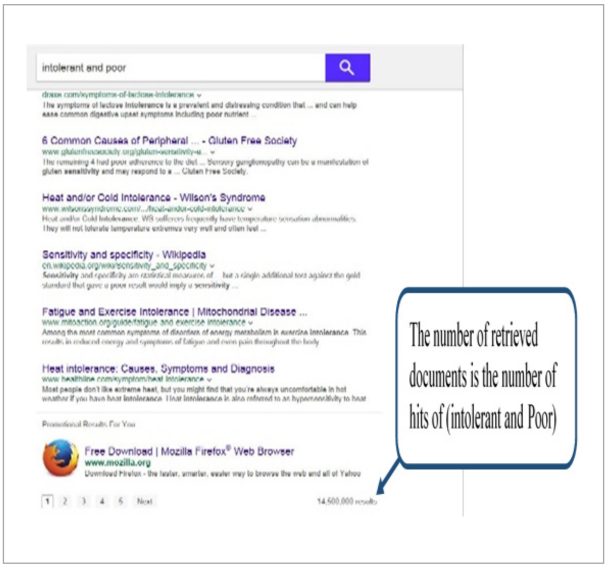
www.mozilla.org

Download Firefox - the faster, smarter, easier way to browse the web and all of Yahoo

The number of retrieved documents is the number of hits of (intolerant and Excellent)

Figure 6

Query results with a search term and negatively associated words



intolerant and poor

draxe.com/symptoms-of-lactose-intolerance

The symptoms of lactose intolerance is a prevalent and distressing condition that ... and can help ease common digestive upset symptoms including poor nutrient ...

6 Common Causes of Peripheral ... - Gluten Free Society

www.glutenfreesociety.org/gluten-sensitivity-a...

The remaining 4 had poor adherence to the diet ... Sensory ganglionopathy can be a manifestation of gluten sensitivity and may respond to a ... Gluten Free Society.

Heat and/or Cold Intolerance - Wilson's Syndrome

www.wilsonssyndrome.com/.../heat-andor-cold-intolerance

Heat and/or Cold Intolerance. WS sufferers frequently have temperature sensation abnormalities. They will not tolerate temperature extremes very well and often feel ...

Sensitivity and specificity - Wikipedia

en.wikipedia.org/wiki/Sensitivity_and_specificity

Sensitivity and specificity are statistical measures of ... but a single additional test against the gold standard that gave a poor result would imply a sensitivity ...

Fatigue and Exercise Intolerance | Mitochondrial Disease ...

www.mitoaction.org/guide/fatigue-and-exercise-intolerance

Among the most common symptoms of disorders of energy metabolism is exercise intolerance. This results in reduced energy and symptoms of fatigue and even pain throughout the body

Heat intolerance: Causes, Symptoms and Diagnosis

www.healthline.com/symptom/heat-intolerance

Most people don't like extreme heat, but you might find that you're always uncomfortable in hot weather if you have heat intolerance. Heat intolerance is also referred to as hypersensitivity to heat

Promotional Results For You

Free Download | Mozilla Firefox® Web Browser

www.mozilla.org

Download Firefox - the faster, smarter, easier way to browse the web and all of Yahoo

1 2 3 4 5 Next 14,500,000 results

The number of retrieved documents is the number of hits of (intolerant and Poor)

6. Evaluation of the DIALS Performance

To evaluate the effectiveness of the DIALS, we have manually labeled 300 positive and 200 negatives tweets collected for Saudi Telecom Companies (STC). We tested the system on the unlabeled tweets, and results were compared with the manually labeled dataset. It was observed that the Tweets having all words available in the lexicon are always classified correctly, hence giving 100% accuracy. However, 17 out of 500 Tweets are incorrectly classified using the Web search, giving an accuracy of 94.33%. An example set used for testing the DIALS is shown in Table 2. The first column in the table shows few Tweets where Web corpus has been used to calculate the labels, the second column shows results obtained through the system, and third and fourth columns show whether the system has labeled the Tweets correctly or incorrectly. It can be observed from the results that the majority of Tweets are labeled by the system correctly; only 5% of tweets are labeled incorrectly. One reason for incorrectly labeled might be the fact that we computed the unknown word polarity in correlation with few extreme positive and negative words

Table 2

Experimental results

Tweets	DIALS Results	Correct	Incorrect
nice service at affordable prices	positive	√	
his comments were very appropriate at the time.	positive	√	
the man is currently employed	positive	√	
the company released their critically acclaimed 2007	positive	√	
a bright and bubbly personality	negative		√
the big-hearted bunch have decided to donate money	negative		√

and there might be the possibility that unknown word does not appear with these words. In the future, we will compute the correlation with other less positive and less negative words to improve the accuracy of the algorithm.

7. Empirical Evaluation

We evaluated the prediction performance of DIALS and compared it with the baseline-lexicon-based [38] and machine learning [30] algorithms using the real-world datasets of three different domains. The reason to evaluate the proposed system on three different datasets is to verify its portability across multiple domains and completely unseen data.

The first dataset we have used to test the baseline algorithms is collected from Amazon product reviews. A collection of 1000 reviews on four products, including digital camera, smartphones, printer, and Bluetooth device, has been collected. This domain is relatively convenient for experiment purposes because of the availability of a large collection of such online reviews and overall summarization of the user's opinion in the form of a 1 to 5-star rating system, where 5-star rating indicates highly positive review and 1-star indicates highly negative review. For evaluation purposes, we have automatically extracted ratings to prepare the labeled dataset for these 1000 reviews. Hence, we did not need to tag this dataset manually. Results generated by the DIALS are compared with the labeled dataset to measure the overall accuracy of each algorithm selected for comparison. For this dataset, we extracted the textual information from the original HTML document format. For text processing, we removed punctuation, numbers, special characters, stop words, and lowercase the capital letters. We built a vocabulary of 2500 words by selecting the most frequent words.

In the second dataset, we collected 1000 movie reviews from the Internet Movie Database (IMDb). After text processing, a vocabulary of 3000 words is generated. To train a machine learning algorithm and performance comparison, we used the tagged dataset for a movie review from the polarity dataset provided by Pang et al. [30] available at http://www.cs.cornell.edu/people/pabo/movie-review-data/. We randomly selected 500 positive and 500 negative reviews.

For the third dataset, we collected 1000 Tweets for a business-centric application (Saudi Telecom Company); after preprocessing, we generated a vocabulary of 1500 words. To evaluate the performance of the baseline algorithm and DIALS on this dataset, we manually labeled 300 positives and 200 negatives. We compared the results generated by the system with the manually labeled dataset.

For supervised learning algorithms, it is essential to perform training and testing on separate datasets where the training dataset is transformed into features that appeared in the text to train the classifier, and the testing dataset is used to evaluate the prediction accuracy of the classifier. Therefore, we divided our dataset into five equal-sized bins, each containing balanced class distributions to train and test the classifier using 5-fold cross-validation with both Support Vector Machine (SVM) and Maximum Entropy (ME) algorithms. However, for the lexicon-based method, such cross-validation was not applicable since we used the predefined lexicon to derive the semantic orientation of the text instead of learning from the text features as in the case of supervised learning methods.

Results from all datasets were obtained by formulating a binary classification task to predict the polarity of each review using traditional lexicon-based and machine learning methods and compared their prediction performance with the DIALS. Since, labeling data manually and training baseline algorithms take much time, we choose a smaller dataset size of 1000 reviews for each dataset. We ran both baseline algorithms on these three datasets one by one and compared results with DIALS to assess how well each algorithm accurately predicts the correct label.

For evaluation, we compared experimental results with the following most relevant machine learning methods.

Support Vector Machines (SVMs): The majority of text classification research built SVMs, trained on a particular dataset using features such as unigram, bigram and/ or part-of-speech labels, and have proven very effective in natural language processing applications [29, 35]. The idea behind the classification procedure is to find a hyperplane to separate the document vectors in one class from others. SVMs has been largely used baseline method to build sentiment classifier. For example, Pang et al. [30] used SVM to classify a movie review dataset as recommended or not recommended; Zhang et al. [44] and Go et al. [14] used SVM for Twitter sentiment analysis problem.

Maximum Entropy (ME): ME is another state-of-the-art classification technique which has been successfully applied to many natural language applications such as text classification, Named Entity Recognition (NER), and Part-of-Speech (POS) tagging [10, 14, 28, 30, 32]. ME estimates the conditional distribution of the class label on given documents. In a two-class problem, the idea is finding the correct class label over feature distributions such as unigram or bigram. We have selected Pang et al. [30] sentiment classification system for movie review due to the availability of datasets.

For lexicon-based methods, we have found the following two most relevant opinion mining systems:

Feature-Based System (FBS): FBS is a lexicon-based method proposed by Hu and Liu [15] for feature-based sentiment analysis. It uses association rule mining to extract frequent features using a noun phrase from a given review. Frequent features are then used to extract potential opinion words (adjectives only) using WordNet [23] synonym/antonym in conjunction with the set of seed words to find actual opinion words. The SO of each word is aggregated into a single score to predict the SO of each opinionated sentence.

Semantic Orientation CALculator (SO-CAL): SO-CAL is a lexicon-based method proposed by Taboada et al. [38]. It uses a list of seeded words and phrases with their associated orientation and then uses these seeded words to compute the sentiment score of each document.

We evaluated the performance of the machine learning and the lexicon-based method using the standard evaluation measures of accuracy, precision, recall, and F-Score. To cross-verify the results, we compared the results generated by these algorithms with the labeled dataset for each domain. Accuracy of baselines machine learning and lexicon-based methods for the selected datasets are shown in Table 3 and Figure 7.

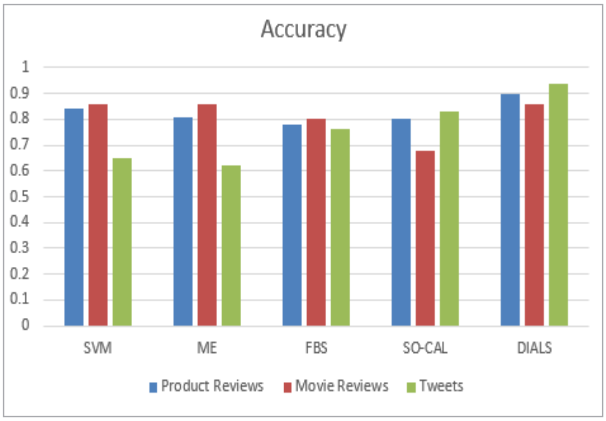
We selected bigram features to train the machine learning models to look for negation words and more contextual information in general. Accuracy results show that SVM and ME outperform lexicon-based models when the model is trained and tested on a single domain, e.g., when the model is trained and test on product review. However, when we trained the machine learning model on movie reviews and tested on product reviews and the Tweets dataset, we observed accuracy dropped by 45%, and recall decreased by 17%. Hence, we were able to conclude that machine learning approaches suffer from cross-domain portability and do not apply in our case of developing a system for cross-domain business applications.

On the other hand, the effectiveness of any lexicon-method depends on the richness of the underlying lexicon; therefore, much research has been done to enrich the underlying lexicon by considering enhancements in linguistic features. Lexicon-based methods results reported in Table 3 show consistent accuracy across all datasets, which suggests that the lexicon-based system could outperform machine learning methods for the cross-domain datasets since lexicon-based methods do not rely on the quality of

Table 3
Accuracy of baseline methods

Methods	Accuracy Evaluation		
	Product Reviews	Movie Reviews	Tweets
SVM	0.84	0.86	0.65
ME	0.81	0.86	0.62
FBS	0.78	0.80	0.76
SO-CAL	0.80	0.68	0.83
DIALS	0.90	0.86	0.94

Figure 7
Graphical representation of the accuracy of baseline methods



training data. On the Twitter dataset, both SVM and ME have shown a low accuracy (0.65, 0.62), one possibility might be the lack of sufficient training examples.

Among lexicon-based methods, FBS outperformed SO-CAL on movie reviews by 12% higher accuracy and underperformed on the Twitter dataset by 17% lower accuracy. DIALS outperformed both FBS and SO-CAL in the accuracy measure of 10–12% on product reviews, 10-14% on movie reviews, and 11-18% on Twitter datasets. The higher accuracy of the DIALS is due to the integration of Web-based PMI statistics for calculating SO of the missing words in the underlying lexicon. Whereas, both FBS and SO-CAL, either incorrectly label or label as neutral for the text with missing words in the underlying dictionary.

Tables 4-6 show precision, recall, and F-score of baseline methods on three datasets selected for empirical evaluation, also results are reported graphically in Figures 8-10 for visual interpretation.

DIALS has shown a 12-14% increase in precision over baseline machine learning and lexicon-based methods on product and movie review datasets. The precision of machine learning is slightly higher on the product and movie review dataset as compared to the Twitter dataset, perhaps due to the availability of sufficient training data on these two domains. However, for Twitter dataset DIALS has shown 24-26% higher precision over machine learning methods, due to unique features of Twitter data and lack of training data to train machine learning algorithms on Twitter dataset. The precision of both lexicon-based methods across all datasets is very close to DIALS (only 10-12% higher), but the recall of DIALS is significantly higher than FBS and SO-CAL. DIALS has outperformed FBS by 18% and SO-CAL by 28% higher in recall on the product review dataset. On movie review, dataset DIALS has shown 38% higher recall as compared to FBS and 29% higher than SO-CAL. On Twitter, dataset DIALS has shown 32% higher recall over FBS and 23% higher over SO-CAL. The significantly higher recall for DIALS over FBS and SO-CAL is due to the following two main differences:

DIALS extract adjectives, adverbs, and negation words to handle contextual information, whereas FBS only uses adjectives from a text as features, and SO-CAL uses a list of seeded words to calculate the SO of unknown words.

Additionally, DIALS uses Web-based PMI statistics by considering the entire Web as a corpus to compute SO of unknown words; therefore, able to identify the SO of a large percentage of text. On the other hand, FBS and SO-CAL either ignore the text with missing or unknown words in the underlying lexicon or label them neutral.

Table 4
Precision, recall, and F-score on product reviews

Amazon Product Reviews			
Methods	Precision	Recall	F-Score
SVM	0.838	0.805	0.82
ME	0.827	0.773	0.735
FBS	0.863	0.656	0.673
SO-CAL	0.896	0.738	0.720
DIALS	0.924	0.917	0.926

Table 5
Precision, recall, and F-score on movie reviews

Movie Reviews			
Methods	Precision	Recall	F-Score
SVM	0.827	0.784	0.805
ME	0.802	0.679	0.735
FBS	0.843	0.56	0.673
SO-CAL	0.809	0.65	0.7201
DIALS	0.913	0.94	0.926

Table 6
Precision, recall, and F-score on the Twitter dataset

Tweets Dataset			
Methods	Precision	Recall	F-Score
SVM	0.783	0.734	0.756
ME	0.684	0.694	0.689
FBS	0.84	0.58	0.687
SO-CAL	0.864	0.67	0.754
DIALS	0.943	0.90	0.9217

Figure 8

Graphical representation of precision, recall, and F-score on product reviews

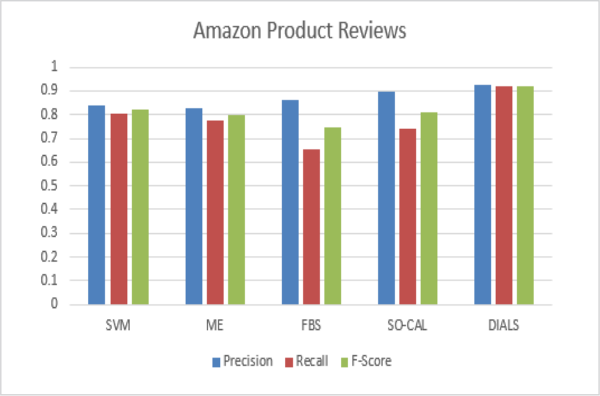


Figure 9

Graphical representation precision, recall, and F-score on movie reviews

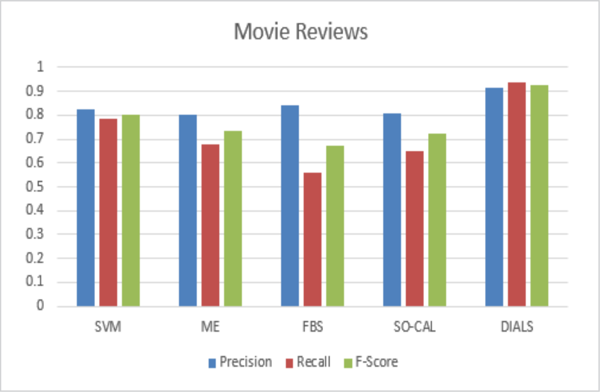
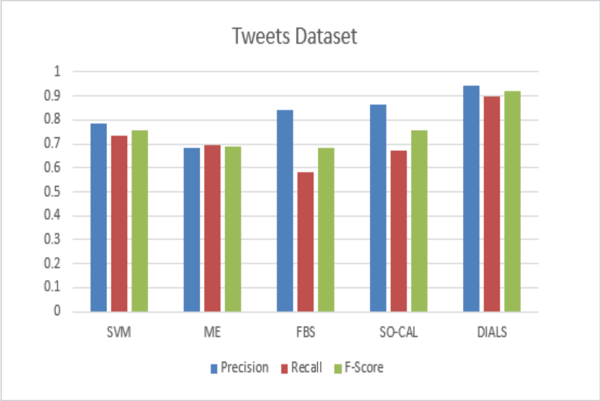


Figure 10

Graphical representation of precision, recall, and F-score on the Twitter dataset



Although DIALS enhances recall significantly by integrating massive text corpus from Web as compared to existing lexicon-based approaches, which mainly use a manually built dictionary; however, PMI computation to collect Web statistics requires a set of search engine queries to get hit counts for every unknown word and requires a substantial amount of computation to process the retrieved results. This computation certainly increases the run time as compared to simple lexicon-based approaches, which are generally superior in terms of performance. However, by integrating the lexicon-based method, the system does not need web-based PMI computation for every iteration due to the following two reasons. First, most of the common words are already available in the predefined lexicon applicable to all domains, for which the system does not need any additional computation. Second, when deployed in a specific business application, initially, the system may need to label domain-dependent unknown words with the help of the search engine. However, once such words are continuously resolved and added to the predefined lexicon automatically, over time, the performance overhead due to web-based PMI computation might decrease.

7. Conclusion and Future Work

In this research, we have developed a purely unsupervised and domain-independent system called DIALS to automate the process of labeling large amounts of business-specific reviews generated by customers on different online sites. The labeled dataset can then be used to summarize customer reviews without any further need for development and/or can be used as a training corpus for text classification in any domain. We have adopted lexicon and Web-based PMI statistics by considering Web as a massive text corpus to compute the semantic orientation of words and phrases, without requiring extensive linguistic knowledge or human intervention. Results have shown robust performance over cross-domain data that is difficult to achieve with machine learning methods or solely relying on the manually crafted lexicon. Compared to other lexicon-based methods, DIALS has shown consistent improvement in recall and F-score when tested on three different datasets, i.e., Amazon product reviews, movie reviews, and Tweets collection.

Moreover, the system can be easily adaptable to any business application to track online discussion and generate sentiment timelines.

The main finding of our work is that the lexicon-based methods are robust, perform better on cross-domain datasets, and can be easily enhanced with multiple sources of knowledge as compared to the machine learning methods. However, we could not get the full benefit of such systems unless underlying lexicon is fully enriched with language constructs, rules and usage patterns, (which is not often the case, since, it takes strenuous manual annotation to develop such lexicon), which causes significant difficulties for existing lexicon-based methods. In the proposed approach, we managed to overcome some linguistic limitations of the lexicon-based method by utilizing external evidence in the form of Web PMI statistics to find the SO of a review in an effective way. However, the system is still in its infancy and needs further research on fine-grain discourse analysis, such as finding SO of sub-topics, extraction of comparative sentences, analysis on additional contextual information, pronoun resolution, and idiomatic phrases. We plan to address these fine-grain discourse analyses as part of future research. Additionally, we plan to extend the system to develop an unsupervised text classifier for Arabic text classification and summarization.

Acknowledgment

This work is supported by the Deanship of Scientific Research (DSR) at King Faisal University under Grant No. 180052. We thank DSR for funding this research project.

References

1. Agerri, R., García-Serrano, A. Q-WordNet: Extracting Polarity from WordNet Senses, in LREC, 2010.
2. Alistair, K., Diana, I. Sentiment Classification of Movie and Product reviews Using Contextual Valence Shifters,. Proceedings of FINEXIN, 2005.
3. Argamon, S., C. Whitelaw, P. Chase, S. R. Hota, N. Garg, Levitan, S. Stylistic Text Classification Using Functional Lexical Features. Journal of the American Society for Information Science and Technology, 2007, 58(6), 802-822. https://doi.org/10.1002/asi.20553
4. Baccianella, S., Esuli, A., Sebastiani, F. Sentiwordnet 3.0: an enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC, 2010, 2200-2204.
5. Bartlett, J., Albright, R. Coming to a Theater Near You! Sentiment Classification Techniques Using SAS Text Miner. SAS Global Forum, 2008.
6. Boiy, E., Hens, P., Deschacht, K., Moens, M.-F. Automatic Sentiment Analysis in On-line Text. ELPUB, 2007, 349-360.
7. Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., Gonçalves, M. A. Combining Link-Based and Content-Based Methods for Web Document Classification. Proceedings of the 12th International Conference on Information and Knowledge Management, 2003, ACM, 394-401. https://doi.org/10.1145/956863.956938
8. Chaovalit, P., Zhou, L. Movie Review Mining: A Comparison Between Supervised and Unsupervised Classification Approaches. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005, 112c-112c.
9. Cheng, X., Yan, X., Lan, Y., Guo, J. BTM: Topic Modeling over Short Texts. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12), 2928-2941. https://doi.org/10.1109/TKDE.2014.2313872
10. Chieu, H. L., Ng, H. T. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. Proceedings of the 19th International Conference on Computational Linguistics, 2002, 1, 1-7. https://doi.org/10.3115/1072228.1072253
11. Church, K.W., Hanks, P. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 1990, 16(1), 22-29.
12. Finn, A., Kushmerick, N., Smyth, B. Genre Classification and Domain Transfer for Information Filtering. European Conference on Information Retrieval, 2002, 353-362. https://doi.org/10.1007/3-540-45886-7_23
13. Gamon, M. Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. Proceedings of the 20th International Conference on Computational Linguistics, 2004, 841. https://doi.org/10.3115/1220355.1220476
14. Go, A., Bhayani, R., Huang, L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford, 2009, 1(12), 2009.

15. Hu, M., Liu, B. Mining and Summarizing Customer Reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, 168-177. https://doi.org/10.1145/1014052.1014073

16. Karlgren, J., Cutting, D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Proceedings of the 15th Conference on Computational Linguistics, 1994, 2, 1071-1075. https://doi.org/10.3115/991250.991324

17. Khatoon, S. Real-Time Twitter Data Analysis of Saudi Telecom Companies for Enhanced Customer Relationship Management. International Journal of Computer Science and Network Security (IJCSNS), 2017, 17(2), 141-147.

18. Koppel, M., Schler, J., Argamon, S. Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, 2009, 60(1), 9-26. https://doi.org/10.1002/asi.20961

19. Liu, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 2012, 5(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

20. Liu, B. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, 2010, 2, 627-666.

21. Medhat, W., Hassan, A., Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. Ain Shams Engineering Journal, 2014, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

22. Mejova, Y., Srinivasan, P. Exploring Feature Definition and Selection for Sentiment Classifiers, in ICWSM, 2011.

23. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. Introduction to WordNet: An On-Line Lexical Database. International Journal of Lexicography, 1990, 3(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

24. Miner, G., Elder J., Hill, T. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, 2012, Academic Press.

25. Mostafa, M. M. More than Words: Social Networks‘ Text Mining for Consumer Brand Sentiments. Expert Systems with Applications, 2013, 40(10), 4241-4251. https://doi.org/10.1016/j.eswa.2013.01.019

26. Mullen, T., Collier, N. Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004.

27. Nie, W., Liu, A., Su, Y. Cross-Domain Semantic Transfer from Large-Scale Social Media. Multimedia Systems, 2016, 22(1), 75-85. https://doi.org/10.1007/s00530-014-0394-9

28. Nigam, K., Lafferty, J., McCallum, A. Using Maximum Entropy for Text Classification. IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, 61-67.

29. Pang, B., Lee, L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2008, 2(1-2), 1-135. https://doi.org/10.1561/1500000011

30. Pang, B., Lee, L., Vaithyanathan, S. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing- 2002, 10, 79-86. https://doi.org/10.3115/1118693.1118704

31. Pennebaker, J. W., Mehl, M. R, Niederhoffer, K. G. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 2003, 54(1), 547-577. https://doi.org/10.1146/annurev.psych.54.101601.145041

32. Phan, X.-H., Nguyen, L.-M., Horiguchi, S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections. Proceedings of the 17th International Conference on World Wide Web, 2008, 91-100. https://doi.org/10.1145/1367497.1367510

33. Qiu, G., Liu, B., Bu, J., Chen, C. Expanding Domain Sentiment Lexicon Through Double Propagation. IJCAI, 2009, 1199-1204.

34. Riloff, E., Wiebe, J. Learning Extraction Patterns for Subjective Expressions. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 2003. https://doi.org/10.3115/1119355.1119369

35. Salvetti, F., Reichenbach, C., Lewis, S. Opinion Polarity Identification of Movie Reviews. Computing Attitude and Affect in Text: Theory and Applications, 2006, 303-316. https://doi.org/10.1007/1-4020-4102-0_23

36. Stamatatos, E. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 2009, 60(3), 538-556. https://doi.org/10.1002/asi.21001

37. Stamatatos, E., Fakotakis, N., Kokkinakis, G. Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics, 2000, 26(4), 471-495. https://doi.org/10.1162/089120100750105920

38. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. Lexicon-Based Methods for Sentiment Analysis, Com-

putational Linguistics, 2011, 37(2), 267-307. https://doi.org/10.1162/COLI_a_00049

39. Tomokiyo, L. M., Jones, R. You‘re Not from ‚Round Here, Are You?: Naive Bayes Detection of Non-native Utterance Text. Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, 2001, 1-8. https://doi.org/10.3115/1073336.1073367
40. Tong, R. M. An Operational System for Detecting and Tracking Opinions in On-line Discussion. Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, 2001.
41. Turney, P. D. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, 417-424. https://doi.org/10.3115/1073083.1073153
42. Wiebe, J., Riloff, E. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. International Conference on Intelligent Text Processing and Computational Linguistics, 2005, 486-497. https://doi.org/10.1007/978-3-540-30586-6_53
43. Xia, R., Zong, C., Li, S. Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification. Information Sciences, 2011, 181(6), 1138-1152. https://doi.org/10.1016/j.ins.2010.11.023
44. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B. Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis. HP Laboratories, Technical Report HPL-2011, 2011, 89.