The Knowledge Engineering Review, Vol. 30:1, 76–105. © Cambridge University Press, 2013
doi:10.1017/S0269888913000167
First published online 3 May 2013

Performance and trends in recent opinion retrieval techniques

SYLVESTER O. ORIMAYE, SAADAT M. ALHASHMI and EU-GENE SIEW

Faculty of Information Technology, Monash University, Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor Darul Ehsan, Malaysia;
e-mail: sylvester.orimaye@monash.edu, alhashmi@monash.edu, siew.eu-gene@monash.edu

Abstract

This paper presents trends and performance of opinion retrieval techniques proposed within the last 8 years. We identify major techniques in opinion retrieval and group them into four popular categories. We describe the state-of-the-art techniques for each category and emphasize on their performance and limitations. We then summarize with a performance comparison table for the techniques on different datasets. Finally, we highlight possible future research directions that can help solve existing challenges in opinion retrieval.

1 Introduction

The purpose of this paper is to review the performance of major influential opinion retrieval techniques on different datasets. The paper shows how opinion retrieval techniques have propagated from the foremost techniques to the recent ones. It enables easy identification of performance and limitations with respect to different techniques.

Since the research on opinion retrieval is relatively new, we identified state-of-the-art opinion retrieval techniques on the Association for Computing Machinery (ACM) portal and Google Scholar in early 2011. Identified techniques were then classified under *text classification approach*, *lexicon-based approach*, *probabilistic approach*, and *other emerging approaches*. The *text classification* and *lexicon-based* approaches have been the two conventional approaches discussed in the literature on opinion retrieval. Recently, *probabilistic* and other *new* approaches emerged and have also become popular in the literature. For all the techniques identified, we did not only consider published works that conducted the original opinion retrieval experiments, but recent papers that show the significance of the techniques.

Recently, some review papers have discussed opinion retrieval to a certain extent. However, these papers have only been successful in describing the theoretical foundations of opinion retrieval and the potential applications of opinion retrieval to wide range of natural language problems. For example, a survey paper on opinion mining and sentiment analysis was written by Pang and Lee (2008). The paper discusses fundamental knowledge on major opinion retrieval and summarization approaches. Another survey paper was written by Liu (2010). The paper introduces a theoretical description for the concepts and approaches in sentiment analysis and subjectivity. In fact, the main idea was to facilitate learning and teaching of opinion retrieval in an academic environment.

In contrast to the review papers written by Pang and Lee (2008) and Liu (2010), our purpose is to discuss the performance, challenges, and limitations of recent opinion retrieval techniques evaluated on different datasets. For this purpose, we provide performance summary tables for leading techniques under major opinion retrieval categories. We believe this will aid decision on

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

suitable and appropriate opinion retrieval technique to be considered for further research and development tasks or actual implementation for commercial applications.

The organization of this paper is as follows: the remaining part of Section 1 discusses the importance of opinion retrieval, its trends in the global research community, and overview of existing opinion retrieval approaches. Sections 2, 3, 4, and 5 discuss *text classification*, *lexicon-based*, *probabilistic*, and *other emerging* opinion retrieval approaches, respectively. In each section, we give an overview of the approach supported by a component diagram. We show an example state-of-the-art work with a description of the algorithm, its performance, and limitations. We then present a performance comparison table for techniques proposed on different datasets. The table also shows the dataset used for the experiments, selected evaluation metrics, and the performance results. Section 6 discusses future research directions, and Section 7 draws summary and conclusions.

1.1 Opinion retrieval and its importance

For the purpose of this review, we define opinion retrieval as a systematic and computational process that identifies subjective information from human-generated textual contents in order to retrieve textual information that contains opinion. Further inference on opinion retrieval describes extraction and judgement analysis on various aspects of opinionated contents (Pang & Lee, 2008; Liu, 2010). However, the arrival of web 2.0 interactive media such as web logs (blogs) and other Social Network sites creates an avenue whereby humans show their diversities in terms of writing styles. When people express opinions on blogs, for example, they write with self-perception often conveyed by different psychological inferences; they present opinions with different languages and styles, and even play upon words. These challenges among many have made opinion retrieval task difficult. It has also led to active research on opinion retrieval, and had since recorded a huge number of research works since early 2000 (Pang *et al*., 2002; Nasukawa & Yi, 2003; Hu & Liu, 2004; Pang & Lee, 2004; Liu *et al*., 2005; Wilson *et al*., 2005; Eguchi & Lavrenko, 2006; Zhang *et al*., 2007b; Abbasi *et al*., 2008; Ding *et al*., 2008; Gerani *et al*., 2009). Some of the challenges highlighted above can be attributed to the different ways of expressing *opinions or sentiments* by humans. This is common to complex interactive opinion-driven sources such as collection of blogs called *blogosphere* (e.g. Google Blogs Search[1] and Technorati Blog Directory[2]) and some popular review Websites such as reviewcenter[3].

Interestingly, there is huge need and growing demand for mining different kinds of knowledge from opinion-driven sources. For example, organizations wish to know relevant and qualitative opinions toward products or services rendered (Liu, 2010), individuals wish to retrieve relevant information for personal use (Lee *et al*., 2010), and the need for understanding human behaviors through trends at real-time is also becoming more apparent and necessary (Munson & Resnick, 2010; Zafarani *et al*., 2010). The New York Times[4] published an article in late 2009 relating the emergence of opinion mining or sentiment analysis to translating human emotion from social media (e.g. blogs) into hard data. Such data can be used to improve the quality of a product; enhance efficiency and productivity of a company; and understand political demands of the masses. Other important usage of opinion retrieval include, analysis of questionnaire or survey responses, personalized search engines, opinion or sentiment analytic, and understanding consumers sentiments from products or movie reviews.

1.2 Global trend in research community on opinion retrieval

Text REtrieval Conference (TREC) has been a leading research community on opinion retrieval techniques (Ounis *et al*., 2008; Macdonald *et al*., 2009). TREC introduced the Blog Track in 2006

[1] http://blogsearch.google.com/

[2] http://technorati.com/

[3] http://www.reviewcentre.com/

[4] http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

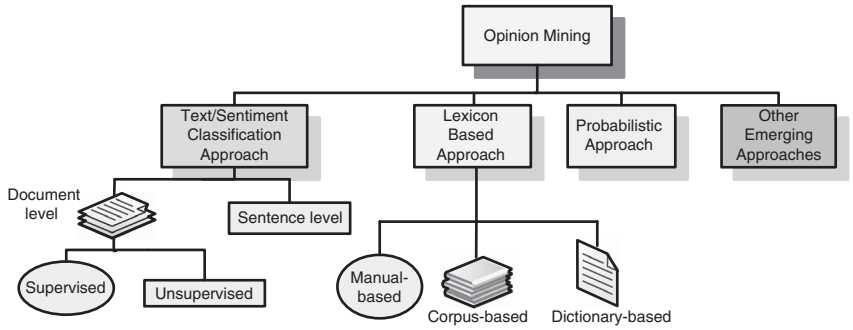


Figure 1 Categories of opinion retrieval approaches

at the University of Glasgow, United Kingdom. Majority of the proposed opinion retrieval techniques are evaluated on TREC blog datasets using the standard TREC blog baselines. The TREC blog dataset contains collection of blog posts from 2006 to 2008. The TREC baselines are usually selected as the best techniques in each year's TREC blog track participation on opinion retrieval task, which is performed on the standard TREC blog dataset.

Alongside opinion retrieval task, the TREC blog track also focuses on other research tasks. These include splogs (spam blogs) detection and/or spam comments detection from blogs, blog distillation, and top stories/news identification. Recent TREC blog tasks include faceted blog distillation (a blog feed search task) and top stories identification. Other than TREC blog datasets, some techniques have evaluated their performance on other domain-specific datasets such as Internet Movie Database (IMDb) dataset for retrieving sentiments from movie reviews, Amazon[5] product reviews for retrieving sentiments on consumer products, and Usenet[6] newsgroups for retrieving sentiments from news. For each opinion retrieval approach, we will show performance comparison on TREC blog datasets alongside other domain-specific datasets.

The selection of dataset for opinion retrieval task may vary according the proposed technique or a domain of interest. However, it is advisable that any proposed opinion retrieval technique should be evaluated on a standard dataset whereby the proposed technique can be easily compared with the available standard baselines.

Finally, performance evaluation of opinion retrieval techniques involves the use of standard information retrieval (IR) evaluation metrics such as Mean Average Precision (MAP) for measuring retrieval precision over a set of queries, precision at K (e.g. P@5 and P@10) for measuring retrieval precision at a fixed level of results, relevance precision (R-Precision) for measuring the best precision from a set of relevant documents, and binary preference (Bpref) that considers non-relevant and relevant documents only from a set of judged documents from the retrieved results. The choice of evaluation metrics may, however, vary according to the proposed opinion retrieval techniques or according to the selected baselines. In any case, the selected evaluation metric must be part of the standard IR evaluation metrics. Researchers are therefore referred to background information on traditional IR evaluation metrics (Manning *et al.*, 2009).

1.3 Overview of existing opinion retrieval techniques

Opinion retrieval techniques can be broadly classified into four categories as shown in Figure 1.

Early research works treated opinion retrieval as *text classification* (TC) problem thereby using machine learning (ML) approach to classify text in documents that contain opinions. The TC technique enables the introduction of both traditional *supervised* and *unsupervised* ML techniques with reasonable efficiencies recorded in most research works (Ounis *et al.*, 2008). *Lexicon-based*

[5] http://www.amazon.com/

[6] http://www.usenet.com/

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

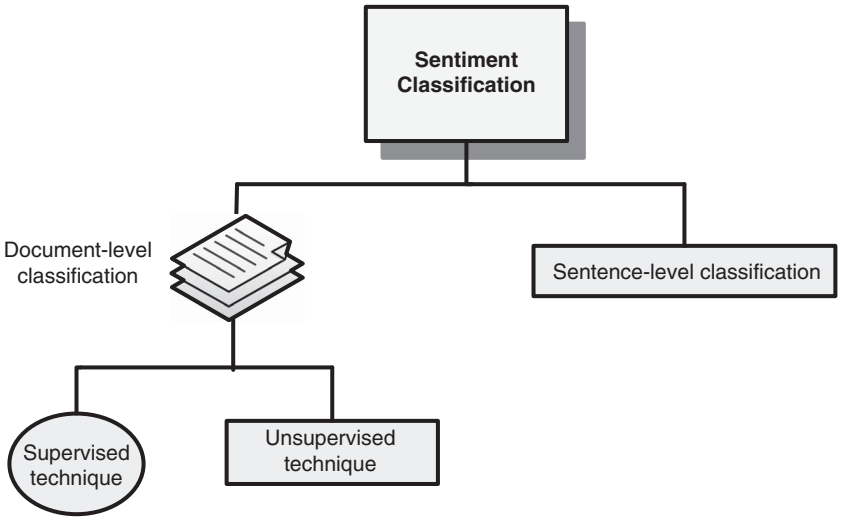


Figure 2 Structural overview of sentiment classification approach

approach was introduced with the aim that it can compensate for the inefficiencies of the TC techniques, and many research works have combined both lexicon-based and TC techniques for effective opinion retrieval (Lee *et al*., 2008; Zhang & Ye, 2008; Bollen *et al*., 2010). *Probabilistic approaches* have also been used to retrieve and rank opinions from documents using statistical inferences (Gerani *et al*., 2010). The idea is to automatically detect large collection of documents that contain opinion without any learning technique. This technique has performed reasonably in ranking opinionated documents from a huge collection. Finally, because it was difficult for the three existing approaches to outperform the standard TREC baselines, *other emerging* opinion retrieval approaches have been proposed with the hope of outperforming the standard baselines (Du & Tan, 2009a, 2009b; Thet *et al*., 2009). Some of the approaches are either completely independent of the three main approaches highlighted earlier or combine some of the main approaches with other newly formed techniques. In the following sections, we will describe the four categories in detail.

2 Text classification approach

2.1 Overview of the approach

The TC approach for opinion retrieval is generally termed *sentiment classification* (Pang & Lee, 2008; Liu, 2010). In this case, a classifier is built based on the occurrences of opinion words and linguistic features (semantic and syntactic) to identify if content from a particular document is subjective or not (Jia *et al*., 2008; O'Hare *et al*., 2009; Siersdorfer *et al*., 2010). For example, comments to political news can be classified as either 'positive' or 'negative' having trained an opinion retrieval system with both positive and negative words. However, the TC approach to opinion or sentiment classification differs from the generic topic-based Web or document classification. In the generic topic-based classification, the amount of query words present in the retrieved documents is important. On the contrary, in opinion or sentiment classification, the amount of subjective words in the retrieved documents is important. Sentiment classification appears to be frequently studied by using different methods within the traditional TC approach. This has attracted a huge number of research works recently (Zhang & Yu, 2006; Zhang & Zhang, 2006; Joshi *et al*., 2006; Jia *et al*., 2008; Gerani *et al*., 2009; Jin *et al*., 2009; Kobayakawa *et al*., 2009; O'Hare *et al*., 2009; Taboada *et al*., 2009; Siersdorfer *et al*., 2010). We will discuss sentiment classification and its sub-classes as shown in Figure 2. We will also identify limitations in using the TC approach for opinion or sentiment retrieval tasks, hence propagating the need for other opinion retrieval approaches.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

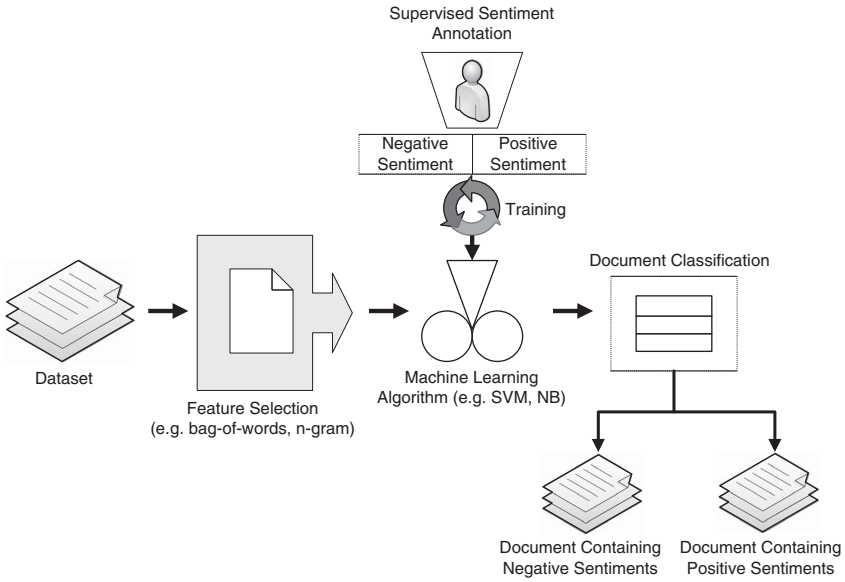


Figure 3 Component architecture for supervised document-level opinion/sentiment classification

Sentiment classification include *document-level* sentiment classification and *sentence-level* sentiment classification (Lin *et al*., 2006). The two classification types differ by input to the respective algorithms. In document-level classification, a document (input) is classified to differentiate between the polarities contained in the document (e.g. positive, negative, and objective). In sentence-level classification, each sentence (input) in a given document is analyzed to show subjectivity or objectivity. We will discuss the two techniques in detail in the following section.

2.2 Document-level classification for opinion retrieval

In document-level classification, documents in a given collection are classified as either 'positive' or 'negative' or in some cases 'objective' with the assumption that the documents contain opinion (Pang *et al*., 2002; Liu, 2010). However, this particular type of classification may only be applicable to documents that focus on a unique or independent discussion, for example, *news review* (Koppel & Shtrimberg, 2006). In this case, the opinion retrieval system is able to classify the news (an entire document) as 'positive' or 'negative' or 'objective'. In some cases, some domain-dependent classification may require documents to be classified as 'bad' or 'good' (Koppel & Shtrimberg, 2006; Lin *et al*., 2006). Also, like the traditional topic-based text classification (Baharudin *et al*., 2010), sentiment classification can be categorized as using *supervised* and *unsupervised* methods (Liu, 2010). We will outline some opinion retrieval techniques that fall under these methods and highlight their performance and limitations on different datasets.

2.2.1 Supervised document-level sentiment classification

This is similar to the traditional topical text classification. Supervised document-level sentiment classification can be achieved by training and testing documents with focused opinions (Liu, 2010). For example, product review with *thumbs-up* and *thumbs-down* (Siersdorfer *et al*., 2010), or *star ratings* (Lim *et al*., 2010). In contrast to understanding *topical orientation* in text classification, opinions or sentiments about the document are derived from opinionated features such as *good*, *beautiful*, *notorious* among others (Nguyen *et al*., 2010). A description of features that can be used for effective document-level classification can be found in Pang & Lee (2008) and Liu (2010). Also, research works such as Pang *et al*. (2002), Lin *et al*. (2006), Zhang and Zhang (2006), Jia *et al*. (2008), Lee *et al*. (2008), Jin *et al*. (2009), Kobayakawa *et al*. (2009), O'Hare *et al*. (2009), Siersdorfer *et al*. (2010), and Nguyen *et al*. (2010) have applied classification algorithms such as support vector machines (SVM)

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

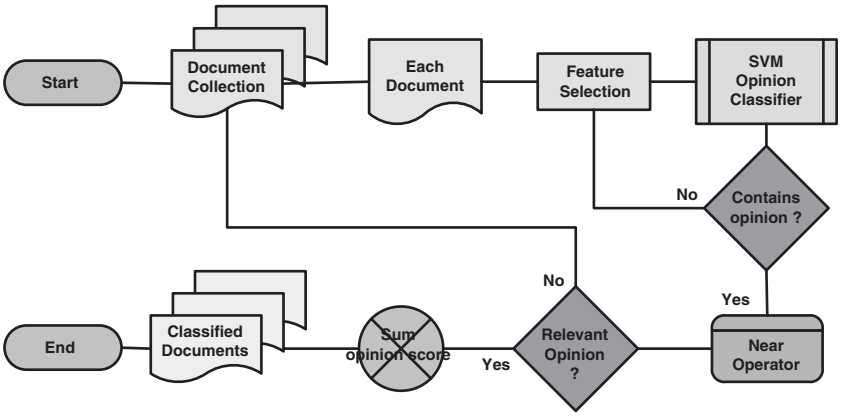


Figure 4 Flow chart describing the algorithm in Jia *et al.* (2008)

(Sun *et al.*, 2002) and Nave Bayes (Zhang *et al.*, 2009) for document-level sentiment classification. The results of these works show classification using SVM gave best performance in most of the experiments. We will discuss an example state-of-the-art supervised technique for opinion retrieval (Figure 3).

2.2.2 Example work (supervised document-level opinion/sentiment classification)

Jia *et al.* (2008)

In this work, a four-step procedure was used for document-level opinion retrieval task at TREC blog track 2008. In the first step, documents that are relevant to the given query are selected based on term or concept similarity. In the second step, common abbreviations are identified to enhance the efficiency of the opinion retrieval technique. In the third step, SVM is used to classify relevant opinions from sentences. Finally, documents are scored and ranked based on the degree of relevance shown by the previous steps. Figure 4 describes the flow of the algorithm.

The work is an improvement on Zhang and Zhang (2006), which was proposed to solve sentence-level opinion retrieval problem. The concept of *NEAR* operator was introduced to determine the degree at which subjective sentences are relevant to the given query. The *NEAR* operator technique simply checks whether proximity of query terms to each subjective sentence occur at a five-sentence window. In the end, the overall opinion polarity of a document is determined based on the degree of relevance shown by using the *NEAR* operator. The result of their experiment recorded 0.4461 and 0.4473 MAPs with a slight improvement of 0.01 and 0.02 over the given baseline. It also recorded 0.4822 and 0.4822 R-Precisions in two different runs, respectively.

Discussion

Despite its significance in the opinion research community, the performance result of the above work shows that it is very unlikely that learning classifiers to retrieve opinions would outperform baselines significantly. A problem with sentiment classification task is the identification of subjective sentences according to the domain for the purpose of training the classifiers. This is still an active area of research. For example, in the iPhone domain, a subjective sentence can read 'The iphone lacks some basic features', whereas in the movie domain, a subjective sentence can read 'The movie is not so interesting'. If a classifier was trained on the iphone domain, such classifiers would have low performance on movie domain as movies cannot *lack basic features*. Therefore, for multiple domain opinion retrieval, it becomes inappropriate to train classifiers based on the subjective expressions from a single domain. Another problem we observed is the use of word-concept relationship to identify subjective sentences. There is need to establish whether such words must occur independently without any dependency on other words in proximity or in the original sentence. For example, in the sentence 'she is amazingly beautiful', the word *beautiful* in this context depends on the word *amazingly* in order to determine the degree of subjectivity of the sentence. However, we think the word *beautiful* and *amazingly* have been considered independently.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

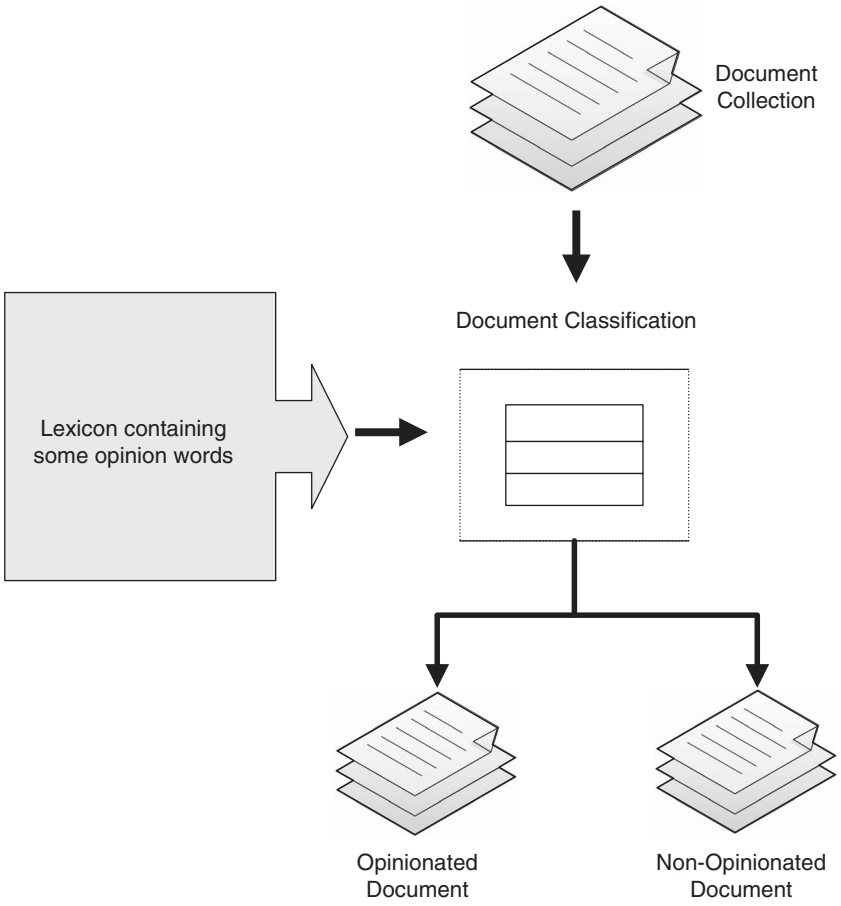


Figure 5 Component architecture for unsupervised document-level opinion/sentiment classification

2.2.3 Unsupervised document-level sentiment classification

Another approach to sentiment or opinion classification is automatic document-level classification (Lin *et al*., 2006). The literature generally refers to it as unsupervised sentiment classification (Pang & Lee, 2008; Liu, 2010). Unlike the supervised approach, the unsupervised approach does not require manual labeling of documents for training, thus it is not labor intensive. Documents that contain opinions are systematically and automatically categorized according to the degree of subjectivity. Recently, some techniques have combined topic model with sentiment model to form a joint sentiment/topic model. For example, Lin and He (2009) proposed a joint sentiment/topic model (JST) by adding a sentiment layer to the popular state-of-the-art topic model, Latent Dirichlet Allocation (LDA; Blei *et al*., 2003). The idea is to detect both sentiments and topic simultaneously from documents unlike the supervised techniques that detect both topic and sentiment as a two-stage process. JST is fully unsupervised and without explicit knowledge of labeled sentiment classes. The approach has similar performance with a popular sentiment classification technique (Pang *et al*., 2002). Other similar sentiment classification approaches using topic model include Lu *et al*. (2011) and Paul and Girju (2010). Both techniques detect multiple aspects of sentiments and topics using various topic models, including LDA.

A technique called lexicon creation has also been used for unsupervised document-level sentiment classification. A lexicon contains collection of words regularly used in a particular domain, thus lexicons are created for different opinion domains. Some lexicons are created based on different categories of state of feelings such as emotion, thoughts, and expressions. An example is the Linguistic Inquiry and Word Count (LIWC), which assign words into psychologically meaningful categories (Tausczik & Pennebaker, 2010). LIWC was originally developed to identify

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

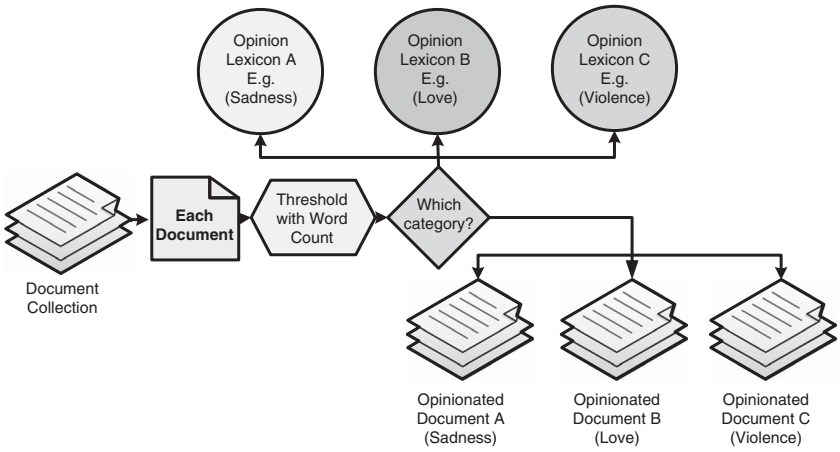


Figure 6 Component description of using LIWC for unsupervised opinion detection technique. LIWC = linguistic inquiry and word count

positive and negative emotion words within documents. It later considered identifying categories of words that show self-reflections and simple opinions. Some lexicons are also created to determine opinion or sentiment polarity of words or sentences. An example is SentiWordNet (Esuli & Sebastiani, 2006). SentiWordNet determines positive, negative, and neutral values for words. Such values are summed over words in a given sentence to determine the polarity of the sentence or its subjective orientation (Esuli, 2008). We will discuss a top performing opinion retrieval technique, which use unsupervised approach for document-level classification (Figure 5).

2.2.4 Example work (unsupervised document-level opinion/sentiment classification)

Tausczik & Pennebaker (2010)

The work uses LIWC to empirically show if a certain word or a subject of discussion belongs to an opinion category. For example, consistent use of first person personal pronouns such as 'I' and 'Me' emphasizes on a *subject of attention* in the given sentence. Also, the use of *positive emotion* words such as *like*, *interesting* and *good* are used in writing about positive events. Therefore, once each word in the document is identified with a category, the percentage of each category is calculated and identified with a psychometric fact gathered through empirical study, such that the leading category describes the likely opinion contained in the document (Figure 6).

For example, Chung *et al*. (2008) used LIWC to discover the level at which individual word or phrase can predict success and failure of weight loss as mentioned in blogs. The LIWC technique claimed significant success at predicting weight loss success by higher percentage of sadness words within the blog.

Discussion

The LIWC technique is dictionary-based and it requires human prior knowledge or assumption for respective sentiment categories that can be associated with each given word. The prior knowledge of the human judges also depends on their perceptions or experiences with the use of such words, which can vary greatly among humans. A major limitation observed in this technique is the use of word counts to represent opinion without considering the grammatical context at which each word had occurred (Krahmer, 2010). In Chung *et al*. (2008), it is unclear whether individual words or phrase matched with 'sadness' lexicon category had specifically occurred in sentences particular about weight loss or failure. For example, these two sentences, '*the fight for academic success*' and '*I will fight you to finish*', have regular occurrence of the word 'fight', which may imply *violence* as an opinion target after a certain frequency threshold, whereas, the word 'fight' has appeared in two different contexts, respectively. That is, the sentence 'the fight for academic success' may imply 'passion for academic excellence', and the sentence 'I will fight you to finish' may imply 'violence'.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Therefore, for a more effective unsupervised opinion or sentiment retrieval, we suggest context-driven subjectivity (subjective words must occur in the context wanted by the user). This can be achieved by using natural language processing or grammatical analysis to differentiate between the orientations of expressions. A good example is demonstrated in Hatzivassiloglou and McKeown (1997) and similar work is also presented in Turney and Littman (2002).

2.2.5 Research issues in supervised and unsupervised document-level classification

The contributions of document-level classification toward opinion retrieval have recorded some level of success (Turney & Littman, 2002; Jia *et al*., 2008; O'Hare *et al*., 2009; Siersdorfer *et al*., 2010). However, some limitations involved in both supervised and unsupervised approaches show the need for more reliable approaches, which can provide effective opinion or sentiment retrieval.

In *supervised document-level sentiment classification* such as Pang *et al*. (2002), the manual labeling of words or phrases with a particular polarity is labor intensive. SentiWordNet (Esuli & Sebastiani, 2006) is now playing an active role in this regard. Also, according to Pang and Lee (2008), subjectivity is a two-state task and can be interpreted differently in some cases. It comes with different levels of challenges, such as 'relevant' or 'not relevant' (Lee *et al*., 2010), 'agreement' or 'disagreement' (Bermingham & Smeaton, 2009), and 'winner' or 'looser' (Tumasjan *et al*., 2010). However, it is still very challenging to effectively determine appropriate subjectivity state in many documents that contain opinion.

Also in Sarmento *et al*. (2009), the use of ironic phrases and inverted polarity in opinionated documents led to lower precision for positive opinions with 77% accuracy. This resulted from the difficulty in determining polarity in opinionated documents (Abbasi *et al*., 2008; Pang & Lee, 2008; Choi *et al*., 2009). An approach to solve this problem is to consider semantic analysis at phrase and sentence level (Kobayakawa *et al*., 2009). Natural language processing techniques such as word-sense disambiguation using Wikipedia and Wiktionary may also be helpful to understand the orientation of opinion contained in a given document (Torsten *et al*., 2009).

Another problem with supervised document-level sentiment classification is feature selection (Gamon, 2004). In product and news reviews, opinions or sentiments can be easily polarized as 'positive' or 'negative' and 'good' or 'bad', respectively (Liu, 2010), thus it is very easy to select positive and negative features for training. However, in opinion sources such as blogs, subjective discussions can be highly diverse (Agarwal & Liu, 2008; Pang & Lee, 2008; Liu, 2010); therefore, selecting training features from such documents for opinion or sentiment classification becomes a daunting task.

In *unsupervised document-level sentiment classification* manual creation of lexicon by set of human judges is often a challenging task (Krahmer, 2010; Tausczik & Pennebaker, 2010). It suffices to say opinion categories such as self-reflection, perception, feeling, and emotion may be too abstract to be detected by simple opinion word count or presence of certain opinion words. Since words are commonly dependent on other words in a sentence (e.g. adjectives that modifies nouns and arguments that describe predicates), a better approach is to consider word–word dependency in the document. For example, proximity between words (Santos *et al*., 2009; Gerani *et al*., 2010) and concept similarity (Jia *et al*., 2008) can be used to further understand subjectivity in phrases or sentences of documents that contain opinion (Table 1).

2.2.6 Performance comparison of document-level techniques across different datasets

The expression of opinions or sentiments differs according to the domain. This is shown in the above comparison table. For classification-based techniques, it is practically challenging to use the same set of features across different domains due to the way people express opinion or sentiments on different domains. However, it could be easier to compare between works that have used the same dataset. For example, Pang *et al*. (2002) and Whitelaw *et al*. (2005) used the same IMDd movie reviews. Whitelaw *et al*. (2005) had improved performance due to an automatic extraction of appraisal features that includes orientation of appraisal adjectives (e.g. very beautiful). Since opinions and sentiments are expressed differently in the datasets, it is almost certain that the

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Table 1 Performance comparison for *document-level* techniques across different datasets

Author	Approach	Dataset	Performance metric	Results
Pang *et al.* (2002)	Supervised	IMDb movie reviews	Classifier accuracy (three different experiment runs)	(i) 78.7% (ii) 77.7% (iii) 82.9%
Zhang and Yu (2006)	Supervised	TREC Blog 2006	MAP, G-MAP, R-Prec, and P@10, respectively	0.1636, 0.0921, 0.2522, and 0.4380
Lin *et al.* (2006)	Supervised	Bitterlemons corpus	Classifier accuracy (two different experiment runs)	(i) 0.9493 (ii) 0.8689
O'Hare *et al.* (2009)	Supervised	Financial blog corpus	Classifier accuracy (three different experiment runs)	(i) 68.2383% (ii) 70.7022% (iii) 74.3683%
Whitelaw *et al.* (2005)	Unsupervised	IMDb Movie Reviews	Classifier accuracy (two different experiment runs)	(i) 87.0% to 87.6% (ii) 90.1% to 90.2%
Fei *et al.* (2004)	Unsupervised	Sports Reviews	Classifier accuracy	86%
Gamon (2004)	Unsupervised	Customers feedback	Classifier accuracy, *F*-measure 'good' and *F*-measure 'bad', respectively	85.47%, 74.62% and 89.82%
Nguyen *et al.* (2010)	Unsupervised	Livejournal	*F*-score (two different experiment runs)	(i) 77.4% (ii) 78.8%

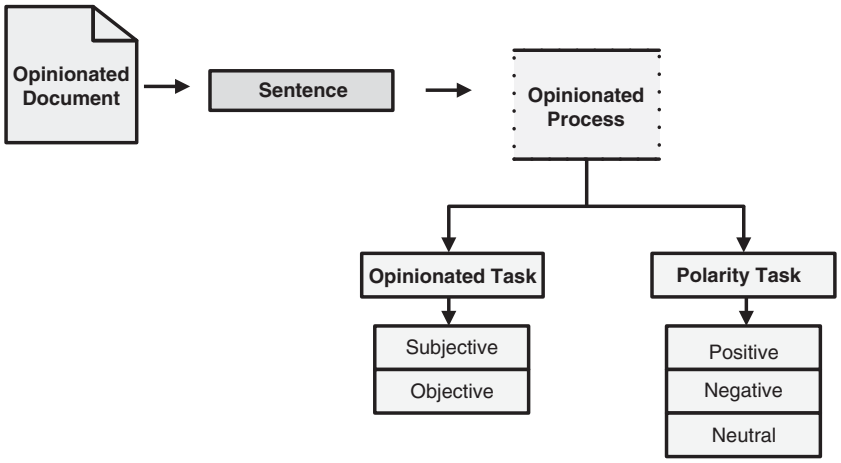


Figure 7 Component architecture for sentence-level classification of opinionated and polarity tasks

performance will vary due to the nature of feature extracted to train classifiers. Overall, domain-specific unsupervised techniques are likely to outperform supervised techniques because of their ability to automatically use semantic and syntactic features, see Gamon (2004).

2.3 Sentence-level classification for opinion retrieval

Sentence-level classification of opinion goes beyond identification of individual opinionated words as studied in some document-level opinion classification tasks (e.g. unsupervised technique). In sentence-level opinion classification, rather than individual words, each sentence in a given document is analyzed and checked to be subjective (Narayanan *et al.*, 2009). There are basically two tasks involved in sentence-level classification. These include *subjectivity or opinionated classification* and *sentence-level sentiment or polarity classification* (Jindal & Liu, 2006; Lin *et al.*, 2006). In subjectivity or opinionated classification, a sentence is checked for subjectivity (contains opinion) or factual (does not contain opinion) (Zhang & Yu, 2006; Ganapathibhotla & Liu 2008; Jin *et al.*, 2009; Santos *et al.*, 2009; Zettlemoyer & Collins, 2009). In sentence-level sentiment or polarity classification, a sentence is checked for *positive* or *negative* or *neutral* sentiment polarity (Jia *et al.*, 2009; Narayanan *et al.*, 2009). Figure 7 shows the component architecture of sentence-level classification for both opinionated and polarity tasks. We will discuss an example state-of-the-art work that has used sentence-level classification.

2.3.1 Example work (sentence-level classification for opinion retrieval)

Zhang and Zhang (2006)

This work performed an opinionated classification task. It is also among the top performing techniques highlighted in TREC 2006 blog track (Ounis *et al.*, 2006). The work combines different IR and NLP techniques to achieve a two-step approach for opinion retrieval system. In the first step, the given query is pre-processed to determine relevant documents. In the second step, documents that contain opinions are retrieved.

The first step involves two query pre-processing techniques (e.g. query expansion) to ensure broader scope of retrieval results (Huang & Efthimiadis, 2009). However, we suggest that the choice of query expansion must be carefully decided. Query expansion techniques may change the meaning of the original query drastically (Santos *et al.*, 2010). This negatively affects the intent of information need and thereby leads to low precision.

The first pre-processing technique used in Zhang and Zhang (2006) is based on concept identification and retrieval from the given query. This involves an algorithm that combines some NLP tools such as Minipar (Lin, 1998), WordNet (Fellbaum, 2010), Wikipedia (Torsten *et al.*, 2009),

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Collins Parser (Collins, 1996), and Google[7] Search Engine. Further explanation of the algorithm is provided in (Zhang *et al*., 2007b). The second query pre-processing technique combines traditional query expansion using pseudo-relevant feedback (Yu *et al*., 2003), query segmentation using Wikipedia (Tan & Peng, 2008), and query expansion using web feedback (Huang & Efthimiadis, 2009).

The second step involves retrieval and ranking of relevant documents regarding the pre-processed query. The technique combines *concept similarity* derived by using Wikipedia concepts to measure similarity between query and documents, and *term similarity* derived by using Okapi formula (Robertson & Walker, 1999). The work gave high priority to the concept similarity and the concept similarity algorithm is described in Liu *et al*. (2004). Sentence classifier is trained using subjective and objective sentences derived by using Wikipedia[8] and Rateitall.com[9]. Sufficient description for selecting subjective and objective sentences using the two medias is presented in Zhang and Zhang (2006) itself. Two relevance ranking and scoring methods were used in this technique, the first method retrieves only documents that contain opinion, and the second method determines sum of the classification scores for relevant sentences within the retrieved documents. We will show the two scores, respectively:

$$Score_{rank}(D,Q) = Sim(D,Q) \times I(D,Q) \tag{1}$$

where $\mathrm{Sim}(D,Q)$ is the relevant score for retrieving each document. $I(D,Q)$ equals 1 if the document contains relevant sentences and 0 otherwise:

$$Score_{rank}(D,Q) = \sum_{S \in \{opinionated\ relevant\ sentences\ in\ D\}} [Score_{classification}(s) \times relevant(s,Q)] \tag{2}$$

where $Score_{classification}(s)$ is the SVM classifier score for each sentence and $relevant(s,Q) = 1$ provided the sentence meets all relevance criteria. Using scoring method highlighted in Equation (1), the technique achieved MAP value of 0.1636 and P@10 value of 0.4380. Using the scoring method highlighted in Equation (2), the technique achieved MAP value of 0.1885 and P@10 value of 0.5120

Discussion

The principal concern is the impact of the query expansion techniques on the nature of opinion retrieved. It is very likely that the use of query expansion techniques may lead to poor performance in terms of context-dependent opinion retrieval. For example, using query expansion techniques on user's query logs, previous queries might be for different query intents, therefore affecting the context of the opinion to be retrieved. Also, the computational power required to combine the two query pre-processing techniques may be high. For example, using pseudo-relevance feedback for query expansion amount to very high computational power and many research works have shown concerns about the technique (Jones *et al*., 2006). More importantly, the use of Wikipedia-based concept similarity technique is still an active area of research, thus it is still difficult to find a benchmark to weigh its performance and contribution to opinion retrieval.

2.3.2 Research issues in sentence-level classification

The main issue in opinion or sentiment classification is the insufficient or domain-independent training features. Some research works have addressed the problem (Abbasi *et al*., 2008; Pang & Lee, 2008). In opinion retrieval, feature selection can be tedious especially in sentence-level classification where each sentence need to be annotated manually for the classifier (Liu, 2010). In fact, one major concern is how the training features can be consistently identified for a well-diverse opinionated document, which is likely to have sentences with diverse concepts or multi-concepts (Pang & Lee, 2008). We think opinion detection in sentences is a natural language problem as

[7] http://www.google.com

[8] http://www.wikipedia.org/

[9] http://www.rateitall.com/

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

many opinions may be too abstract to be detected without proper linguistic analysis. This makes research in opinion retrieval more challenging as opinion retrieval task requires various NLP pre-processing techniques in order to ensure qualitative results (e.g. word sense disambiguation and syntactic analysis; Zhang & Zhang, 2006; Santos *et al.*, 2009).

In addition, some sentences can be too ambiguous for complete opinion retrieval processes (e.g. idioms, metaphor, and sarcasm). This makes it difficult to retrieve absolute opinion that a document may contain. Some research works avoid such sentences to avoid ambiguities in experiments. However, this omission means that the complete opinions or sentiments in the document cannot be fully realized, thus creating bias in the overall opinion retrieval process (Sarmento *et al.*, 2009). Therefore, retrieving opinions or sentiments from ambiguous sentences could be an active area of research (Table 2).

2.3.3 Performance comparison of sentence-level techniques across different datasets

The sentence-level techniques also attracted different datasets like the document-level techniques. However, the use of TREC Blog datasets (i.e. TREC Blog 06, 07, and 08) was persistent as a result of the growing influence of the TREC Blog tracks. Sentence-level techniques seem to perform relatively low on TREC blog datasets. The reason could be because TREC Blog datasets have multiple domains (i.e. it contains blog posts discussing on different domains such as music, movies, products, leisure, etc.), thus there is indeed the possibility of inappropriate feature selection for all the domains that might exist in the documents. Again, unlike techniques used on multi-domain datasets, we think domain-specific opinion or sentiment classification techniques are more likely to have better performance due to familiar and fixed features used to train the classifiers. If we must compare between techniques used on TREC Blog dataset, it will be appropriate to compare between techniques that have used TREC blog dataset of the same year and also with the same performance metrics. For example, Jia *et al.* (2008) used their supervised technique on TREC Blog 2008 and outperforms both Santos *et al.* (2009) and Zhang and Zhang (2006); Zhang and Ye (2008) techniques using TREC Blog 2006 and 2007 datasets, respectively. While comparison cannot be made directly due to the difference in the years of the datasets (TREC datasets differs by number of blog posts over the period of collection; Macdonald *et al.*, 2010), we think Jia *et al.* (2008) is an effective algorithm to retrieve substantial opinions at sentence level.

3 Lexicon-based approach

3.1 Overview of the approach

The presence of opinion words in documents creates an easy process for selecting and combining such words to be used as lexicon for automatic opinion or sentiment retrieval tasks. The process is called *lexicon* or *dictionary generation* (Esuli & Sebastiani, 2006; Lee *et al.*, 2008; Liu, 2010). The lexicon generated can then be used to identify opinion or sentiments in documents that contain such opinionated words. This process is called lexicon-based opinion or sentiment retrieval. A lexicon-based approach is commonly used when a learning-based technique is to be improved by domain specific opinion words. This has improved domain-specific opinion retrieval (Kanayama & Nasukawa, 2006; Yang *et al.*, 2007; Ding *et al.*, 2008; He *et al.*, 2008; Tan *et al.*, 2008; Zhang & Ye, 2008; Du & Tan, 2009a; Na *et al.*, 2009; Santos *et al.*, 2009; Pan *et al.*, 2010; Vechtomova, 2010). It has also shown significant improvements in detecting opinions that are commonly expressed with adjectives (e.g. beautiful, bad) or adverbs (e.g. severely, permanently; Ding *et al.*, 2008; Strapparava & Mihalcea, 2008), thereby making it suitable for domain-independent opinion retrieval as well. Similar technique is also used in Tausczik and Pennebaker (2010), constructing a lexicon of 80 different sentiment categories. The sentiment categories include words that describe personal reflection, personal emotion, and personal feelings. Other widely used lexicons for opinion retrieval include General Inquirer (Stone *et al.*, 1966) and Chinese Network Sentiment Dictionary (CNSD) or NTU Sentiment Dictionary (Ku *et al.*, 2006). Figure 8 shows the structural overview of the lexicon-based opinion retrieval approach.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Table 2 Performance comparison for *sentence-level* techniques across different datasets

Author	Approach	Dataset	Performance metric	Results
Jia *et al.* (2008)	Supervised	TREC Blog 2008	MAP and R-Precision, respectively (two different experiment runs)	(i) 0.4461, 0.4473 (ii) 0.4822, 0.4822
Siersdorfer *et al.* (2010)	Supervised	YouTube Comments	Precision and recall, respectively	0.8598 and 0.4
Fortuna *et al.* (2007)	Unsupervised	Usenet newsgroups	Classifier accuracy	85%
Zhang and Ye (2008)	Unsupervised	TREC Blog 2006 and 2007	R-Accuracy	0.1619, 0.2517
Jin *et al.* (2009)	Unsupervised	Amazon review and Hus corpus	Precision, recall, and *F*-score, respectively (two different experiment runs)	(i) 91.41%, 85.81%, 88.52% (ii) 85.58%, 69.17%, 76.50%
Kobayakawa *et al.* (2009)	Unsupervised	Random TV show opinion	Classifier accuracy (two different experiment runs)	(i) 100% (ii) 78%
Ganapathibhotla and Liu (2008)	Unsupervised	Product reviews and forum posts	Precision, recall, and *F*-score, respectively	0.967, 0.961, and 0.964
Santos *et al.* (2009)	Unsupervised	TREC Blog 2006	MAP, R-Prec, bPref, and P@10, respectively	0.2990, 0.3432, 0.3644, and 0.6060
Lloret *et al.* (2010)	Text summarization	Bank reviews	*F*-measure	0.563

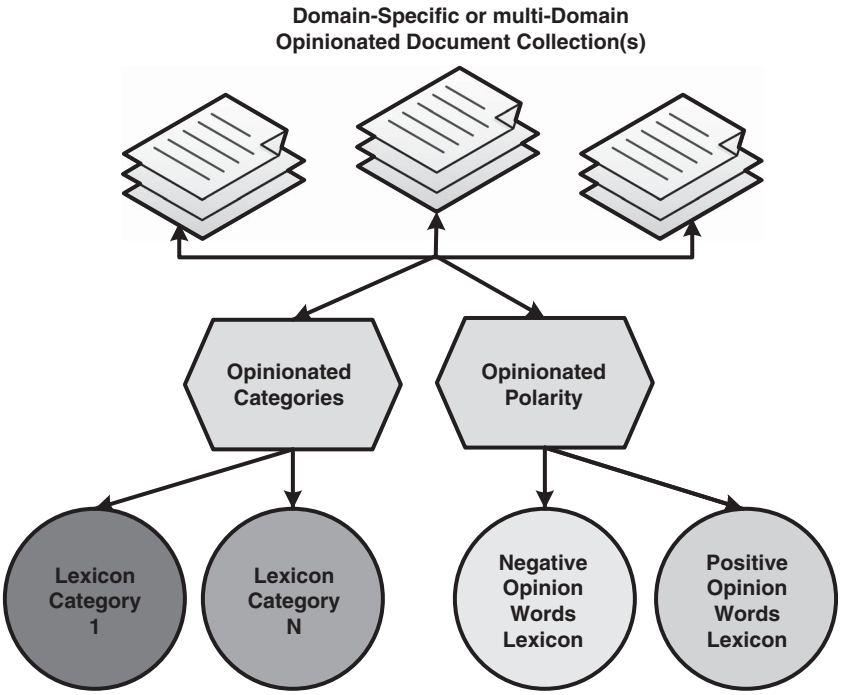


Figure 8 Structural overview of lexicon-based opinion retrieval approach

It is also possible to create polarity-based lexicon for opinion or sentiment retrieval (e.g. SentiWordNet) (Esuli & Sebastiani, 2006). SentiWordNet is a publicly available lexical resource that shows appropriate numeric polarity values for positive, negative, and objective words. It has made significant contributions in lexicon-based opinion or sentiment retrieval (Gerani *et al*., 2010; Siersdorfer *et al*., 2010). In many lexicon-based opinion retrieval techniques, opinion words with positive orientation are used to denote *agreement*, while opinion words with negative orientation are used to denote *disagreement*. However, the presence of ambiguous sentences in documents that contain opinion has been one of the limitations of the lexicon-based approach. It is often challenging to understand the orientation of ambiguous words or to determine whether the words contain opinion or not. For example, there has been limited success in retrieving opinion from phrases of inverted polarities, irony, idioms, and metaphor (Sarmento *et al*., 2009).

3.2 Example work (lexicon-based opinion retrieval)

Ding *et al*. (2008)

In this work, a holistic lexicon-based approach is used to determine sentiment orientation of product reviews using external evidence and linguistic rules. This work addresses the problem of context-dependent semantic orientations of opinion words. It also addresses the methodology for combining opinion words found within the same sentence. The idea is to use the opinion words at proximity to each product feature in a sentence to detect the orientation of the opinion. The main steps involved in their algorithm are described as follows:

1. Determine opinion words or phrases that express negative and positive opinion.
   a. Use opinion lexicon comprising of adjectives, adverbs, nouns, and verbs.
   b. Use list of context-dependent opinion words.
   c. Perform part-of-speech tagging for opinion words in the lexicon.
   d. Identify positive, negative, and dependent idioms.
   e. Identify non-opinion phrases containing opinion words.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

2. Aggregating opinions for a product feature.
   a. Using the above list of context-dependent opinion words, identify positive, negative, and neutral opinion orientation on product feature in each sentence.
   b. The opinion orientation score is computed as $score(f) = \sum_{w_i.w_i \in S \wedge w_i \in V} \frac{w_i.SO}{dis(w_i,f)}$
   c. Where w_i is the opinion word, V is the set of all opinion words and idioms, s is the sentence that contains product feature f, $dis(w_i, f)$ is the distance between feature f and opinion word w_i, and $w_i.SO$ is the semantic orientation of the opinion word w_i.
   d. Using the *multiplicative inverse* in the formula, low weights is assigned to opinion words that are far away from the feature f.
   e. With a positive score, the opinion on the feature in the given sentence is positive, otherwise negative.
3. Determine context-dependent opinions using linguistic rules.
   a. Use intra-sentence conjunction rule which shows that a sentence only expresses one opinion orientation.
   b. Use pseudo-intra (indirect) sentence conjunction rule to detect implicit conjunction in opinionated sentences.
   c. Should (*a*) and (*b*) not hold, use inter sentence conjunction rule to determine opinion orientation for the previous or next sentences.
   d. Use synonym and Antonym rule to identify implicit opinion orientation.

Discussion

The lexicon-based work described above shows a detailed algorithm that can capture the orientation of opinions regardless of the domain. The strength is particularly in the context-dependent component, which ensures opinion orientation is detected in the context needed by the user. This is an advantage over a classification-based opinion or sentiment retrieval technique that uses random frequency of words to train its classifier. Some recent works have also introduced domain specific techniques for detecting opinion or sentiments from a specific domain. This is quite effective as opinion words mostly used in the domain are also used to create the needed lexicon. For example, an attempt to retrieve opinions or sentiments from digital camera sales review may require creating lexicon that consist of words or terminology commonly used in the digital camera sales business (e.g. megapixel, memory, and lens), see Du and Tan (2009a).

3.3 Research issues in lexicon-based approach

Generated lexicons are mostly based on heuristic rules that limit the words in the lexicon to certain lexical categories. For example, some lexicons only contain *adjectives* and *adverbs* (Hu & Liu, 2004). The lexicon-based approach introduced in Ding *et al.* (2008) considered external evidences and the analysis of natural language expressions to understand the semantic orientation of opinion words. However, the work explicitly relies on set of certain opinionated words selected based on heuristic rules to generate the lexicon. Using the heuristic rules, opinionated words are selected from each sentence and combined to form lexical features. There are two major concerns about this technique. First, although some research works have shown significant improvements by considering adjectives and adverbs as part of their retrieval strategies, we think it rather limits the scope of sentiment or opinion that can be retrieved. Second, it is worth studying a model that finds a common relationship between all the lexical and semantic resources in a given sentence. This could include function words, which contribute greatly to the context or orientation of a sentence (Krahmer, 2010). We believe opinionated words derived from a certain sentence may not completely and independently express the overall opinion or sentiment in the sentence. Consideration must be given to inter-dependencies between function words and opinion words toward the overall opinion. For example, *object to subject* linguistic relationships that may exist between words in the sentence may be considered (Krahmer, 2010).

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

3.4 Performance comparison of lexicon-based techniques across different datasets

The use of TREC Blog dataset is consistent as shown in the comparison table above. In terms of MAP, one could make direct comparison between He *et al.* (2008), Zhang and Ye (2008), and Na *et al.* (2009) since they all use the same TREC Blog 2006 dataset. He *et al.* (2008) outperforms Zhang and Ye (2008) by almost 40% and outperforms Na *et al.* (2009) by almost 16%. We think the significant performance by He *et al.* (2008) is due to the statistical approach that automatically detect evidence of subjectivity. On the other hand, Zhang and Ye (2008) performed significantly with TREC blog 2007 topics using a generation model that unifies topic relevance and opinion generation through a quadratic combination. Another interesting observation is that opinion retrieval techniques tend to perform better on higher versions of TREC Blog datasets and topics (i.e. TREC Blog 2007 and 2008). We think this could be due to the wider coverage of blog documents collected between year 2006 and 2008.

It could also be observed that other than the TREC Blog dataset, some lexicon-based techniques have been experimented on reviews datasets. In terms of precision, Kanayama and Nasukawa (2006) outperforms Ding *et al.* (2008) but with very low recall. However, Ding *et al.* (2008) showed substantial performance for both precision and recall. This makes Ding *et al.* (2008) the likely suitable lexicon-based technique especially on review datasets (Table 3).

4 Probabilistic approach

4.1 Overview of the approach

Probabilistic approaches have been used to measure the degree of opinion in opinionated documents (Elsas *et al.*, 2007; Kim *et al.*, 2009; Gerani *et al.*, 2010; Jiang *et al.*, 2010). In this approach, opinionated documents are ranked in order of relevance by calculating the probability that each document contains opinions using the frequency of opinionated words. A typical generative model in IR is followed by estimating the likelihood of generating a document given a query, that is, $p(d|q)$. For improved opinion retrieval performance, the probabilistic approach is usually combined with lexicon-based approach. One would also prefer the probabilistic approach over the sentiment classification approach if a high number of documents are to be retrieved. In other words, a probabilistic approach is likely to give high recall more than precision.

The probabilistic approach is also known as *probability ranking principle*, formulated by Robertson (1977). According to the early literatures (Hiemstra, 2000; Sparck Jones *et al.*, 2000a, 2000b), the probabilistic approach is dichotomous and independent of every other document within a chosen collection. For example, the relevance of a document to a given query should be either relevant or not relevant, and must be derived within each document without the need to consider other documents, see Lavrenko (2010). For example, in a collection of 100 documents, if we assume 20 out of 100 documents are relevant to the word 'happy', then the probability of relevance for a document taken at random is $1/20 = 0.05$. However, for queries with more than one terms, the binary independence assumption is considered (Hiemstra, 2000). For example, a query with two terms 'happy' and 'family' would have four possible binary states. That is, happy AND family, happy NOT family, family NOT happy, and happy NOT family and NOT (happy or family). Figure 9 shows the structural overview of probabilistic opinion retrieval approach.

The idea of probabilistic opinion retrieval is that words within a given document are ranked in terms of their contributions to searching and returning opinionated information in order of relevance. For this purpose, a two-state process is involved. First, relevant document is identified based on frequency of query words. Second, from the set of relevant documents, words that express opinion and their *frequency of occurrence or proximity to query words* are essential to determining the opinionated relevance of each document. Lexicons of words that express opinion are used to determine the frequency of opinionated words. A precise description for the probabilistic

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Table 3 Performance comparison for *lexicon-based* techniques across different datasets

Author	Approach	Dataset	Performance Metric	Results
Kanayama and Nasukawa (2006)	Lexicon-based	Products and movie reviews	Precision and Recall respectively	96.5% and 1.28%
Ding *et al.* (2008)	Lexicon-based	Amazon and Cnet review	Precision, Recall, and F-score respectively	0.93, 0.92, 0.93
He *et al.* (2008)	Dictionary-based	TREC Blog 2006	MAP (2 different experiment runs)	(i) 0.3749 (ii) 0.3671
Tan *et al.* (2008)	Lexicon-based and Machine Learning	Computer, education, house reviews	Classifier accuracy (three different experiment runs)	(i) 0.9186 (ii) 0.9210 (iii) 0.8230
Zhang and Ye (2008)	Lexicon-based	TREC Blog 2006 and 2007	MAP, R-Prec, and P@10 Respectively (two different experiment runs)	(i) 0.2257, 0.3038, 0.512 (ii) 0.3371, 0.3896, 0.606
Santos *et al.* (2009)	Dictionary-based	TREC Blog 2006	MAP, R-Precision, bPref, and P@10 respectively	0.2990, 0.3432, 0.3644, and 0.6060
Na *et al.* (2009)	Lexicon-based	TREC Blog 2006 and 2007	MAP (two different experiment runs)	(i) 0.3159, 0.3471 (ii) 0.4399, 0.4248
Yang *et al.* (2007)	Lexicon-based	Yahoo! Kimo Blog dataset	Precision and recall, respectively	70.15% and 63.42%
Vechtomova (2010)	Lexicon-based	TREC 2007 and 2008	MAP, P@10, and R-Prec	0.4229, 0.6840, and 0.4601

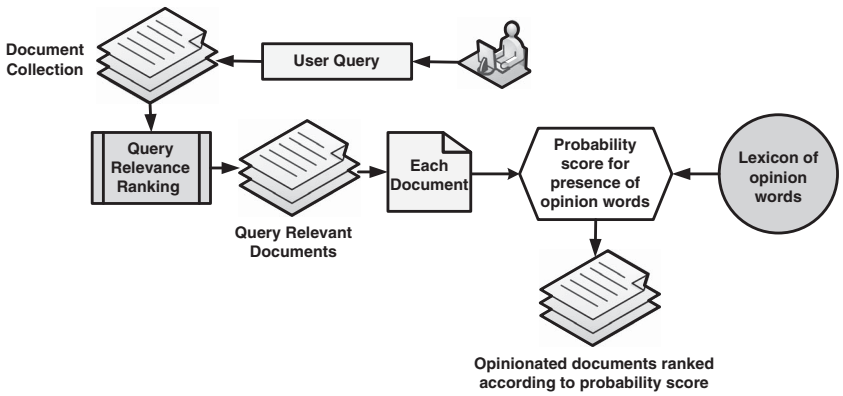


Figure 9 Structural overview of probabilistic opinion retrieval approach

approach to opinion retrieval is to assume user information is represented as query Q, which is a set of query term t. Ranking of relevant document d in document collection D is largely based on the frequency of $t_1 \ldots t_n$ in d. The degree of opinion contained in d is largely based on the frequency of words $w_1 \ldots w_n$ that express opinion as contained in the lexicon.

4.2 Example work (probabilistic opinion retrieval)

Gerani *et al.*, 2010

In this work, a probabilistic model for assigning relevant opinion scores to blog documents was proposed. The work used general opinion lexicon and proposed the use of proximity information to capture opinion term relatedness. The proximity information forms opinion density that signifies the probability of opinion about the query term. The work combined probabilistic and lexicon-based approaches and requires that opinion weights be determined based on a list of generated lexicon. The steps involved in their algorithm can be described as follows:

1. Determine proximity opinion score.
   a. Use opinion lexicon to identify opinionated words.
   b. Calculate the opinion density function at each position in the document by accumulating the opinion density from different opinion terms at a particular position.
   c. The accumulated density is the probability of the opinion to be expressed on a term at a particular position.
   d. The probability that a term at position j is about the query term at position i is computed as:

$$p(j|i,d) = \frac{k(j,i)}{\sum_{j'=1}^{|d|} k(j',i)} \tag{3}$$

   e. Where $k(j, i)$ denotes the kernel function that gives the weight of opinion from position j to position i.
   f. The probability that the document contains opinion at position i is computed as:

$$p(o|i,d) = \sum_{j=1}^{|d|} p(o|t_j)p(j|i,d) \tag{4}$$

2. Calculate the probability that the document contains opinion about the query $p(o|d,q)$.
   a. $p(o|d,q)$ is computed as:

$$p(o|d,q) = \sum_{i=1}^{|d|} p(o,i|d,q)p(i|d,q) = \sum_{i=1}^{|d|} p(o|i,d,q)p(i|d,q) \tag{5}$$

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

b. $p(i|d,q)$ is computed as:

$$p(i|d,q) = \begin{cases} \frac{1}{|pos(q)|} & if\ t_i \in q \\ 0 & otherwise \end{cases} \quad (6)$$

3. Refine the proximity model by performing smoothing.
   a. Perform smoothing with non-proximity-based estimation as follows:

$$p(o|q,d) = (1-\lambda)p(o|q,d) + \lambda p(o|d) \quad (7)$$

   b. The smoothing process ensures proximity is determined at different ranges.

Discussion

The above work showed significant performance on TREC Blog 2006 dataset with about 2.5% improvement in terms of MAP over the best run in TREC Blog 2008 opinion finding task. An interesting research outcome in this work is that proximity information of opinionated words to query words does contribute greatly to determining the degree of opinion in a particular document. However, it is important to normalize the topic-relevance score before applying it to the proximity-based opinion retrieval model. Such normalization strategy would also depend on the selected technique for the topic-relevance retrieval (e.g. BM25).

4.3 Research issues in probabilistic approach

A common problem about the probabilistic approach is the parameter tuning for the smoothing processes. One needs to compare between different parameter values before deciding on the best parameter that gives best performance. Normalization is often another problem in probabilistic approaches. Care must be taken if existing relevant scores are to be transformed to their respective probability estimates. An appropriate normalization technique would depend on the selected relevant scores to be used as baselines. While probabilistic approaches would always be an essential retrieval strategy for opinion retrieval, we think it is more likely to be effective where there are high chances of frequency of observations as applicable to common probabilistic approaches such as *BM25* and *Divergence from Randomness* probabilistic models (Amati & Rijsbergen, 2002; Blei *et al*., 2003; Pickens & MacFarlane, 2006; Ounis *et al*., 2008; Macdonald *et al*., 2009; Robertson & Zaragoza, 2009; Xu *et al*., 2009; Vechtomova, 2010) (Table 4).

4.4 Performance comparison for probabilistic techniques across different datasets

The TREC Blog 2006 dataset is commonly used by most probabilistic approaches. It is important to note that there are variations as to the probabilistic approaches used by individual works. What is more important is that the techniques follow the same fundamental *probability ranking principle* as we had discussed earlier. From the comparison table above, it was observed that most techniques do not have significant performance in terms of MAP. Kim *et al*. (2009) gave the most significant performance on TREC Blog 2006 with MAP value of 0.4242. We think the introduction of semantic and local context proximities as features could be responsible for the improvement over other techniques. Hannah *et al*. (2007) also gave similar performance with MAP value of 0.4160 with little improvement on P@10 value compared to the P@10 given by Kim *et al*. (2009). This improvement is reported to have been influenced by the introduction of statistical approach such as divergence from randomness (DFR) weighting model that uses weighted dictionary for opinion retrieval tasks (McCreadie *et al*., 2009). Gerani *et al*. (2010) also showed substantial performance on TREC Blog 2006 dataset with MAP value of 0.4292 by using the best proximity method to calculate opinion proximity density at different positions in a document. The performance is still encouraging considering the fact that it is really challenging to significantly outperform opinion retrieval baselines on TREC Blog datasets, see Macdonald *et al*. (2010).

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Table 4 Performance comparison for *probabilistic* techniques across different datasets

Author	Approach	Dataset	Performance Metric	Results
Yang *et al.* (2007)	Probabilistic fusion results	TREC Blog 2006	MAP, MRP and P@10, respectively	0.2052, 0.2881 and 0.512
Mishne (2006)	Topic-relevance	TREC Blog 2006	MAP, R-Prec and bpref, respectively	0.1795, 0.2771, 0.2625
Eguchi and Lavrenko (2006)	Probabilistic language-model	MPQA Opinion Corpus	Bpref	0.2278, 0.2441
Hannah *et al.* (2007)	Divergence From Randomness (DFR)	TREC Blog 2006	MAP, P@10 (two different experiment runs)	(i) 0.4160, 0.7200 (ii) 0.3264, 0.5520
Jiang *et al.* (2010)	Relevance-model	TREC Blog 2008	MAP, R-prec, and P@10, Respectively	0.0853, 0.1317, 0.1231
Kim *et al.* (2009)	Term-weight	TREC Blog 2006	MAP, R-Precision, and P@10, respectively	0.4242, 0.4579, 0.6640
Gerani *et al.* (2010)	Probabilistic model	TREC Blog 2006	MAP, R-Prec, bPref, and P@10, respectively	0.4292, 0.4578, 0.4485 and 0.7140
Huang and Croft (2009)	Relevance model	TREC Blog 2006 and COAE08	MAP, R-Prec, bPref and P@10 (Blog06 & COAE08, respectively)	(i) 0.3147, 0.3546, 0.3418, 0.5670 (ii) 0.3697, 0.4311, 0.4069, 0.7900
Lee *et al.* (2010)	Language model and Query likelihood	TREC Blog 2008 and New York Times Annotated Corpus	MAP, P@5, and P@10, respectively	0.1957, 0.3673 and 0.3364

5 Other emerging opinion retrieval approaches

5.1 Overview of the approach

Other than the existing opinion retrieval approaches described in the previous sections, there are other emerging approaches for retrieving opinions. These approaches are designed to overcome the shortcomings of the existing main approaches since study has shown that it is very difficult to significantly outperform the baselines by using the existing approaches that were discussed earlier (Ounis *et al.*, 2008). More importantly, it is difficult for us to show a component diagram that shows the overview of these approaches as they vary depending on the retrieval task and domain of retrieval. These emerging approaches have either proposed a completely new technique or combine two or more existing techniques. We will discuss an example state-of the-art work and then highlight the performance and limitations.

5.2 Example work (other opinion retrieval approaches)

Du and Tan (2009)

In Du and Tan (2009), an information theoretic approach to 'finer-grained' opinion retrieval was proposed. Rather than traditional template extraction method, the technique used an iterative based heuristics to identify hidden sentiments that are not explicitly clear or observable within each sentence in the review document. In addition, the technique allows document features and opinion targets to converge simultaneously and iteratively. Co-occurrence of terms and similar contextual information is considered for this purpose. The steps involved in their algorithm are as follows:

1. Identify hidden sentiment association.
   a. Detect the sentiment association between set of product features F and set of opinion words O with polarity labels.
   b. Build a weighted bipartite graph from F and O denoted by $G = F, O, R$, where $R = [r_{ij}]$ denotes the $m * n$ link weight matrix, which contains all the pair-wise weights between set F and O.
   c. Calculate r_{ij} as the frequency of co-occurrence of f_i and O_j at clause level.
   d. Take F and O as two random variables, and then find the compression C between F and O such that the mutual information between C and O is sufficiently large.
   e. To find the compression C, use iterative reinforcement approach that employs improved information bottleneck algorithm.
2. Use improved information bottleneck algorithm.
   a. Unlike the traditional information bottleneck approach, the divergence between two data objects is known by using co-occurrence information between the two variables F and O.
   b. Do a linear combination of the co-occurrence information and the semantic information as the final distance between two data objects F and O.

Discussion

The strength of this work lies in the interpolation of semantic information with the co-occurrence information between domain specific features and opinionated words. We think the work could be very useful for context-dependent opinion retrieval whereby the content of the opinionated documents retrieved has the same semantic context with the given query topic. However, the precision attained by the work is 78.90%, which gives more room for further improvements. Since this work was evaluated on review dataset, it is difficult to assume that the heuristic created can perform efficiently on domain independent opinionated documents such as blogs.

5.3 Research issues in other emerging opinion retrieval approaches

Emerging opinion retrieval techniques still perform lower than the state-of-art techniques that use the existing approaches that have been discussed in the previous sections. We think this is because

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Table 5 Performance comparison for *emerging* techniques across different datasets

Author	Approach	Dataset	Performance metric	Results
Ernsting *et al.* (2007)	Time-based and frequency-based	TREC Blog 2006	MAP, P@10, and P@30, respectively	0.1605, 0.3111, and 0.2170
Bermingham and Smeaton (2009)	Fusion of approaches	TREC Blog 2006	MAP, R-Prec, and P@10, respectively	0.1644, 0.2074, 0.2041
McCreadie *et al.* (2009)	Divergence from randomness and voting model	TREC Blog 2008	NDCG@10, P@10	0.518, 0.168
Du and Tan (2009b)	Iterative reinforcement approach	Hotel reviews	Precision	78.90%
Thet *et al.* (2009)	Clause-level linguistic	Imdb movie review	Precision, recall and *F*-score Respectively	87%, 100%, and 93%
Xu *et al.* (2009)	Facet identification	TREC Blog 2008	MAP and R-Prec	0.2399 and 0.2863
He *et al.* (2008)	Fitting score distribution method	TREC Blog 2006	MAP (two different experiment runs)	(i) 0.3088 (ii) 0.3096
Huang and Croft (2009)	Relevance model	TREC Blog 2006 and COAE08	MAP, R-Prec, bPref and P@10 (Blog06 & COAE08, respectively)	(i) 0.3147, 0.3546, 0.3418, 0.5670 (ii) 0.3697, 0.4311, 0.4069, 0.7900
Lee *et al.* (2010)	Sentiment-relevance flow	TREC Blog 2006	P@1, P@3 and P@5 (two different experiment runs)	(i) 0.5800, 0.5800, 0.5573 (ii) 0.6333, 0.5689, 0.5520

of the difficult nature of opinion retrieval task. However, emerging approaches that tend to use NLP techniques using semantic and syntactic information seems promising. A common challenge across many opinion retrieval techniques is their ability to detect implicit opinions or sentiments, which could be difficult without effective NLP technique (Table 5).

5.4 Performance comparison for emerging techniques across different dataset

The performance of these techniques varies on different datasets. However, the emerging techniques experimented on TREC Blog datasets can be easily compared for performance improvements over the baselines. Based on the comparison table above, He *et al*. (2008) significantly outperforms Bermingham and Smeaton (2009) by almost 60%. However, performance results in terms MAP for all the emerging techniques stated in the table above are significantly low compared with the performance results given by the existing approaches discussed earlier (i.e. text classification, lexicon-based, and probabilistic approaches). As the research on opinion retrieval gets mature, these emerging techniques are more likely to use new and efficient methods that can outperform the various state-of-the-art techniques.

6 Future research directions

From this review, we believe opinion retrieval could be treated as natural language problem. Recently, Krahmer (2010) suggested a linguistic approach toward providing effective solution to some of the research issues identified in this paper. As mentioned earlier, opinionated documents are diverse such that there is need for proper understanding of what each sentence means (Munson & Resnick, 2010). Moreover, opinions are context-dependent, for example, one sentence may have multiple meaning, and it would be difficult to extract opinion from such sentence without effective linguistic approach. Therefore, we think semantic and syntactic features could play important role in opinion retrieval (Pang & Lee, 2008). From the series of literatures discussed in this paper, we could observe that the success of any opinion retrieval technique would depend on high degree of semantic relevance of opinionated documents to the underlying meaning of the given query topic.

An important development is to see emerging opinion retrieval approaches divert from the conventional approaches that use frequency of query words or frequency of opinion words. The nature of problems that come with new opinion retrieval tasks require multi-faceted approach to detect opinions at different inclinations (Liu, 2010). Inclination describes what kind of opinionated documents are to be retrieved (Macdonald *et al*., 2010). For example, some people may prefer to retrieve opinionated documents that discuss about personal life experiences rather than quality of a product. In this case, the inclination of interest is *personal* rather than *commercial* as in product reviews. To solve the multi-faceted opinion retrieval problem, it could be less effective to rely on approaches that use frequency of query words or frequency of opinion words alone. We think multi-sentence semantic dependencies approach should be used to partition opinions that occur at different inclinations.

Detecting sentences that contain opinion is still a key challenge in opinion retrieval. This affects selection of training features for the text classification approach. We think it is worth studying an automatic and effective approach to detect sentences that contains opinion regardless of the domain. Many attempts have been made to use the lexicon-based approach to solve this problem (Ding & Liu, 2007; Tan *et al*., 2008; Na *et al*., 2009). However, it is not yet clear how the lexicon-based approach can be domain-independent. For example, the domains from which the words in the lexicon are selected are limited in most cases to certain domains of interest. Certain lexicons created by set of human judges, for example, LIWC (Tausczik & Pennebaker, 2010), have also been limited to few sentiment categories. It is obvious that with the fast growing rate of activities on the World Wide Web, documents that contain opinion are getting more diverse and dynamic. This creates possibility that lexicon created for certain domains could be outdated within a very

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

short period. We think it could be very useful to explore how lexical relationships between words in a given sentence can be used to detect opinions.

Finally, we are aware that proximity between query and opinion terms has been explored (Santos *et al.*, 2009; Gerani *et al.*, 2010). However, many words in proximity might be independent of each other. We think it is better to identify words proximity in the semantic context at which opinionated information has been requested and not words proximity by ordinary presence of query or opinion words.

7 Summary and conclusions

Current review papers on opinion retrieval discuss theoretical foundations and possible applications of opinion retrieval systems. In this paper, we have discussed the major opinion retrieval approaches and their performance on different datasets. We have also highlighted some limitations and questions that demand answers in future research efforts. Some of the approaches discussed in this paper have contributed immensely to solving opinion retrieval research problems. However, there are still many challenges as most emerging techniques still perform significantly lower than the state-of-the-art techniques.

Based on the performance summary tables presented in this paper, majority of the opinion retrieval techniques are evaluated on TREC Blog datasets, which aid the performance comparison. However, regardless of the approach used, we could observe most opinion retrieval techniques in the research community are still largely based on frequency or presence of individual query or opinion terms within documents. The use of frequency or presence of query terms negatively affect context-dependent opinion retrieval (Krahmer, 2010). We also note that state-of-the-art techniques perform similarly except for few instances. This shows research on opinion retrieval is still very open to developing effective opinion retrieval techniques that can perform with better precision.

From our review work, it is still difficult to conclude that a particular approach is the best. Even though there may be some limitations in some of the works, we have seen instances where the *text classification approach* showed good performance (Turney & Littman, 2002; Jia *et al.*, 2008; Zagibalov & Carroll, 2008; O'Hare *et al.*, 2009; Siersdorfer *et al.*, 2010); *lexicon-based approach* (Kanayama & Nasukawa, 2006; Yang *et al.*, 2007; Ding *et al.*, 2008; He *et al.*, 2008; Tan *et al.*, 2008; Zhang & Ye, 2008; Du & Tan, 2009; Santos *et al.*, 2009; Vechtomova, 2010); *probabilistic approach* (Eguchi & Lavrenko, 2006; Bermingham & Smeaton, 2009; Elsas & Dumais, 2010; Gerani *et al.*, 2010); and *other emerging approaches* (Du & Tan, 2009b; Huang & Croft, 2009; Santos *et al.*, 2009; Thet *et al.*, 2009; Xu *et al.*, 2009; Yu *et al.*, 2010).

We believe that regardless of the approach used for opinion retrieval, effective opinion retrieval with sufficient semantic relevance between the opinionated documents and the query topic is

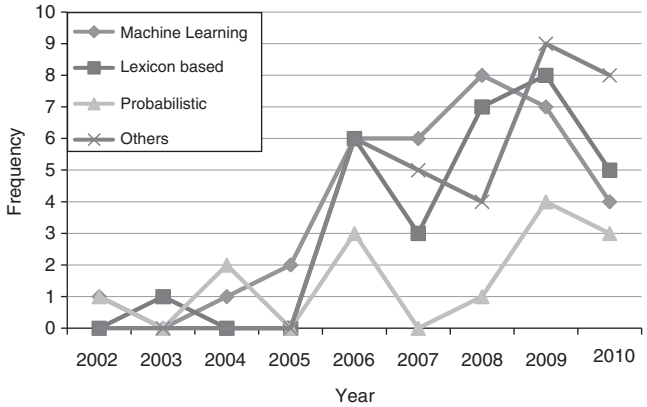


Figure 10 Frequency of opinion retrieval approaches over 90 research works published in research communities between year 2002 and 2010

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

paramount. More importantly, performance evaluation can be based on appropriate IR performance metrics as discussed earlier in this paper. Experimentation should also be conducted on popular and standard datasets such that the performance results can be compared with other techniques (Ounis *et al*., 2006; Macdonald *et al*., 2007; Ounis *et al*., 2008; Macdonald *et al*., 2009; Macdonald *et al*., 2010). In conclusion, Figure 10 shows how the discussed opinion retrieval approaches have trended over the past 8 years.

The above figure shows trends of different approaches used in opinion retrieval over a period of 8 years as collected from ACM and Google Scholar Digital Libraries. It could be observed that machine learning or text classification approach moved at a very high pace since the inception of opinion retrieval in the early 2000 (Pang & Lee, 2008; Liu, 2010). However, the use of the approach dropped around year 2008 as other approaches such as lexicon-based and probabilistic approaches began to receive attention of the research community. Since 2006, the probabilistic approach has received considerable attention in opinion retrieval research community due to its scalability. Currently, other independent approaches (i.e. approaches other than machine learning or text classification, lexicon based, and probabilistic) are now getting more attention within the opinion retrieval research community. However, as mentioned earlier, none of the independent approaches has significantly outperformed the state-of-the-art baselines for opinion retrieval.

Acknowledgement

This work is supported by MONASH University Sunway Campus Higher Degree by Research (HDR) Scholarship.

References

Abbasi, A., Chen, H. & Salem, A. 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems* **26**(3), 1–34.

Agarwal, N. & Liu, H. 2008. Blogosphere: research issues, tools, and applications. *SIGKDD Explorations Newsletter* **10**(1), 18–31.

Amati, G. & Rijsbergen, C. J. V. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems* **20**(4), 357–389.

Baharudin, B., Lee, L. H. & Khan, K. 2010. A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology* **1**(1), 4–20.

Bermingham, A. & Smeaton, A. F. 2009. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 784–785.

Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning and Research* **3**, 993–1022.

Bollen, J., Pepe, A. & Mao, H. 2010. Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina.

Choi, Y., Kim, Y. & Myaeng, S-H. 2009. Domain-specific sentiment analysis using contextual feature generation. In *Proceeding of the 1st international CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*. ACM, 37–44.

Chung, C. K., Jones, C., Liu, A. & Pennebaker, J. W. 2008. Predicting success and failure in weight loss blogs through natural language use. *Association for the Advancement of Artificial Intelligence*, 180–181.

Collins, M. J. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 184–191.

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*. ACM, 811–812.

Ding, X., Liu, B. & Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining*. ACM, 231–240.

Du, W. & Tan, S. 2009a. Building domain-oriented sentiment lexicon by improved information bottleneck. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. ACM, 1749–1752.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Du, W. & Tan, S. 2009b. An iterative reinforcement approach for fine-grained opinion mining. In *Proceedings of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, 486–493.

Eguchi, K. & Lavrenko, V. 2006. Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 345–354.

Elsas, J., Aguello, J., Callan, J. & Carbonell, J. 2007. Retrieval and Feedback Models for Blog Distillation. In *Proceedings of the Text Retrieval Conference (TREC)*, National Institute of Standards and Technology, MD, USA.

Elsas, J. L. & Dumais, S. T. 2010. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of the Third ACM International Conference on Web search and Data Mining*. ACM, 1–10.

Esuli, A. 2008. Automatic generation of lexical resources for opinion mining: models, algorithms and applications. *SIGIR Forum* **42**(2), 105–106.

Esuli, A. & Sebastiani, F. 2006. SENTIWORDNET: a publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, Italy.

Fei, Z., Liu, J. & Wu, G. 2004. Sentiment classification using phrase patterns. In *Proceedings of the 4th IEEE International Conference on Computer Information Technology*, 1147–1152.

Fellbaum, C. 2010. WordNet. Theory and Applications of Ontology. In *Computer Applications*, Poli, R., Healy, M. & Kameas, A. (eds). Springer, 231–243.

Fortuna, B., Rodrigues, E. M. & Milic-Frayling, N. 2007. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. ACM, 877–880.

Gamon, M. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*. ACL, 841.

Ganapathibhotla, M. & Liu, B. 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Association for Computational Linguistics, vol. 1, 241–248.

Gerani, S., Carman, M. J. & Crestani, F. 2009. Investigating learning approaches for Blog Post opinion retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Springer-Verlag, 313–324.

Gerani, S., Carman, M. J. & Crestani, F. 2010. Proximity-based opinion retrieval. In *Proceedings of the 33rd International Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*. ACM, 403–410.

Hannah, D., Macdonald, C., Peng, J., He, B. & Ounis, I. 2007. *University of Glasgow at Trec 2007: Experiments in Blog and Enterprise Tracks with Terrier*. TREC.

Hatzivassiloglou, V. & McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 174–181.

He, B., Macdonald, C., He, J. & Ounis, I. 2008. An effective statistical approach to blog post opinion retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*. ACM, 1063–1072.

Hiemstra, D. 2000. Using language models for information retrieval. PhD Thesis, Centre for Telematics and Information Technology.

Hu, M. & Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 168–177.

Hu, M. & Liu, B. 2004. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*. AAAI Press, 755–760.

Huang, J. & Efthimiadis, E. N. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*. ACM, 77–86.

Huang, X. & Croft, W. B. 2009. A unified relevance model for opinion retrieval. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*. ACM, 947–956.

Jia, L., Yu, C. & Meng, W. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*. ACM, 1827–1830.

Jiang, P., Zhang, C., Yang, Q. & Niu, Z. 2010. Blog opinion retrieval based on topic-opinion mixture model. In *Advances in Knowledge Discovery and Data Mining*, Zaki, M., Yu, J., Ravindran, B. & Pudi, V. (eds). Springer, vol. 6119, 249–260.

Jin, W., Ho, H. H. & Srihari, R. K. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1195–1204.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Jindal, N. & Liu, B. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 244–251.

Jia, L., Yu, C. & Zhang, W. 2008. UIC at TREC 2008 Blog Track. In *Text REtrieval Conference* 2008.

Jones, R., Rey, B., Madani, O. & Greiner, W. 2006. Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland, 387–396.

Joshi, H., Bayrak, C. & Xu, X. 2006. UALR at TREC: Blog track. In *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Kanayama, H. & Nasukawa, T. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 355–363.

Kim, J., Li, J.-J. & Lee, J.-H. 2009. Discovering the discriminative views: measuring term weights for sentiment analysis. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, vol. 1, 253–261.

Kobayakawa, T. S., Kumano, T., Tanaka, H., Okasaki, N., Kim, J.-D. & Tsujii, J. 2009. Opinion classification with tree kernel SVM using linguistic modality analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, Hong Kong, China, 1791–1794.

Koppel, M. & Shtrimberg, I. 2006. Good news or bad news? Let the market decide. Computing attitude and affect in text. In *Theory and Applications*, Shanahan, J., Qu, Y. & Wiebe, J. (eds). Springer, vol. 20, 297–301.

Krahmer, E. 2010. What computational linguists can learn from psychologists (and vice versa). *Association for Computational Linguistics* **36**(2), 285–294.

Ku, L.-W., Liang, Y.-T. & Chen, H.-H. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *American Association for Artificial Intelligence*.

Lavrenko, V. P. 2010. Introduction to probabilistic models in IR. In *Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, 905–905.

Lee, S.-W., Lee, J.-T., Song, Y.-I. & Rim, H.-C. 2010. High precision opinion retrieval using sentiment-relevance flows. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, 817–818.

Lee, Y., Jung, H.-y., Song, W. & Lee, G.-H. 2010. Mining the blogosphere for top news stories identification. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland, 395–402.

Lee, Y., Na, S.-H., Kim, J., Nam, S.-H., Jung, H.-y. & Lee, J.-H. 2008. KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval. In *TREC* 2008.

Lin, C. & He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM CIKM*. ACM, 375–384.

Lin, D. 1998. Dependency-Based Evaluation of Minipar. `http://www.cfilt.iitb.ac.in/archives/minipar_evaluation.pdf`

Lin, W.-H., Wilson, T., Wiebe, J. & Hauptmann, A. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 109–116.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B. & Lauw, H. W. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of CIKM ACM*, Toronto, Ontario, Canada.

Liu, B. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing* **2**, 568.

Liu, B. 2010. Sentiment analysis: a multi-faceted problem. *IEEE Intelligent Systems* **25**(3), 76–80.

Liu, B., Hu, M. & Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web*, Chiba, Japan, 342–351.

Liu, S., Liu, F., Yu, C. & Meng, Y. 2004. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, 266–272.

Lloret, E., Saggion, H. & Palomar, M. 2010. Experiments on summary-based opinion classification. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. ACL, 107–115.

Lu, B., Ott, M., Cardie, C. & Tsou, B. K. 2011. Multi-aspect sentiment analysis with topic models. *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, Vancouver, Canada, 81–88.

Macdonald, C., Ounis, I. & Soboroff, I. 2007. Overview of the TREC2007 Blog Track. *TREC* 2007.

Macdonald, C., Ounis, I. & Soboroff, I. 2009. Overview of the TREC2009 Blog Track. *TREC* 2009.

Macdonald, C., Santos, R. L. T., Ounis, I. & Soboroff, I. 2010. Blog track research at TREC. *SIGIR Forum* **44**(1), 58–75.

McCreadie, R., Macdonald, C., Ounis, I., Peng, J. & Santos, R. L. 2009. *University of Glasgow at Trec 2009: Experiments with Terrier*. Glasgow University, UK.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Manning, C. D., Raghavan, P. & Schtze, H. 2009. *An Introduction to Information Retrieval*. Cambridge University Press.

Mishne, G. 2006. Multiple ranking strategies for opinion retrieval in blogs. In *Online Proceedings of TREC*.

Munson, S. A. & Resnick, P. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the 28th international Conference on Human Factors in Computing Systems*. ACM, 1457–1466.

Na, S.-H., Lee, Y., Nam, S.-H. & Lee, J.-H. 2009. Improving opinion retrieval based on query-specific sentiment lexicon. In *Advances in Information Retrieval*, Boughanem, M., Berrut, C., Mothe, J. & Soule-Dupuy, C. (eds). Springer, vol. 5478, 734–738.

Narayanan, R., Liu, B. & Choudhary, A. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 180–189.

Nasukawa, T. & Yi, J. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*. ACM, 70–77.

Nguyen, T., Phung, D., Adams, B., Tran, T. & Venkatesh, S. 2010. Classification and pattern discovery of mood in Weblogs. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 283–290.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C. & Smeaton, A. F. 2009. Topic-dependent sentiment analysis of financial blogs. In *Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*. ACM, 9–16.

Ounis, I., Macdonald, C. & Soboroff, I. 2008. Overview of the TREC2008 Blog Track. *TREC* 2008.

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. & Soboroff, I. 2006. Overview of the TREC-2006 Blog Track. *TREC* 2006.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q. & Chen, Z. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 751–760.

Pang, B. & Lee, L. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 271.

Pang, B. & Lee, L. 2008. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval* **2**(1–2), 1–135.

Pang, B., Lee, L. & Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 79–86.

Paul, M. & Girju, R. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, Georgia, USA, 545–550.

Pickens, J. & MacFarlane, A. 2006. Term context models for information retrieval. In *Proceedings of the 15th ACM international Conference on Information and KnowledgeMmanagement*. ACM, 559–566.

Robertson, S. E. 1997. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., 281–286.

Robertson, S. & Zaragoza, H. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* **3**(4), 333–389.

Robertson, S. E. & Walker, S. 1999. Okapi/keenbow at trec-8. In *Proceedings of TREC*, volume 8.

Sarmento, L., Carvalho, P., Silva, M. J. & de Oliveira, E. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, Hong Kong, 29–36.

Santos, R. L. T., Ben, H., Macdonald, C. & Ounis, I. 2009. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Springer-Verlag, 325–336.

Santos, R. L. T., Macdonald, C. & Ounis, I. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 881–890.

Siersdorfer, S., Chelaru, S. & Pedro, J.-S. 2010. How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. *International World Wide Web Conference*, Raleigh, North Carolina, USA, 891–900.

Sparck Jones, K., Walker, S. & Robertson, S. E. 2000a. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management* **36**(6), 779–808.

Sparck Jones, K., Walker, S. & Robertson, S. E. 2000b. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management* **36**(6), 809–840.

Stone, P. J., Dunphy, D. C. & Smith, M. S. 1966. *The general inquirer: a computer approach to content analysis*. MIT Press.

Strapparava, C. & Mihalcea, R. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*. ACM, 1556–1560.

Sun, A., Lim, E.-P. & Ng, W.-K. 2002. Web classification using support vector machine. In *Proceedings of the 4th International Workshop on Web Information and Data Management*. ACM, 96–99.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press

Taboada, M., Brooke, J. & Stede, M. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the SIGDIAL 2009 Conference*: *The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 62–70.

Tan, B. & Peng, F. 2008. Unsupervised query segmentation using generative language models and wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 347–356.

Tan, S., Wang, Y. & Cheng, X. 2008. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 743–744.

Tausczik, Y. R. & Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* **29**(1), 24–54.

Thet, T. T., Na, J.-C., Khoo, C. S. G. & Shakthikumar, S. 2009. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*. ACM, 81–84.

Torsten, Z., Christof, M. & Iryna, G. 2009. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary, *LREC*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. 2010. Predicting elections with twitter: what 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC.

Turney, P. & Littman, M. L. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report, EGB-1094, National Research Council Canada.

Vechtomova, O. 2010. Facet-based opinion retrieval from blogs. *Information Processing & Management* **46**(1), 71–88.

Whitelaw, C., Garg, N. & Argamon, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany. ACM, 625–631.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. 2005. OpinionFinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Association for Computational Linguistics, 34–35.

Xu, X., Liu, Y., Xu, H., Yu, X., Song, L., Guan, F., Peng, Z. & Cheng, X. 2009. ICTNET at Blog Track TREC 2009. In *TREC* 2009.

Yang, C., Lin, K. H.-Y. & Chen, H.-H. 2007. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, 133–136.

Yu, S., Cai, D., Wen, J.-R. & Ma, W.-Y. 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th International Conference on World Wide Web*. ACM, 11–18.

Yu, X., Liu, Y., Huang, X. & An, A. 2010. A quality-aware model for sales prediction using reviews. In *Proceedings of the 19th International Conference on World wide web*. ACM, 1217–1218.

Zafarani, R., Cole, W. & Huan, L. 2010. Sentiment propagation in social networks: a case study in LiveJournal. In *Advances in Social Computing*, Chai, S.-K., Salerno, J. & Mabry, P. (eds). Springer, vol. 6007, 413–420.

Zagibalov, T. & Carroll, J. 2008. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Association for Computational Linguistics, 1073–1080.

Zettlemoyer, L. S. & Collins, M. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, vol. 2, 976–984.

Zhang, C., Xue, G.-R., Yu, Y. & Zha, H. 2009. Web-scale classification with naive bayes. In *Proceedings of the 18th international Conference on World Wide Web*. ACM, 1083–1084.

Zhang, E. & Zhang, Y. 2006. UCSC on TREC 2006 Blog Opinion Mining. In *TREC* 2006.

Zhang, M. & Ye, X. 2008. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 411–418.

Zhang, W., Liu, S., Yu, C., Sun, C., Liu, F. & Meng, W. 2007a. Recognition and classification of noun phrases in queries for effective retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. ACM, 711–720.

Zhang, W. & Yu, C. 2006. UIC at TREC 2006 Blog Track. In *TREC*, 2006.

Zhang, W., Yu, C. & Meng, W. 2007b. Opinion retrieval from blogs. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. ACM, 831–840.

https://doi.org/10.1017/S0269888913000167 Published online by Cambridge University Press