Parlar *et al. Hum. Cent. Comput. Inf. Sci.* (2018) 8:10
https://doi.org/10.1186/s13673-018-0135-8

Human-centric Computing and Information Sciences

RESEARCH Open Access

QER: a new feature selection method for sentiment analysis

Tuba Parlar[1*], Selma Ayşe Özel[2] and Fei Song[3]

*Correspondence:
tparlar@mku.edu.tr
[1] Department of Mathematics, Mustafa Kemal University, Antakya, Hatay, Turkey
Full list of author information is available at the end of the article

Abstract

Sentiment analysis is about the classification of sentiments expressed in review documents. In order to improve the classification accuracy, feature selection methods are often used to rank features so that non-informative and noisy features with low ranks can be removed. In this study, we propose a new feature selection method, called query expansion ranking, which is based on query expansion term weighting methods from the field of information retrieval. We compare our proposed method with other widely used feature selection methods, including Chi square, information gain, document frequency difference, and optimal orthogonal centroid, using four classifiers: naïve Bayes multinomial, support vector machines, maximum entropy modelling, and decision trees. We test them on movie and multiple kinds of product reviews for both Turkish and English languages so that we can show their performances for different domains, languages, and classifiers. We observe that our proposed method achieves consistently better performance than other feature selection methods, and query expansion ranking, Chi square, information gain, document frequency difference methods tend to produce better results for both the English and Turkish reviews when tested using naïve Bayes multinomial classifier.

Keywords: Sentiment analysis, Feature selection, Machine learning, Text classification

Introduction

"What other people think" has always been an important piece of information for most of us during the decision making process [1]. The Internet and social media provide a major source of information about people's opinions. Due to the rapidly-growing number of online documents, it becomes both time-consuming and hard to obtain and analyze the desired opinionated information. Turkey is among the top 20 countries with the highest numbers of Internet users according to the Internet World Stats.[1] The exploding growth in the Internet users is one of the main reasons that sentiment analysis for different languages and domains becomes an actively-studied area for many researchers [2–6].

Sentiment analysis (SA) is a natural language processing task that classifies the sentiments expressed in review documents as "positive" or "negative". In general, SA is considered as a two-class classification problem. However, some researchers use "neutral" as

[1] http://www.internetworldstats.com/.

the third class label. There are a number of studies about sentiment analysis that use different approaches for data preprocessing, feature selection, and sentiment classification [1, 3, 4, 6–10]. The statistical methods such as Chi square (CHI2) and information gain (IG) are used to eliminate unnecessary or irrelevant features so that the classification performance can be improved [11]. Supervised learning methods including naïve Bayes (NB), support vector machines (SVM), decision trees (DT), and maximum entropy modelling (MEM) are used to classify the sentiments of the reviews.

Although SA can be considered as a text classification task, it has some differences from the traditional topic-based text classification. For example, instead of saying: "This camera is great. It takes great pictures. The LCD screen is great. I love this camera" in a review document, people are more likely to write: "This camera is great. It takes breathtaking pictures. The LCD screen is bright and clear. I love this camera." [8]. As can be seen, sentiment-expressing words like "great" are not so frequent within a particular review, but can be more frequent across different reviews, and a good feature selection method for SA should take this observation into account.

In this paper, we propose a new feature selection method, called query expansion ranking (QER) which is especially developed for reducing dimensionality of feature space of SA problems. The aim of this study is to show that our proposed method is effective for SA from review texts written in different languages (e.g., Turkish, English) and domains (e.g., movie reviews, book reviews, kitchen appliances reviews, etc.). QER is based on query expansion term weighting methods used to improve the search performance of information retrieval systems [12, 13] and to evaluate its effectiveness as a feature selector in SA, we compare it with other common feature selection methods, including CHI2, IG, document frequency difference (DFD), and optimal orthogonal centroid (OCFS), along with four text classifiers: naïve Bayes multinomial (NBM), SVM, DT, and MEM, over ten different review documents datasets. Our goal is to examine whether these feature selection methods can reduce the feature sizes and improve the classification accuracy of sentiment analysis with respect to different document domains, languages, and classifiers.

The rest of the paper is organized as follows. "Related work" reviews the related work on sentiment analysis. "Methods" presents the methods that we used for our study, including the new feature selection method we proposed. "Experiments and results" describes the experimental settings, datasets, performance measures, and testing results. Finally, "Conclusion" concludes the paper.

Related work

SA is an important topic in Natural Language Processing and Artificial Intelligence. Also known as opinion mining, SA mines people's opinions, sentiments, evaluations, and emotions about entities such as products, services, organizations, individuals, issues, and events, as well as their related attributes. This kind of analysis has many useful applications. For example, it determines a product's popularity according to the user's reviews. If the overall sentiments are negative, further analysis may be performed to identify which features contribute to the negative ratings so companies can reshape their businesses. Numerous studies have been done for sentiment analysis in different domains, languages, and approaches [3–5, 8–10, 14–17]. Among these studies,

the machine learning approaches are more popular since the models can be automatically trained and improved with the training datasets. Pang et al. [4] apply supervised machine learning methods such as NB and SVM to sentiment classification. NB, SVM, MEM, and DT are some of the commonly used machine learning approaches [4, 7–9, 14]. Feature selection methods are used to rank features so that non-informative features can be removed to improve the classification performance [18]. Some researchers have investigated the effects of feature selection for sentiment analysis [3, 8–10, 19–25]. For example, Yang and Yu [3] examine IG for feature selection and evaluate its performance using NB, SVM, and C4.5 (popular implementation for DT) classifiers. Nicholls et al. [8] compare their proposed DFD feature selection method against other feature selection methods, including CHI2, OCFS [26], and count difference using the MEM classifier. Agarwal et al. [9] investigate minimum redundancy maximum relevancy (mRMR) and IG methods for sentiment classification using NBM and SVM classifiers. The results show that mRMR performs better than IG for feature selection, and NBM performs better than SVM in accuracy and execution time. Abbasi et al. [22] examine a new feature selection method called entropy weighted genetic algorithm (EWGA) and compare the performance of this method using information gain feature selection method. EWGA achieves a relatively high accuracy of 91.7% using SVM classifier. Xia et al. [24] design two types of feature sets: POS based and word relation based. Their word relation based method improves an accuracy of 87.7 and 85.15% on movie and product datasets. Bai [25] proposes a Tabu heuristic search-enhanced Markov blanket model that provides a vocabulary to extract sentiment features. Their method achieves an accuracy of 92.7% for the movie review dataset. Mladenovic et al. [16] propose a feature selection method that is based on mapping of a large number of related features to a few features. Their proposed method improves the classification performance using unigram features with 95% average accuracy. Zheng et al. [27] perform comparative experiments to test their proposed improved document frequency feature selection method. Their method achieves significant improvement in sentiment analysis of Chinese online reviews with an accuracy of 97.3%.

Most of the SA studies listed above focus on the English language. Only few studies have been done on SA for the Turkish language [6, 10, 19, 28–31]. The Turkish language belongs to the Altaic branch of the Ural-Altaic family of languages and is mainly used in the Republic of Turkey. Turkish is an agglutinative language similar to Finnish and Hungarian, where a single word can be translated into a relatively longer sentence in English [32]. For instance, word “karşılaştırmalısın” in Turkish can be expressed as “you must make (something) compare” in English. As Turkish and English have different characteristics, methods developed for SA in English need to be tested for Turkish. Among the few researchers who investigate the effects of feature selection on the SA of Turkish reviews, Boynukalın [29] applies Weighted Log Likelihood Ratio (WLLR) to reduce feature space with NB, Complementary NB, and SVM classifiers for the emotional analysis using the combinations of n-grams where sequences of n words are considered together. It is shown that WLLR helps to improve the accuracy with reduced feature sizes. Akba et al. [19] implement and compare the performance of reduced feature sizes using two feature selection methods: CHI2 and IG with NB and SVM classifiers. They show that feature selection methods improve the classification accuracy.

Our aim is to propose a new feature selection method for the SA of Turkish and English reviews. We presented an initial version of this method in [10] where we employ only product review dataset in Turkish and compare our method with CHI2 and DFD by using only one classifier. We now extend it to more datasets for Turkish, and also investigate the performance of our method in English datasets to show that our method is language independent. We further include more feature selection methods especially developed for SA and compare the performance of our proposed method using NBM, SVM, MEM, and DT classifiers along with statistical analysis to prove that our method is classifier independent.

Methods

Machine learning algorithms

For sentiment classification, we use the Weka [33] data mining tool, which contains the four classifiers we use in our experiments, i.e., NBM, SMO for SVM, J48 for C4.5, and LR for MEM. We choose NBM, SVM, LR, and J48 classification methods due to the following reasons: (i) many researchers use NBM for text classification because it is computationally efficient [9, 10, 14] and performs well for large vocabulary sizes [34]; (ii) SVM tends to perform well for traditional text classification tasks [3, 4, 7, 14, 35]; (iii) LR is known to be equivalent to MEM which is another method used in SA studies [8]; (iv) J48 is a well-known decision tree classifier for many classification problems and is used for SA [3, 30].

Feature selection

Feature Selection methods have been shown to be useful for text classification in general and sentiment analysis in specific [11, 18]. Such methods rank features according to certain measures so that non-informative features can be removed, and at the same time, the most valuable features can be kept in order to improve the classification accuracy and efficiency. In this study, we consider several feature selection methods, including information gain, Chi square, document frequency difference, optimal orthogonal centroid, and our new query expansion ranking (QER) so that we can compare their effectiveness for the sentiment analysis.

Feature sizes are selected in the range from 500 to 3000 with 500 increments, compared with the total feature sizes ranging from 8000 to 18,000 for the Turkish review datasets and from 8000 to 38,000 for English review datasets. In our previous study [10], we observed that feature sizes up to 3000 tend to give good classification performance improvement; therefore we choose these feature sizes in our experiments.

Information gain

Information gain is one of the most common feature selection methods for sentiment analysis [3, 9, 19, 35], which measures the content of information obtained after knowing the value of a feature in a document. The higher the information gain, the more power we have to discriminate between different classes.

The content of information can be calculated by the entropy that captures the uncertainty of a probability distribution for the given classes. Given m number of classes: $C = \{c_1, c_2, \ldots, c_m\}$ the entropy can be given as follows:

$$H(C) = -\sum_{i=1}^{m} P(c_i) \log_2 P(c_i) \tag{1}$$

where $P(c_i)$ is the probability of how many documents in class c_i. If an attribute A has n distinct values: $A = \{a_1, a_2, \ldots, a_n\}$, then the entropy after the attribute A is observed can be defined as follows:

$$H(C|A) = \sum_{j=1}^{n} \left(-P(a_j) \sum_{i=1}^{m} P(c_i|a_j) \log_2 P(c_i|a_j) \right) \tag{2}$$

where $P(a_j)$ is the probability of how many documents contain the attribute value a_j, and $P(c_i|a_j)$ is the probability of how many documents in class c_i that contain the attribute value a_j. Based on the definitions above, the information gain for an attribute is simply the difference between the entropy values before and after the attribute is observed:

$$IG(A) = H(C) - H(C|A) \tag{3}$$

For sentiment analysis, we normally classify the reviews into positive and negative categories, and for each keyword, it either occurs or does not occur in a given document; so the above formulas can be further simplified. Nevertheless, we can cut down the number of features in the same way by choosing the keywords that have high information gain scores.

Chi square (CHI2)

Chi square measures the dependence between a feature and a class. A higher score implies that the related class is more dependent on the given feature. Thus, a feature with a low score is less informative and should be removed [3, 8, 10, 19]. Using the 2-by-2 contingency table for feature f and class c, where A is the number of documents in class c that contains feature f, B is the number of documents in the other class that contains f, C is the number of documents in c that does not contain f, D is the number of documents in the other class that does not contain f, and N is the total number of documents, then the Chi square score can be defined in the following:

$$\chi^2(f, c) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{4}$$

The Chi square statistics can also be computed between a feature and a class in the dataset, which are then combined across all classes to get the scores for each feature as follows:

$$\chi^2(f) = \sum_{i=1}^{m} P(c_i) \chi^2(f, c_i) \tag{5}$$

One problem with the CHI2 method is that it may produce high scores for rare features as long as they are mostly used for one specific class. This is a bit counter-intuitive, since rare features are not frequently used in text and thus do not have a big impact for text

classification. For SA, however, this is not a big issue since many sentiment-expressing features are not frequently used within an individual review.

Document frequency difference

Inspired by the observation that sentiment-expressing words tends to be less frequent within a review, but more frequent across different reviews, Nicholls and Song [8] propose the DFD method that tries to differentiate the features for positive and negative classes, respectively, across a document collection. More specifically, DFD is calculated as follows:

$$Score_f = \frac{|DF_+^f - DF_-^f|}{N} \tag{6}$$

where DF_+^f is the number of documents in the positive class that contain feature f, DF_-^f is the number of documents in the negative class that contain f, and N is the total number of documents in the dataset. Note that all scores are normalized between 0 and 1; so they should be proportional for us to rank the features in a document collection. For example, a non-sentiment word may have similar document frequencies in both positive and negative classes, and will get a low score, but a sentiment word for the positive class may have a bigger difference, resulting in a higher score. One limitation of the DFD method is that it requires an equal or nearly equal number of documents in both classes, which is more or less true for the datasets used in our experiments.

Optimal orthogonal centroid (OCFS)

OCFS method is an optimized form of the orthogonal centroid algorithm [26]. Documents are represented as high dimensional vectors where the weights of each dimension correspond to the importance of the related features, and a centroid is simply the average vector for a set of document vectors. OCFS aims at finding a subset of features that can make the sum of distances between all the class means maximized in the selected subspace. The score of a feature f by OCFS is defined in the following [8]:

$$Score_f = \sum_c \frac{N_c}{N} \left(m_c^f - m^f \right)^2 \tag{7}$$

where N_c is the number of documents in class c, N is the number of documents in the dataset, m_c is the centroid for class c, m is the centroid for the dataset D, and m^f, m_c^f are the values of feature f in centroid m, m_c respectively. The centroids of m and m_c are calculated as follows:

$$m_c = \frac{\sum_{x_i \in c} x_i}{N_c} \tag{8}$$

$$m = \frac{\sum_{x_i \in D} x_i}{N} \tag{9}$$

Query expansion ranking

Query expansion ranking method is our proposed feature selection method inspired by the query expansion methods from the field of information retrieval (IR). Query

expansion helps to find more relevant documents for a given query. It does so by adding new terms to the query. The new terms are selected from documents that are relevant to the original query so that the expanded query can retrieve more relevant documents. More specifically, terms from the relevant documents are extracted along with some scores, and those with the highest scores are included in the expanded query.

We propose a new feature selection method inspired by the query expansion technique developed for probabilistic weighting model proposed by Harman [12]. Harman [12, 36] studies how to assign scores to terms extracted from relevant documents for a given query Q so that high scored terms are used to expand the original query and improve precision of information retrieval strategy. In this method, first, query Q is sent to the information retrieval system, and then the system returns documents that are found as relevant to the user. Then, user examines the returned documents and marks the ones that are relevant with the query. After that, all the terms in the relevant documents are extracted and they are assigned scores by using a score formula as proposed by Harman [12], and top scored k terms are chosen as the most valuable terms to expand the query. Then, the expanded query Q', which includes the terms in the original query plus the k new terms that have the top-k scores, is sent to the information retrieval system to return more relevant documents to the original query Q. Equation 10 presents the score formula developed by Harman [12] to calculate ranking score of a term f extracted from the set of relevant documents for a given query Q.

$$Score_f = log_2 \frac{p_f(1-q_f)}{(1-p_f)q_f} \tag{10}$$

where p_f is the probability of term f in the set of relevant documents for query Q, and q_f is the probability of term f in the set of non-relevant documents for query Q. These probability scores are computed according to Robertson and Sparck Jones [13].

We revise the above score computation method to develop an efficient feature selector for SA. In our feature selection method, we propose a score formula given in Eq. 11 to compute scores for features:

$$Score_f = \frac{p_f + q_f}{|p_f - q_f|} \tag{11}$$

where p_f is the ratio of positive documents containing feature f and q_f is the ratio of negative documents containing feature f, which are computed according to Eqs. 12, 13, respectively:

$$p_f = \frac{DF_+^f + 0.5}{N^+ + 1.0} \tag{12}$$

$$q_f = \frac{DF_-^f + 0.5}{N^- + 0.5} \tag{13}$$

where DF^f_+ and DF^f_- are the raw counts of documents that contain f in the positive and negative classes, respectively and N^+ and N^- are the numbers of documents in the positive and negative classes, respectively. In the probability calculations, we add small constants to the numerators and denominators in Eqs. 12, 13 following Robertson and Sparck Jones [13] who add similar constants to avoid having zero probabilities. Such a method is known as data smoothing in statistical language processing.

In QER feature selection method, scores of features are computed before the features having the lowest scores are selected and used in the classification process. When a feature has low score, the difference between the probabilities for the positive and negative classes is high; therefore the feature is more class specific and more valuable for classification process. Among the feature selection methods we considered, we notice that IG and OCFS are good at distinguishing multiple classes, while CHI2, DFD, and QER are restricted to two classes, although all of them are suitable for sentiment analysis. IG is considered as a greedy approach since it favors those that can maximize the information gain for separating the related classes. Although CHI2 tries to identify the features that are dependent to a class, it can also give high values to rare features that only affect few documents in a given collection. OCFS has been shown to be effective for traditional topic-based text classification, but it depends on the distance/similarity measures between the vectors of the related documents. Since sentiment-expressing features do not happen frequently within a review, as illustrated by the example in the introduction, they may not be favored by the OCFS method. QER is similar to DFD in that they both rely on the differences of the document frequencies of a given feature between the two classes. However, QER is different from DFD in that it normalizes the document frequencies of a feature in both classes into probabilities and uses the ratio of the sum over the difference for these two probabilities.

Experiments and results

Datasets

We use Turkish and English review datasets in our experiments. The Turkish movie reviews are collected from a publicly available website (http://www.beyazperde.com) [30]. The dataset has 1057 positive and 978 negative reviews. The Turkish product review dataset is collected from an e-commerce website (http://www.hepsiburada.com) from different domains [28]. It consists of four subsets of reviews about books, DVDs, electronics, and kitchen appliances, each of which has 700 positive and 700 negative reviews. To compare our results with existing work for sentiment analysis, we use similar datasets for English reviews. The English movie review dataset is introduced by Pang and Lee [7], and consists of 1000 positive and 1000 negative reviews. English product review dataset is introduced by Blitzer et al. [37] and also has four subsets: books, DVDs, electronics, and kitchen appliances, with 1000 positive and 1000 negative reviews for each subset. In order to keep the same dataset sizes with Turkish product reviews, we randomly select 700 positive and 700 negative reviews from each subset of the English product reviews.

Performance evaluation

The performance of a classification system is typically evaluated by F measure, which is a composite score of precision and recall. Precision (P) is the number of correctly

Table 1 Baseline results in *F* measure for the Turkish and English review datasets

	Turkish review datasets					English review datasets				
	Features	**NBM**	**SVM**	**J48**	**LR**	**Features**	**NBM**	**SVM**	**J48**	**LR**
Movie	18,578	0.8248	0.8161	0.6954	–	38,869	0.8129	0.8480	0.6769	–
DVDs	11,343	0.7957	0.7320	0.6886	–	17,674	0.7836	0.7649	0.6789	–
Electronics	10,911	0.8155	0.7707	0.7371	–	9010	0.7629	0.7856	0.6750	–
Book	10,511	0.8317	0.7955	0.7019	–	18,306	0.7619	0.7485	0.6407	–
Kitchen	9447	0.7762	0.7407	0.6647	–	8076	0.8099	0.8136	0.7093	–

classified items over the total number of classified items with respect to a class. Recall (R) is the number of correctly classified items over the total number of items that belong to a given class. Together, the F measure gives the harmonic mean of precision and recall, and is calculated as follows [33]:

$$F = 2 \times \frac{P \times R}{P + R} \tag{14}$$

Since we are doing multi-fold cross validations in our experiments, we use the micro-average of F measure for the final classification results. This is done by adding the classification results for all documents across all five folds before computing the final P, R, and the F.

Experimental settings

We conduct the experiments on a MacBook Pro with 2.5 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3. We use Python with NLTK [38] library in our experiments. After tokenizing text into words along with case normalization, we keep some punctuation marks and stop words, as they may express sentiments (e.g., punctuation marks like exclamation and question marks, and stop words like "too" in "too expensive"). In addition, we do not apply stemming as Turkish is an agglutinative language and the polarity of a word is often included in the suffixes. Therefore, we can have a large feature space and it becomes important to apply feature selection methods to reduce this space. For sentiment classification, we use the Weka [33] data mining tool, which contains the four classifiers we use in our experiments, i.e., NBM, SMO for SVM, J48 for C4.5, and LR for MEM. Since our datasets are relatively small with at most a couple of thousands of documents, we apply the fivefold cross validation, which divides a dataset into five portions: four of them are used for training and the remaining one for testing, and then these portions are rotated to get a total of five F measures. Table 1 the average F measures for all the classifiers where the whole feature spaces are used for each dataset, except the LR classifier since it requires too much memory to handle the whole feature spaces for these datasets. As can be seen in Table 1, the total number of features without any reduction ranges from 9000 to 18,000 for the Turkish review datasets, and 8,000–38,000 for the English review datasets. These results form the baselines of our study and any new results obtained with feature selection methods by applying five folds cross validation can be compared for possible improvements.

Performance of feature selection methods for Turkish reviews

We tested five feature selection methods: QER, CHI2, IG, DFD, and OCFS on both Turkish and English review datasets. For each feature selection method, we tried six feature sizes at 500, 1000, 1500, 2000, 2500, and 3000, since this is the range typically considered for text classification, and in terms of total features, we have 9000–18,000 for the Turkish review datasets, and 8000–38,000 for English review datasets from our baseline systems. In our previous study [10], we also observed that feature sizes up to 3000 tend to give good classification performance. For all feature selection methods, we pick the top-ranked features of a desirable size n based on the scores of the related formulas for these methods. All of these settings are run against four classifiers: NBM, SVM, LR, and J48, resulting in a total of 120 experiments for each review dataset. Table 2 summarizes the best results for all pairs of feature selection methods and Turkish review datasets. For each pair, we show the best micro-average F measure along with the corresponding classifier and feature size. Also, the best results for each review dataset are given in bold-face.

As observed in Table 2, our new method QER is the best performer for each review dataset. CHI2 and IG have almost the same performance for the Turkish reviews and have better results than DFD and OCFS for the movie, book, DVDs, and kitchen review datasets. DFD with NBM classifier has better results than CHI2, IG, and OCFS for the electronics review dataset. Also, CHI2, IG, and QER tend to work well with smaller feature sizes, while DFD and OCFS tend to favour bigger feature sizes. Note that DFD does reasonably well across all review datasets, which confirms our intuition that sentiment-expressing words usually have low frequencies within a document, but relatively high frequencies across different documents. Although OCFS is quite robust for traditional topical text classification as reported in Cai and Song [39], it is not doing well for sentiment analysis, perhaps for the same intuition as we just explained for DFD. Once again, NBM remains to be the best for most of our experiments except that SVM does the best for the kitchen reviews when analysed with the CHI2 and IG methods. When analysed by univariate ANOVA and post hoc tests for the book, DVDs, electronics, and kitchen review datasets, we found that there are significant differences between three groups (Baseline and OCFS), (DFD, CHI2, and IG) and (QER) at 95% confidence level. Within each group, however, there are no significant differences. For the movie review dataset, there are significant differences between two groups (Baseline and OCFS), and (DFD, CHI2, IG, and QER) at the 95% confidence level. Overall, feature selection methods are shown to be effective for sentiment analysis, improving significantly over the baseline results.

To examine the effects of text classifiers, we show the best classification results for pairs of feature selection methods and text classifiers on the electronic review dataset in Table 3. Note that NBM does the best for all review datasets; J48 the worst; and SVM and LR in between, although LR is consistently better than SVM except for the QER method. One reason that the decision-tree-based solution J48 does not do well for text classification in general [40] and sentiment analysis in specific is that it is a greedy approach, always trying to find the features that separate the given classes the most. As a result, the classifier may use a much smaller set of features, even though there are many more relevant features are available. SVM typically does well for the traditional topic-based text

Table 2 The best classification results for pairs of feature selection methods and the Turkish review datasets

	QER		DFD		OCFS		CHI2		IG	
	Size	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**
Movie	3000	NBM:**0.9112**	3000	NBM:0.8864	3000	NBM:0.8447	1500	NBM:0.8883	1500	NBM:0.8883
DVDs	1500	NBM:**0.9136**	3000	NBM:0.8650	3000	NBM:0.8129	500	NBM:0.8671	500	NBM:0.8671
Electronics	1500	NBM:**0.8996**	1500	NBM:0.8567	2000	NBM:0.8337	1000	NBM:0.8564	1500	NBM:0.8551
Book	1500	NBM:**0.9150**	1500	NBM:0.8771	3000	NBM:0.8506	1000	NBM:0.8864	1000	NBM:0.8864
Kitchen	1000	NBM:**0.8790**	3000	NBM:0.8314	3000	NBM:0.8017	500	SVM:0.8378	500	SVM:0.8378

Table 3 Detailed results for the Turkish electronics review dataset

	NBM		SVM		LR		J48	
	Size	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**
QER	1500	0.8996	2000	0.8715	1000	0.7927	2000	0.6734
CHI2	1000	0.8564	1000	0.8505	500	0.7969	1000	0.7435
IG	1500	0.8551	1000	0.8505	500	0.8156	1500	0.7428
DFD	1500	0.8567	1500	0.8128	2500	0.7829	500	0.7399
OCFS	2000	0.8337	1000	0.7729	3000	0.7643	1500	0.7371

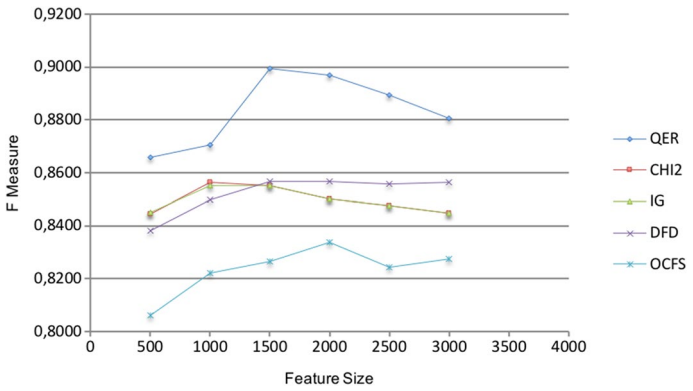

Fig. 1 Detailed results of feature sizes for the Turkish electronic review dataset

classification by finding a hyperplane that clearly separates the two classes [40]. In order to do this, we need to represent documents as weighted vectors so that we can measure the distances or similarities between the documents. For sentiment analysis, however, we are favouring features that have low frequencies within a document, but relatively high frequencies across different documents (as illustrated by the example of "great" in the introduction), making the distance/similarity measures less effective. Both NBM and LR are based on the probabilities of the features in the given dataset. In particular, LR is equivalent to the maximum entropy modelling and is capable of handling dependent features, whereas NBM makes the naïve assumption that all features are independent of each other. In our experiments, NBM does better than LR, which could be due to the same reason as we just explained for SVM above.

To see the impacts of feature sizes for different feature selection methods, we plot our results for the Turkish electronic review dataset in Fig. 1. Clearly, OCFS lags behind other feature selection methods across all feature sizes. DFD tends to do better with bigger feature sizes, while CHI2 and IG tend to favour smaller feature sizes. In addition, the results for CHI2 and IG are sufficiently close, although they are slightly different for certain feature sizes. Our new method QER does reasonably well across all other methods. For Turkish electronics review dataset, QER is the best performer and the selected features include 7.7% of the punctuation patterns and 25% of the stop words; the features selected by DFD method include 61.5% of the punctuation patterns and 59% of the stop words; the features selected by CHI2 method include 15% of the punctuation patterns and 90% of the stop words; and the features selected by OCFS method include 69.2% of

the punctuation patterns and 49.6% of the stop words. Therefore, CHI2 method tends to favor stop words but not punctuation patterns, while DFD and OCFS tend to choose more punctuation patterns and fewer stop words. In addition, when we compare the features selected by QER and CHI2 methods, we observe that 5.7% of selected features are the same, and for QER and DFD methods, there are 6.9% of the features that are common, and for QER and OCFS methods, there are 7% of the features that are common. However, for DFD and CHI2 methods, we observe that 49.8% of the selected features are the same, and for DFD and OCFS methods, there are 76.7% of the features that are common, and for CHI2 and OCFS methods, there are 34% of the features that are common. Note that although we only show the results on specific datasets in Table 3 and Fig. 1, similar trends are observed for other datasets as well, and to save space these results are not included.

Performance of feature selection methods for English reviews

Using similar settings as described in "Performance of feature selection methods for Turkish reviews", we also carried out experiments on the English review datasets. As shown in Table 4, QER achieved the best performance with LR classifier for the movie review dataset and NBM classifier for other datasets. CHI2 and IG achieved better performance with NBM for all five datasets. Once again, the results are basically the same for CHI2 and IG, indicating that the two methods are also strongly correlated for the English review datasets. Compared with the Turkish movie reviews, the feature size for the best performer of the English movie reviews is 3000, which is achieved with QER for the LR classifier. This is likely due to the bigger vocabulary of the English movie reviews over that of the Turkish movie reviews as can be observed in Table 1. Also compared with the Turkish review datasets, DFD is not as good as CHI2 and IG for the English review datasets, even though the performance is close for the kitchen reviews and generally better than OCFS. Furthermore, the best results for DFD are achieved with different classifiers for different datasets: SVM for the movie reviews and LR for the kitchen reviews. Statistical analysis with univariate ANOVA and post hoc tests show similar results as those for the Turkish reviews: there are significant differences between three groups (Baseline and OCFS), (DFD), and (CHI2, IG, and QER) at 95% confidence level for the movie, DVDs, electronic, and kitchen review datasets, but for the book review dataset, there are significant differences between two groups (Baseline and OCFS) and (DFD, CHI2, IG, and QER) at the 95% confidence level.

For text classifiers, Table 4 shows that similar trends are observed for the English reviews as those for the Turkish reviews, although LR and SVM can over-perform NBM for some feature selection methods. For different feature sizes, similar trends are also observed, as illustrated in Fig. 2. Once again, in Table 5 and Fig. 2, we only show the results for specific datasets, but the trends are similar to other datasets as well.

In summary, we see some similarities between Turkish and English reviews in that for data pre-processing, we should keep punctual patterns and stop words, and not perform stemming, leading us to use the same setting as the baselines for further study. In addition, NBM seems to be the most suitable classifier for sentiment analysis since sentiment-expressing words tend to have low frequencies within a document, but relatively high frequencies across different documents. For feature selection methods, our

Table 4 The best classification results for pairs of feature selection methods and the English review datasets

	QER		DFD		OCFS		CHI2		IG	
	Size	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**
Movie	3000	LR:**0.9550**	2500	SVM:0.8640	3000	SVM: 0.8285	2500	NBM:0.9150	2500	NBM:0.9150
DVDs	2500	NBM:**0.9169**	3000	NBM:0.8502	1000	NBM:0.7996	1000	NBM:0.8964	1000	NBM:0.8964
Electronics	2000	NBM:**0.8878**	1500	NBM:0.8221	2000	SVM: 0.7821	1000	NBM:0.8621	1000	NBM:0.8621
Book	3000	NBM:**0.9162**	3000	NBM:0.8628	3000	NBM:0.7899	1000	NBM:0.8879	1000	NBM:0.8879
Kitchen	2000	NBM:**0.9106**	3000	LR:0.8893	1500	SVM: 0.8157	500	NBM:0.8964	500	NBM:0.8964

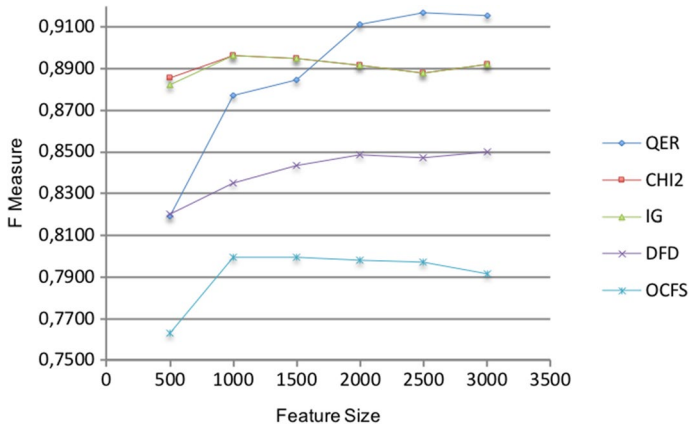

Fig. 2 Detailed results of feature sizes for the English DVDs review dataset

Table 5 Detailed results for the English DVD review dataset

	NBM		SVM		LR		J48	
	Size	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**	**Size**	***F* measure**
QER	2500	0.9169	3000	0.8724	2000	0.8977	2000	0.5481
CHI2	1000	0.8964	500	0.8650	3000	0.6976	3000	0.6799
IG	1000	0.8964	1000	0.8614	2000	0.6970	500	0.6769
DFD	3000	0.8502	1000	0.8293	3000	0.7600	500	0.6771
OCFS	1000	0.7996	1000	0.7714	500	0.6800	2000	0.6829

proposed QER achieves best performances with feature sizes between 2000 and 3000. CHI2 and IG are strongly correlated and tend to work well with smaller feature sizes, while DFD also works reasonably well, but with bigger feature sizes. For differences, the English review datasets usually have bigger vocabulary, resulting in relatively bigger feature sizes for feature selection. Moreover, SVM and LR can also perform well for some English review datasets, while NBM looks like a dominant classifier for the Turkish reviews. Finally, the performance results for the English reviews are generally higher than those for the Turkish reviews, possibly related to the differences between the two languages in terms of vocabularies, writing styles, and the agglutinative property of the Turkish language. The limitation of QER is that it is only suitable for classifying two classes since it is especially developed for sentiment analysis with the observation that sentiment-expressing words are usually more frequent across different reviews. The contribution of QER is that, as it is shown in the experimental results, the method is both language and classifier independent and can select better features than other methods for sentiment analysis.

Comparison of our proposal with the previous studies

It is generally difficult to directly compare the results of different studies since there are often differences in partitioning and preprocessing the datasets for training and testing, as shown in the studies by Pang et al. [4]. That is why we tried different combinations of feature selection methods and text classifiers on multiple datasets in our research so that

Table 6 Summary of related work on the sentiment analysis for the same datasets

Paper	Dataset	Baseline accuracy (%)	Best accuracies observed (%)	Classifier
[4]	Movie	78.7		NB, SVM
[7]	Movie		87.1 minimum cut	SVM
[8]	Movie Product	79.9 74.3	85.7 CHI2; 86.9 DFD; 80.9 OCFS 73.7 CHI2; 75 DFD; 73.8 OCFS	MEM
[9]	Movie Product	84.2 80.9 Book; 78.9 DVD; 80.8 El	91.8 92.5 Book; 91.5 DVD; 91.8 El mRMR with composite features	BNBM, SVM
[23]	Product	70.1	84.2% Kitc. semantic orientation	SVM
[24]	Movie Product	84.8 74.7 Book; 77.2 DVD; 80.8 El.; 83.3 Kitc	87.7 81.8 Book; 83.8 DVD; 85.9 El.; 88.7 Kitc word relation based method	NB, SVM, MEM
[25]	Movie	84.1	92.7% Tabu search-enhanced Markov blanket model	NB, SVM, MEM
Our study	Movie Product	84.8 76.2 Book; 78.4 DVD; 78.6 Elect; 81.4 Kitc	91.5 CHI2-IG; 87.1 DFD; 82.9 OCFS; 95.5 91.6 Book; 91.7 DVD; 88.8 Elect; 91.1 Kitc proposed QER	NBM, SVM, MEM, DT

we can compare their performance collectively and accurately. However, we do agree that it is helpful to describe the results from the related studies so that we can put our results into a suitable context. Table 6 includes a summary for comparison of our results with that of the previous studies which have used the same datasets with our study. For the English movie review dataset, Nicholls and Song [8] obtained a baseline accuracy of 79.9% with the MEM classifier, and better classification accuracies of 86.9, 85.7, and 80.9% when combined with DFD, CHI2, and OCFS feature selection methods, respectively. Dang et al. [23] examined their proposed semantic oriented method on the product dataset [37]. They achieved an accuracy of 84.2% for the kitchen dataset. Also, Xia et al. [24] improved the classification performances from 84.8 to 87.7% using their proposed word relation based feature selection method. Bai [25] improved the accuracies from baseline 84.1–92.7% using their proposed Tabu search-enhanced Markov blanket model for the movie review dataset. Pang et al. [4] obtained accuracy around 78.7% with NB using the document frequency of 4 to eliminate the rare features. Agarwal et al. [9] improved the accuracies from baseline 82.7–89.2% using IG feature selection method with Boolean NBM. Our proposed QER method showed an improvement from the baseline of 81.3–91.1% with NBM in terms of F measures.

For the Turkish movie review dataset, the best classification result of 82.58% is obtained with the SVM classifier [30]. As shown in the previous studies, classification accuracy is improved by applying feature selection, and NB based classifier performs the best in the majority of the cases. The proposed feature selection method is also computationally efficient and easy to implement as it only computes scores for features by counting document frequencies.

Conclusions

In this paper, we proposed a new feature selection method query expansion ranking (QER) for the sentiment analysis and compared it with the common feature selection methods for sentiment classification, including DFD and OCFS, CHI2 and IG. All of these methods are tested against five datasets of Turkish reviews, using four common

text classifiers, including NBM, SVM, logistic regression (LR), and decision trees (J48). Similar experiments are also conducted for English reviews so that we can compare their differences with the Turkish reviews. Our results show that for all Turkish review datasets, the best results are all obtained with the NBM classifier, and for some English review datasets, LR and SVM have the best performance. For feature selection, our proposed QER method helps to achieve the best performance compared with all other feature selection methods for both Turkish and English reviews. For feature selection, our experiments show that our proposed QER method helps to achieve the best performance among all other feature selection methods. We found that CHI2 and IG have almost the same performance for the Turkish reviews and they tend to work well with smaller feature sizes compared with other feature selection methods. DFD does reasonably well across all review datasets, but it tends to favour bigger feature sizes. This confirms our intuition that sentiment-expressing words usually have low frequencies within a document, but relatively high frequencies across different documents. Although OCFS is quite robust for traditional topical text classification, it does not do well for sentiment analysis since it relies on word frequencies to measure the distances between documents. Once again, NBM remains the best performer for most of our experiments when analysed with QER method. Overall, feature selection methods are shown to be effective for sentiment analysis, improving significantly over the baseline results.

Following a similar process, we also carried out experiments on English review datasets and NBM seems to be the most suitable classifier for sentiment analysis. For feature selection methods, CHI2 and IG are strongly correlated and tend to work well with smaller feature sizes, while DFD also works reasonably well, but with bigger feature sizes. Our proposed query expansion ranking method achieves the best performances for the English datasets as well. As for differences, the English review datasets usually have a bigger vocabulary, resulting in relatively bigger feature sizes for feature selection. Moreover, LR and SVM also perform well for some English review datasets, while NBM looks like a dominant classifier for the Turkish reviews. The performance results for the English reviews are generally higher than those for the Turkish reviews, possibly related to the differences between the two languages in terms of vocabularies, writing styles, and the agglutinative property of the Turkish language. Finally, the experimental results show that our proposal QER method is language, domain and classifier independent and improve the classification performance better than other FS methods for sentiment analysis.

Authors' contributions

TP drafted this manuscript, conducted experiments using the datasets and analyzed the results. SAO and FS suggested the methods used in this study and provided guidelines in drafting the manuscript. FS edited and corrected the manuscript. All authors read and approved the final manuscript.

Authors' information

TP received her Ph.D. degree in Computer Engineering from Çukurova University in 2016. She received a Bachelor of Engineering degree in Computer Engineering from Hacettepe University, and she holds a M.Sc. in Management Information Sciences and a M.Sc. in Mathematics. She studied for 4 months of 2015 as a visiting researcher in University of Guelph, Canada with a scholarship supporting by The Scientific and Technological Research Council of Turkey (TUBITAK). She is currently working as a senior lecturer and head of the Computer Technologies Department, Antakya Vocational School, Mustafa Kemal University. Her research interest is in sentiment analysis, data mining, machine learning, and applying text processing techniques to medical data extraction and integration.

SAO received her Ph.D. and Bachelor of Science degrees both in Computer Engineering from Bilkent University, Turkey, in 2004 and 1996, respectively. Currently she is a professor and head of the Department of Computer Engineering, Çukurova University, Turkey. Her research interests include text mining, information retrieval systems, and applying biological and nature inspired computing to text mining.

FS received his Ph.D. degree in Computer Science from the University of Waterloo in Canada. He is currently an associate professor in the School of Computer Science, University of Guelph in Canada. His interests are mostly in Natural Language Processing, working on a wide range of topic areas, including information retrieval, text classification, topic modeling, key phrase extraction, text segmentation, sentiment analysis, text summarization, and document clustering. More recently, he is also interested in applying text processing techniques to privacy policy analysis and medical data extraction and integration.

Author details
[1] Department of Mathematics, Mustafa Kemal University, Antakya, Hatay, Turkey. [2] Department of Computer Engineering, Çukurova University, Adana, Turkey. [3] School of Computer Science, University of Guelph, Guelph, Canada.

Acknowledgements
This research is supported by TUBITAK-2214-A.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
This research is supported by Çukurova University Fund of Scientific Research Projects under Grant No. FDK-2015-3833, and Mustafa Kemal University Fund of Scientific Research Projects under Grant No. 15426.

Publisher's Note

Received: 10 February 2018 Accepted: 16 April 2018
Published online: 09 May 2018

References

1. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2:1–135. https://doi.org/10.1561/1500000011
2. Tripathy A, Anand A, Rath SK (2017) Document-level sentiment classification using hybrid machine learning approach. Knowl Inf Syst 53:805–831. https://doi.org/10.1007/s10115-017-1055-z
3. Yang D-H, Yu G (2013) A method of feature selection and sentiment similarity for Chinese micro-blogs. J Inf Sci 39:429–441. https://doi.org/10.1177/0165551513480308
4. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? In: Proceedings of the ACL-02 conference on empirical methods in natural language processing—EMNLP'02. Association for computational linguistics, Morristown, pp 79–86
5. Mullen T, Collier N (2004) Sentiment analysis using support vector machines with diverse information sources. Conf Empir Methods Nat Lang Process. https://doi.org/10.3115/1219044.1219069
6. Kaya M, Fidan G, Toroslu IH (2012) Sentiment analysis of Turkish political news. In: 2012 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology. IEEE, Macau, pp 174–180
7. Pang B, Lee L (2004) A sentimental education. In: Proceedings of the 42nd annual meeting on association for computational linguistics—ACL'04. Association for Computational Linguistics, Morristown, p 271–es
8. Nicholls C, Song F (2010) Comparison of feature selection methods for sentiment analysis. In: Advances in artificial intelligence. Springer, Berlin, pp 286–289
9. Agarwal B, Mittal N (2016) Prominent feature extraction for review analysis: an empirical study. J Exp Theor Artif Intell 28:485–498. https://doi.org/10.1080/0952813X.2014.977830
10. Parlar T, Ozel SA (2016) A new feature selection method for sentiment analysis of Turkish reviews. In: International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, Sinaia, pp 1–6
11. Fattah MA (2017) A novel statistical feature selection approach for text categorization. J Inf Process Syst 13:1397–1409. https://doi.org/10.3745/JIPS.02.0076
12. Harman D (1992) Relevance feedback revisited. In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval—SIGIR'92. ACM Press, New York, pp 1–10
13. Robertson SE, Jones KS (1976) Relevance weighting of search terms. J Am Soc Inf Sci 27:129–146. https://doi.org/10.1002/asi.4630270302
14. Aldoğan D, Yaslan Y (2017) A comparison study on active learning integrated ensemble approaches in sentiment analysis. Comput Electr Eng 57:311–323. https://doi.org/10.1016/J.COMPELECENG.2016.11.015
15. Singh J, Singh G, Singh R (2017) Optimization of sentiment analysis using machine learning classifiers. Hum centric Comput Inf Sci 7:32. https://doi.org/10.1186/s13673-017-0116-3
16. Mladenović M, Mitrović J, Krstev C, Vitas D (2016) Hybrid sentiment analysis framework for a morphologically rich language. J Intell Inf Syst 46:599–620. https://doi.org/10.1007/s10844-015-0372-5
17. Asgarian E, Kahani M, Sharifi S (2018) The impact of sentiment features on the sentiment polarity classification in Persian reviews. Cognit Comput 10:117–135. https://doi.org/10.1007/s12559-017-9513-1

18. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. https://doi.org/10.1016/j.aca.2011.07.027
19. Akba F, Uçan A, Sezer E, Sever H (2014) Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews. In: 8th European conference on data mining. Lisbon, Portugal, pp 180–184
20. Liu Y, Bi JW, Fan ZP (2017) Multi-class sentiment classification: the experimental comparisons of feature selection and machine learning algorithms. Expert Syst Appl 80:323–339. https://doi.org/10.1016/j.eswa.2017.03.042
21. Sagar K, Saha A (2017) Qualitative usability feature selection with ranking: a novel approach for ranking the identified usability problematic attributes for academic websites using data-mining techniques. Hum centric Comput Inf Sci 7:29. https://doi.org/10.1186/s13673-017-0111-8
22. Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans Inf Syst 26:1–34. https://doi.org/10.1145/1361684.1361685
23. Dang Y, Zhang Y, Chen H (2010) A Lexicon-enhanced method for sentiment classification: an experiment on online product reviews. IEEE Intell Syst 25:46–53. https://doi.org/10.1109/MIS.2009.105
24. Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci (Ny) 181:1138–1152. https://doi.org/10.1016/j.ins.2010.11.023
25. Bai X (2011) Predicting consumer sentiments from online text. Decis Support Syst 50:732–742. https://doi.org/10.1016/j.dss.2010.08.024
26. Yan J, Liu N, Zhang B, et al (2005) OCFS: optimal orthogonal centroid feature selection for text categorization. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval—SIGIR'05. ACM Press, New York, p 122
27. Zheng L, Wang H, Gao S (2018) Sentimental feature selection for sentiment analysis of Chinese online reviews. Int J Mach Learn Cybern 9:75–84. https://doi.org/10.1007/s13042-015-0347-4
28. Demirtas E, Pechenizkiy M (2013) Cross-lingual polarity detection with machine translation. In: Second international workshop on issues of sentiment discovery and opinion mining—WISDOM'13. ACM Press, New York, pp 1–8
29. Boynukalin Z (2012) Emotion analysis of Turkish texts by using machine learning methods. M.Sc. Thesis, Middle East Technical University
30. Sevindi BI (2013) Türkçe Metinlerde Denetimli ve Sözlük Tabanlı Duygu Analizi Yaklaşımlarının Karşılaştırılması. M.Sc. Thesis, Gazi University
31. Parlar T, Özel SA, Song F (2018) Interactions between term weighting and feature selection methods on the sentiment analysis of Turkish reviews. In: Computational linguistics and intelligent text processing. CICLing 2016. Lecture Notes in computer Science, vol 9624. Springer, Cham, pp 335–346
32. Çakici R (2009) Wide-coverage parsing for Turkish. Ph.D. Thesis, University of Edinburgh
33. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
34. McCallum A, Nigam K (1998) A comparison of event models for naive Bayes text classification. In: AAAI/ICML-98 workshop on learning for text categorization. pp 41–48
35. Zhao X, Li D, Yang B et al (2015) A two-stage feature selection method with its application. Comput Electr Eng 47:114–125. https://doi.org/10.1016/J.COMPELECENG.2015.08.011
36. Harman D (1988) Towards interactive query expansion. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval—SIGIR'88. ACM Press, New York, pp 321–331
37. Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: 45th annual meeting-association for computational linguistics. pp 440–447
38. Bird S, Klein E, Loper E (2009) Natural language processing with Python. O'Reilly, Newton
39. Cai J, Song F (2008) Maximum entropy modeling with feature selection for text categorization. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). pp 549–554
40. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. Springer, Berlin, pp 137–142