Make It Possible: Multilingual Sentiment Analysis without Much Prior Knowledge

Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Songbo Tan, Xueqi Cheng
CAS Key Laboratory on Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences,
Beijing, 100190, China
Email: {linzheng, jinxiaolong, xuxueke, wangyuanzhuo, tansongbo, cxq}@ict.ac.cn

***Abstract*—Sentiment analysis is a hard problem, while multilingual sentiment analysis is even harder due to the different expression styles in different languages. Although many methods for multilingual sentiment analysis have been developed in the open literature, most of them suffer from two major problems. The first is their excessive dependence on external tools or resources (e.g., machine translation systems or bilingual dictionaries), which may not be readily obtained, especially for minority languages; The second is conflictive sentiments, i.e., the sentiment polarity of some parts of a text is inconsistent with its overall sentiment polarity. It is observed that in a product or service review there usually exist a few sentences which play a more important role in determining its sentiment polarity, as compared to others. Therefore, differentiating key sentences from trivial ones may be helpful to improve sentiment analysis. Inspired by this observation in this paper we propose a novel framework to estimate the sentiment polarity of reviews by virtue of opinion lexica and key sentences automatically extracted from unlabelled data. This framework cannot only overcome the problem of excessive dependence on external resources, but also is able to capture the overall sentiment polarity of reviews. Experimental results on realistic review datasets demonstrate that the proposed framework is effective and competitive with the representative baselines.**

I. Introduction

Sentiment analysis [10] aims to automatically identify the sentiment polarity of given texts, which has broad applications, including recommendation systems [23], sentiment summarization [7], opinion retrieval [17], and so on. Given the explosively growing number of online reviews in different languages, multilingual sentiment analysis has recently attracted a great deal of attention from both academia and industries [3], [8], [16], [26]. According to the resources employed, existing methods for multilingual sentiment analysis can basically be categorized into two types, namely, machine-translation-based methods and bilingual-dictionary-based methods.

Machine translation (MT) has been widely employed in cross-language related work. For example, it is often used to translate the labelled data in a source language into a target language [2], [4], [25]. However, such machine-translation-based methods are confronted with three problems: First, they are inefficient when dealing with massive data; Second, current MT systems are not powerful to achieve accurate results. Particularly, they usually generate one best translation, which may not be suitable for the situation at hand; Third, the models used in statistical MT rely on a set of characteristics observed on training examples, but large-scale bilingual parallel corpora for a specific domain are not available in some cases.

Utilizing bilingual dictionaries [12] in multilingual sentiment analysis could be effective as the methods using a high-quality MT system [19], [22]. Bilingual dictionaries cannot only reduce workload for labelling data, but also allow one integrating various term weighting and selection methods. However, comprehensive bilingual dictionaries may not be always available, especially for minority language pairs, while generating a bilingual dictionary is difficult and laborious.

In addition to the above issue of resource dependency, another grand challenge of multilingual sentiment analysis is sentiment analysis itself. Sentiment analysis is a hard problem, because many reviews are sentimentally ambiguous for many reasons. For instance, objective statements interleaved with subjective statements can be confusing for learning methods, and subjective statements with conflictive sentiments further make sentiment analysis more complicated [29]. Take a book review for example:

> This book is beautiful.
> .....
> Zusak's novel, set in a small town outside Munich during World War II, chronicles the story of Liesel Meminger, a German girl taken into Hans Huberman's household as a foster child. As likeable as she is well-developed, it's amazing to watch a young girl like that remain so strong in the face of human tragedy, impossible hatred.....

Here, the reader describes the trivial plot using negative words such as "war" and "tragedy". But, s/he enthusiastically expresses that s/he likes the book at the beginning of the review. In this case, the overall sentiment polarity of the review is positive, but is apt to be labelled as a negative one if all sentences are treated equally. In the case of multilingual sentiment analysis where the different expression styles in different languages and cultures are considered, the conflictive sentiments problem becomes more difficult.

To solve the above problems, in this paper we propose a novel multilingual sentiment analysis framework. In the proposed framework, no manually labelled corpus is needed and all extracted information is domain-dependent. In general, the contributions of this study can be summarized as follows:

1) We propose a statistical method for opinion lexicon extraction based on a few seed words, which can be easily transplanted to almost any language and does not need to refer to synonyms and antonyms dictionaries;

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.83

2) On the basis of the extracted opinion lexicon, we propose a key sentence extraction method for capturing the overall opinion of reviews, which solves the problem of conflictive sentiments;
3) We further propose a Self-Supervised Learning (SSL) method for sentiment classification, which combines unsupervised and supervised techniques together by virtue of the above extracted opinion lexicon and key sentences;
4) Finally, extensive experiments on multilingual datasets in different domains well demonstrate the effectiveness of the proposed methods.

The rest of this paper is organized as follows. Section II introduces related work. Section III describes in detail the proposed framework and methods for multilingual sentiment analysis. In Section IV, experimental results are presented and analyzed. Finally, the paper is concluded in Section V.

II. Related Work

In this section, we make a brief introduction to related studies on multilingual sentiment analysis. We first present traditional monolingual sentiment analysis and then introduce pilot studies on multilingual sentiment analysis. Finally, we review related work on opinion lexicon mining.

A. Monolingual Sentiment Analysis

Sentiment analysis aims to identify the sentiment polarity of given texts, such as, reviews, blogs, and microblogs, etc. According to the training mode, the existing methods for sentiment analysis can be roughly categorized into two types, namely, supervised methods and unsupervised methods.

Supervised methods usually regard polarity identification as a classification task and use a labelled corpus to train a sentiment classifier. For instance, Pang et al. [18] conducted polarity classification of reviews using Support Vector Machines, Naive Bayes and Maximum Entropy classifiers. Gamon [11] demonstrated that using large feature vectors in combination with feature reduction, high accuracy can be achieved in even very noisy data of customer feedback. Koppel and Schler [14] used neutral reviews to help improve the classification of positive and negative reviews.

Unsupervised methods usually conduct sentiment classification using a sentiment lexicon. Turney and Littman [24] estimated the sentiment polarity of a review by the average semantic orientation of its phrases that contain adjectives or adverbs. Zagibalov and Carroll [30] described an automatic seed word selection method for unsupervised sentiment classification of product reviews in Chinese. Qiu et al. [21] used a lexicon-based iterative process to iteratively enlarge an initial sentiment dictionary, but they started with a much larger HowNet Chinese sentiment dictionary as the initial lexicon.

In general, monolingual sentiment analysis methods are often dependent on external resources, whilst these resources are unbalanced among different languages. Therefore, traditional monolingual sentiment analysis methods are usually not appropriate for multilingual sentiment analysis.

B. Multilingual Sentiment Analysis

Several pilot studies have been carried out for multilingual sentiment analysis [8], [16], [19], [25], [26]. Existing approaches usually rely on machine translation systems, parallel corpora or labelled data. For instance, Wan [25] applied a co-training method to leverage resources from both source and target languages, where an online translation service was used to translate labelled English reviews into Chinese and unlabelled Chinese reviews into English. These translation-based methods are over dependent on machine translation tools. Michalea et al. [16] discussed different shortcomings of a lexicon-based translation scheme and proposed to use a parallel corpus to train a new classifier. However, building a parallel corpus is time-consuming and a parallel corpus may be scarce for some language pairs especially for minority languages. Shi et al. [22] transferred classification knowledge by translating model features and using an expectation maximization algorithm to automatically learn feature translation probabilities from labelled text in a source language and unlabelled text in a target language. Unlike the above work, the framework proposed in this paper is self-supervised, requiring no external resources. What it needs are unlabelled corpora of the target language and a few seed words.

C. Opinion Word Mining

Extensive studies have been conducted on sentiment analysis at word level. Wiebe [28] and Wiebe et al. [27] proposed an approach to find subjective adjectives using the results of word clustering according to their distributional similarity. However, they did not analyze the sentiment polarities of the found subjective adjectives. Qiu et al. [20] proposed a bootstrapping method to extract target and opinion words using a dependency parser. However, external tools and resources such as dependency parsers are available only for a handful of languages, which confines the applicability of these approaches. In addition, Hassan and Radev [13] applied a Markov random walk model to a large word graph where words are connected if they occur in the same WordNet synset. However, the dictionary-based methods are unable to find domain-dependent opinion words, because most entries in dictionaries are domain-independent.

In general, most related work on opinion lexicon extraction heavily relies on advanced Natural Language Processing (NLP) tools (e.g., syntactic parsers [20], search engines [24]) or broad-coverage external resources (e.g., WordNet [9], [13]). Moreover, these methods are designed to work well in a single language and are difficult to be transplanted to other languages [5], as resources of different languages are unbalanced in terms of both quality and quantity. Unlike these existing methods, in this paper we focus on mining opinion words from corpora with few resources and tools, which is both domain- and language-independent.

III. The Proposed Approach

As aforesaid, there are two major problems in multilingual sentiment analysis, namely, excessive dependence on external resources and conflictive sentiments. In order to solve these problems, we propose a simple and cost-efficient framework for multilingual sentiment analysis, as shown in Figure 1, which can be divided into three key components:

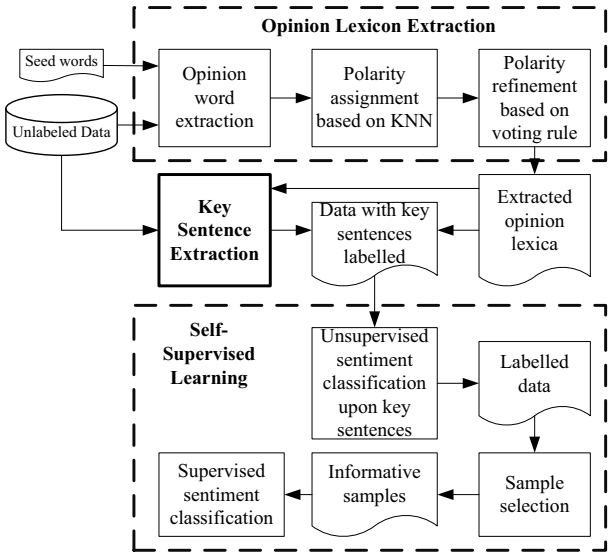


Fig. 1. The schematic diagram of the proposed framework for multilingual sentiment analysis.

- First, an opinion lexicon extraction method (see Subsection III-A) is proposed to overcome the problem of excessive dependence on external resources. Actually, extensive work in the open literature has been done on opinion lexicon extraction. Here, we do not intend to beat existing methods in terms of performance, but take a different perspective and focus on developing an unsupervised and language-independent method for resource-poor languages. In this method, opinion words are first extracted from unlabeled data based on only a few seed words. Next, the sentiment polarities of all extracted opinion words are initialized using a K-Nearest-Neighbor (KNN) method and further refined by a voting mechanism among multiple domains.
- Second, a key sentence extraction method (see Subsection III-B) is developed based on the extracted opinion lexicon to cope with conflictive sentiments. One of the main challenges for document-level sentiment classification is that not every part of the document is equally informative for inferring the polarity of the whole document. Therefore, we believe that making a difference between key sentences and trivial ones will be helpful to improving the performance of sentiment classification.
- Third, a Self-Supervised Learning (SSL) method (see Subsection III-C) is devised to generate a sentiment classifier for multilingual sentiment analysis. In this method, an unsupervised mechanism is first designed to label the sentiment polarity of those data where the key sentences of reviews have been clearly marked. Next, a selection strategy is developed to select some informative reviews with clear sentiment polarity. Finally, the selected reviews are employed to train a sentiment classifier using a certain supervised learning algorithm.

A. Opinion Lexicon Extraction

An opinion lexicon is a list of opinion words, such as *good* and *poor*, along with sentiment polarities. Traditional methods for extracting opinion lexica rely on language-dependent NLP tools (e.g., POS tagger), external resources (e.g., WordNet), or labelled data. In this paper, we will extract opinion lexica from unlabelled data using heuristic information among different languages.

It is known that for any language there are adverbs of degree. In this paper, we adopt only two adverbs of degree to extract opinion words. The first one is *very* in English and the corresponding words in other languages, for example, *très* in French. This kind of adverbs is frequently used to modify the degree of opinion words. The second word is *highly* in English. In this paper, most opinion words are extracted by the first adverb of degree *very*, while the second adverb of degree *highly* plays an auxiliary role to recall more opinion words. For non-English language, the two words are obtained by translating *very* and *highly* via Google Translate.

The process of opinion word extraction is quite straightforward. Take English for example, we first extract a candidate opinion word, denoted as w_i, according to two patterns, "*very* w_i" and "*highly* w_i". Next, we remove stop words from the candidates of opinion words. Here, stop words are identified according to their high frequency in the data for each language. Finally, we adopt the remained words that appears more than once as opinion words. Note that as online reviews are complex and diverse, the words that only appear once are likely to be noise.

Next, we need to identify the sentiment polarity of each extracted opinion word. In this paper, the process follows two steps. First, for a given domain, an initial polarity is assigned to each opinion word according to the polarity information of its K Nearest Neighbors (KNN). Since only a few seed words are used, we apply an iterative KNN strategy to propagate their polarity to all the other words. Here, the similarity between two words is measured by Pointwise Mutual Information (PMI), which is wildly used in the NLP field:

$$sim(o_i, o_j) = \log \frac{p(o_i, o_j)}{p(o_i)p(o_j)}, \tag{1}$$

where $p(o_i, o_j)$ is the probability of words o_i and o_j co-occurring in the same sentence, $p(o_i)$ and $p(o_j)$ are the probabilities of words o_i and o_j occurring in any sentence, respectively. Second, the polarity of an opinion word is refined by a voting rule upon multiple domains. The underlying idea is that when we have little prior knowledge of the target language, it may be reasonable to try multiple domains in hope of that combining their initial results would lead to good performance. In the voting process, in order to refine the sentiment polarity of an opinion word in a domain, two auxiliary domains are involved for voting. That is to say, if we want to refine the sentiment polarity of an opinion word in a domain, the initial polarity information of this word in other two domains is taken into account. Specifically, for an opinion word, the voting results can be summarized into three cases:

1) The initial polarities in all the three domains are the same.

2) The initial polarities in the two auxiliary domains are the same, but different from that in the domain at hand.
3) The initial polarity in the domain at hand is the same as that in one auxiliary domain, but different from that in the other one.

In Cases 1 and 3, the polarity of the main domain is confirmed. Since the polarity of an opinion word is actually domain-dependent, even if its initial polarity in the domain at hand is different from those in the two auxiliary domains, it may still be appropriate for that domain. Therefore, in Case 2 we do not simply change the polarity of the opinion word in the domain at hand to the one in the two auxiliary domains in order to prevent from ignoring its domain-dependent property. Instead, we need to further judge its sentiment polarity by comparing its polarity score (see Equation (3)) in the given domain to those in the auxiliary domains. In more detail, if its polarity score in the given domain is greater than the sum of the polarity scores in the auxiliary domains, its initial polarity in the given domain is confirmed. Otherwise, it is changed to the one in the auxiliary domains. We have experimentally verified that the above voting rule improves the polarity assignment precision especially for the general (i.e., domain-independent) opinion words, while for domain-dependent opinion words the above re-confirmation mechanism prevents polarity modification from over correction.

B. Key Sentence Extraction

As we mentioned before, one challenge for document-level sentiment analysis is that not every part of a review is equally informative for inferring its sentiment polarity. Generally, the polarity of a review mainly depends on the overall evaluation of the reviewer rather than the details about some specific aspects. Therefore, for a given review, we hope to extract the most important sentences that express the overall sentiment or attitude of the reviewer. It should be pointed out that key sentence extraction in this paper is different from sentiment summarization [1], [15] in two aspects: First, sentiment summarization aims to generate a summary that well represents the overall sentiment of a set of documents, while key sentence extraction in this paper intends to select sentences from a single document that well express the overall sentiment polarity of the author; Second, a sentiment summary is subject to predefined length constraints. But, there is no length constraint for key sentence extraction.

Next, before elaborating on the details of the key sentence extraction approach, we first explore the characteristics of key sentences from multiple perspectives. First of all, from the position perspective, we have observed that key sentences are usually located in the beginning or the end of a review; Second, from the content perspective, key sentences often have stronger sentiment polarity than trivial sentences; Finally, from the representation style perspective, key sentences may contain some conclusive words or phrases, such as, *"overall"* and *"in general"*. Consequently, in the key sentence extraction algorithm, we carefully take these features into account by designing three feature functions, respectively. The overall score of any sentence is the sum of the values calculated based on the three feature functions. The corresponding algorithm is simple and can be extended to the cases with more features. In what follows, we present the three feature functions.

- *Position Feature Function:* It is observed that a sentence at the beginning or the end of a review is more likely to be a key one than those in the middle. Therefore, the position feature function should reward sentences at both ends of the document. Intuitively, the curve of a Gaussian probability density function is bell-shaped, and its negative form may fit with the characteristic of the position function. Therefore, given a sentence, s, its position feature function is defined as a negative Gaussian probability density function, namely,

$$f_1(s) = -\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}} \quad (1 \leq s \leq len), \qquad (2)$$

  where μ is the mean (location of the peak), σ is the standard deviation, len is the length (i.e., the number of sentences) of a review. In the experiments in Section IV, μ is set to $len/2$ and σ is set to 1.

- *Content Feature Function:* As key sentences should have strong and clear sentiment polarity, the content feature function is defined as

$$f_2(s) = \frac{\sum_{t \in s} opinion\_lexicon(t)}{\sum_{t \in s} |opinon\_lexicon(t)|} \qquad (3)$$

  where $opinion\_lexicon(t)$ denotes the sentiment polarity of word t in sentence s, if it is an opinion one. Specifically, if t is a positive word, $opinion\_lexicon(t) = 1$; if t is a negative word, $opinion\_lexicon(t) = -1$; otherwise, $opinion\_lexicon(t) = 0$. Here, the opinion lexicon used is acquired by the method described in Section III-A. It can be seen from Equation (3) that a sentence that contains opinion words of the same polarity has a high absolute sentiment score, whilst a sentence with mixed polarity words or no polarity words has a low absolute sentiment score.

- *Representation Style Feature Function:* This function is actually the accumulation of conclusive expressions, defined as

$$f_3(s) = \sum_{t \in s} conclusive\_expressions(t), \qquad (4)$$

  where $conclusive\_expressions(t)$ denotes if t is a conclusive snippet in a sentence. Here, a snippet is a single word or a string of words that is split by punctuations in a sentence. For example, the sentence, *"Overall, I love this book!"* has two snippets, *"overall"* and *"I love this book"*. In the experiments, all conclusive expressions are acquired by a statistical technique rather than manual collection. Two psycholinguistic and psychophysical experiments showed that in order to efficiently extract polarity of written texts, one should concentrate computational efforts on messages in the final position of the text [6]. Therefore, we extract the conclusive expressions from the last sentences of all documents in the whole corpus.

We first calculate the frequency of all snippets in all the last sentences. Those frequently used snippets in the last sentences are likely to be conclusive expressions. We then choose the snippet whose frequency is beyond a pre-defined threshold into the list of conclusive expressions.

Upon these feature functions, we select the top N sentences with the highest values, namely, $f_1(s) + f_2(s) + f_3(s)$, as key sentences. Note that if the number of sentences in a review is less than N, all the sentences are considered as key ones. In Section IV-B3, we will particularly examine the impact of the number N of key sentences on the performance of the proposed sentiment analysis method.

C. Self-Supervised Learning

As mentioned before, supervised methods usually regard sentiment polarity identification as a classification problem and use a labelled corpus to train a sentiment classifier. In order to reduce the burden of manually labelling data, we propose a Self-Supervised Learning (SSL) method for sentiment classification by combining the strengthes of both unsupervised and supervised techniques. In more detail, an unsupervised technique is first employed to label a portion of reviews from the given corpus, where the key sentences of all reviews have been clearly marked; Next, some informative sample reviews with clear sentiment polarities are selected, based on which we train a supervised classifier for multilingual sentiment classification.

In this SSL method, informative sample reviews are selected according to the clarity of their sentiment polarities using the extracted opinion lexicon. It is obvious that, for a review d, the larger the difference between the number of positive words, denoted as $T^P(d)$, and the number of negative words, denoted as $T^N(d)$, the more confident we are about its sentiment polarity. However, the difference between $T^P(d)$ and $T^N(d)$ is significantly affected by the length of the review. It is often the case that the longer the review, the larger its $T^P(d)$ or $T^N(d)$, and thus the larger the difference between its $T^P(d)$ and $T^N(d)$. Therefore, we adopt the normalized absolute difference between $T^P(d)$ or $T^N(d)$ as the the criterion function for selecting sample reviews in ordre to avoid the effect of the length of different reviews:

$$\mathrm{InfRev}(d) = \frac{|T^N(d) - T^P(d)|}{T^P(d) + T^N(d)}. \quad (5)$$

IV. Experimental Validation

In this section, we validate the proposed framework and methods via extensive experiments. Specifically, we first investigate the validity of the method for identifying the sentiment polarity of opinion words. Next, we examine the effectiveness and performance of the developed SSL method by applying it to sentiment classification tasks and comparing with four representative baselines. Finally, we study the impact of the number of key sentences on the performance of the SSL method.

A. Experimental setup

1) Datasets and Prior Knowledge: For the purpose of validation, we carried out experiments with multilingual sentiment corpora, including English, French, and German. In order to highlight the domain-specific nature of opinion words, we collect reviews not only from different languages, but also from different domains, namely, books, DVD, and music. The multilingual sentiment corpora are collected from Amazon by Prettenhofer and Stein [19][1]. The dataset for each domain contains 2000 labelled samples for training, 2000 labelled samples for testing, and a large number of unlabelled samples. Both of the training and testing datasets contain two classes, i.e., the positive and the negative. Tabel I presents the statistics of unlabelled samples from different domains and different languages.

TABLE I. The statistics of unlabelled samples of each domain in each language.

	English	French	German
Books	50,000	32,870	165,470
DVD	30,000	9,358	91,516
Music	25,220	15,940	60,392

In the KNN-based algorithm for polarity identification described in Section III-A, we tried different K values to refine the performance of the algorithm, and finally set K to 60, with which the algorithm achieves the best result.

For sentiment analysis in English, we selected two positive seed words (*good* and *recommend*) and two negative seed words (*bad* and *disappointed*) with the highest frequency from the corpora. For other languages, the seed opinion words are obtained through translating the English ones by Google Translate. Table II presents the lists of seed words of all languages, which are the prior knowledge used in the proposed approach.

TABLE II. The prior knowledge of all languages used in the experiments.

	Adverb of degree	Positive word	Negative word
English	very, highly	good, recommend	bad, disappointed
French	très, fortement	bonne, recommander	mauvaise, désappointé
German	sehr, stark	gut, empfehlen	schlecht, enttäuscht

2) Baselines for Comparison: In the experiments, for the self-supervised learning (SSL) method (see Section III-C) we employ a Naive Bayesian classifier for supervised learning and select top 30 % informatively predicted reviews for training, with which the classifier exhibits the best performance. In order to validate the effectiveness of the SSL method, we compared it with the following four representative baselines:

- Universal Opinion Lexicon (UOL), where the sentiment polarity of a review is estimated by existing universal opinion lexica. For English, the opinion lexicon is collected from an open source[2]. For French and German, the opinion lexica are translated from the English one using Google Translate.
- Naive Bayesian (NB), where the sentiment polarity of a review is estimated by a sentiment classifier, which

[1] http://www.webis.de/research/corpora/webis-cls-10

[2] http://www.keenage.com/html/e_index.html

is trained on manually labelled data by the Naive Bayesian classification method.

- Extracted Lexicon on All Sentences (ELAS), where the sentiment polarity of a review is estimated on all sentences using the extracted opinion lexicon. Specifically, the sentiment polarity is determined by counting the numbers of its positive and negative words. The opinion lexicon being used is extracted from unlabelled data.
- Extracted Lexicon on Key Sentences (ELKS), where the sentiment polarity of a review is estimated only on key sentences using the extracted opinion lexicon.

Among the above four baselines, UOL, ELAS, and ELKS are unsupervised methods, while NB is a typical supervised one.

B. Experimental results

1) Effectiveness of the Voting Rule: In this experiment, we validated the effectiveness of the voting rule for identifying the sentiment polarity of opinion words by evaluating the accuracy of the extracted English opinion lexicon. For the purpose of evaluation, we manually labelled the sentiment polarity of all extracted opinion words, and used them as the reference. In order to highlight the domain-specific nature of an opinion lexicon, we labelled opinion words according to their domain characteristics. Considering the reliability of this labeling process, we invited three annotators to label the semantic orientation (positive or negative) of the same opinion lexicon. For each opinion word, if there is disagreement about the annotated results, the minority is subject to the majority.

TABLE III. THE ACCURACY OF THE POLARITY ASSIGNMENT OF ENGLISH OPINION WORDS.

English	Before Voting	After Voting
Books	0.6931	0.8053
DVD	0.7263	0.7835
Music	0.7512	0.7708
Average	0.7235	0.7865

Table III presents the accuracy of polarity assignment of opinion words before and after voting, respectively. It can be seen in Table II that starting from four frequently used seed words, the proposed method (before voting) for polarity assignment of opinion words achieves acceptable results. When the voting rule is adopted, the accuracy of the English opinion lexicon is further improved by 6.3% on average, which verifies the effectiveness of the voting rule. Note that in the voting process, only in Case 2 the polarity of an opinion word in a domain may be changed, while in other cases it is remained. In more detail, for general (i.e., domain-independent) opinion words their sentiment polarity in the domain at hand is changed to be consistent with that in other two domains, while for domain-specific opinion words their sentiment polarity in the domain at hand remains unchanged. Moreover, we have experimentally found that most opinion words are general and only a small number of opinion words are domain-specific. Therefore, we can conclude that the above improvement mainly benefits from the polarity correction of general opinion words.

2) Sentiment Classification: In this experiment we applied the SSL method to practical sentiment classification in order to evaluate its performance. Specifically, we first studied if the extracted opinion lexicon is effective as compared to a universal opinion lexicon. Next, we examined if the extracted key sentences are helpful for sentiment classification by comparing to the polarity assignment on full texts. Finally, we investigated the effectiveness of the SSL method through comparison with the supervised and unsupervised baselines. Tables IV-VI present the experimental results of sentiment classification in different domains and languages by the SSL method and the four baselines. In these tables, the accuracy of sentiment classification is listed for each method and each domain. In particular, the bottom row of each table presents the accuracy of each method averaged on different domains.

To study the effectiveness of the extracted opinion lexicon in practical sentiment classification, we compared ELAS and UOL. In Table IV, we can see that ELAS consistently outperforms UOL, suggesting that the extracted opinion lexicon is effective. In other words, although the universal opinion lexicon is relatively comprehensive for English opinion words, the lexicon extracted in this paper performs as good as or even better than it. Actually, it has been noted that the polarity of an opinion word may vary from domain to domain. Therefore, extracting an opinion lexicon directly from the given corpus can achieve more satisfactory results. In Tables V and VI, we can see that for both French and German, ELAS outperforms UOL observably, which indicates that the extracted opinion lexica in different languages are all effective.

In order to measure the usefulness of the extracted key sentences in sentiment classification, we compared ELKS and ELAS. From Table IV, for English, we can observe that ELKS outperforms ELAS by performance improvement of 2.03% on average. For French and German, Tables V and VI show that ELKS performs better than ELAS by average improvement of 4.11% and 4.17%, respectively. These observations suggest that key sentences are more confident and concise for identifying the overall sentiment polarity of reviews, while the full text is sentimentally ambiguous due to the interference of detailed statements whose sentiment polarity may be different from that of the whole text. Particularly, the above performance improvement demonstrates two advantages of the key sentence extraction method. First, the concept of key sentences is reasonable and key sentences are useful for dealing with conflictive sentiments; Second, the key sentence extraction method is effective and language-independent.

Finally, we compared SSL with NB and ELKS to investigate its performance and merits. In Table IV, we can see that by combining supervised and unsupervised techniques together, SSL performs better than ELKS by performance improvement of 2.84%. Compared with the supervised baseline, NB, the performance of SSL is also acceptable. From Tables V and VI, it can be noted that the performance of sentiment analysis in French and German is improved by 2.35% and 5.18%, respectively, as compared to ELKS. The improvement for German is more significant than that for French. This is because the scale of German corpora is larger than that of French. For sentiment classification in both English and German, the average performance of SSL is inferior to that of NB. This is reasonable and intuitive, as SSL is completely unsupervised

TABLE IV. THE ACCURACY OF SENTIMENT ANALYSIS IN ENGLISH.

English	UOL	NB	ELAS	ELKS	SSL
Books	0.6985	0.7790	0.7163	0.7370	0.7679
DVD	0.6966	0.7665	0.7180	0.7315	0.7428
Music	0.6779	0.7785	0.6897	0.7163	0.7595
Average	0.6910	0.7747	0.7080	0.7283	0.7567

TABLE V. THE ACCURACY OF SENTIMENT ANALYSIS IN FRENCH.

French	UOL	NB	ELAS	ELKS	SSL
Books	0.6259	0.8315	0.6849	0.7112	0.7463
DVD	0.6646	0.8145	0.6817	0.7238	0.7415
Music	0.6824	0.8355	0.6698	0.7247	0.7423
Average	0.6576	0.8272	0.6788	0.7199	0.7434

TABLE VI. THE ACCURACY OF SENTIMENT ANALYSIS IN GERMAN.

German	UOL	NB	ELAS	ELKS	SSL
Books	0.6746	0.8070	0.6832	0.7423	0.7800
DVD	0.6482	0.8010	0.6971	0.7257	0.7880
Music	0.6537	0.8120	0.6783	0.7158	0.7710
Average	0.6588	0.8067	0.6862	0.7279	0.7797

and does not use any labelled data. But, they are generally comparable. As far as sentiment classification in French, the results are barely satisfactory, because the unlabelled data in French is the least among the three languages especially for the DVD domain. Although we knew nothing about French and German, we successfully conduct classification through a few seed words and some unlabelled data.

In summary, from the above comparison we noticed that SSL performs better than all the three unsupervised baselines, UOL, ELAS and ELKS, and achieves comparable results with the supervised baseline NB, which well demonstrates its effectiveness and merits as a useful and cost-efficient tool in practical sentiment analysis.

3) Impact of the Number of Key Sentences: This experiment intends to investigate the impact of the number of extracted key sentences in the SSL method on sentiment classification. Specifically, we examined the accuracy of sentiment classification using the SSL method in different domains and languages and with different numbers of key sentences. Figure 2 presents the experimental results, where the accuracy curves of sentiment classification averaged on all domains are also depicted. From this figure, we can find that when the number of extracted key sentences is in the region $[1, 3]$, the SSL method exhibits the best performance. When the number of extracted key sentences further increases beyond this region, its performance monotonically decreases. That is to say, too many key sentences cannot lead to better performance. Actually, more sentences mean more noises that deteriorates the identification of the overall sentiment of a review. Based on this finding, in the above experiments the number of key sentences was set to 2.

V. CONCLUSIONS

In this paper, we have proposed a novel multilingual sentiment analysis framework, which contains three specific methods, namely, an opinion lexicon extraction method for extracting opinion words and identifying their sentiment polarities, a key sentence extraction method for extracting key sentences in given texts, and a Self-Supervised Learning (SSL) method, which combines supervised and unsupervised techniques together for training sentiment classifiers. The merits of the proposed framework lie in that no manually labelled corpora is needed and all extracted information is domain-dependent. Finally, extensive experiments on multilingual datasets in different domains have well demonstrated the effectiveness of the proposed framework and methods. Specifically, we have found through comparison with unsupervised and supervised baselines that (1) the proposed SSL method performs better than all the unsupervised baselines; (2) the SSL method is comparable to the supervised baseline based on the Naive Bayesian algorithm. However, as the supervised baseline requires manually labelled corpora, the SSL method is more cost-efficient.

ACKNOWLEDGMENT

This work is supported by National Grand Fundamental Research 973 Program of China (No. 2012CB316303, 2013CB329602) and National Natural Science Foundation of China (No. 61232010, 61100175, 61173008).

REFERENCES

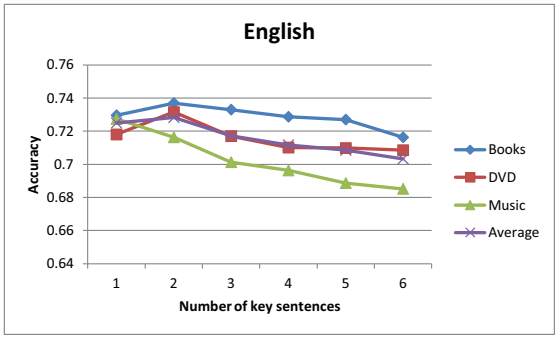
[1] S.-A. Bahrainian and A. Dengel. Sentiment analysis and summarization of twitter data. In *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pages 227–234. IEEE, 2013.

[2] A. Balahur and M. Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, 2014.

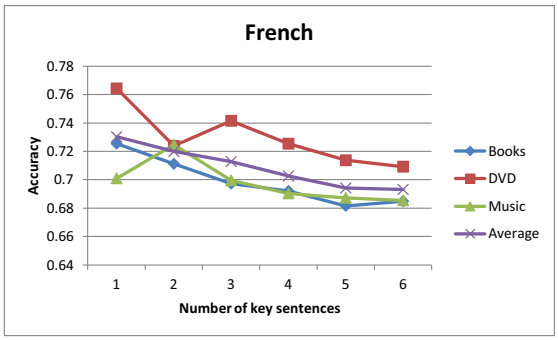
[3] C. Banea, R. Mihalcea, and J. Wiebe. Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 28–36, 2010.

[4] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135, 2008.

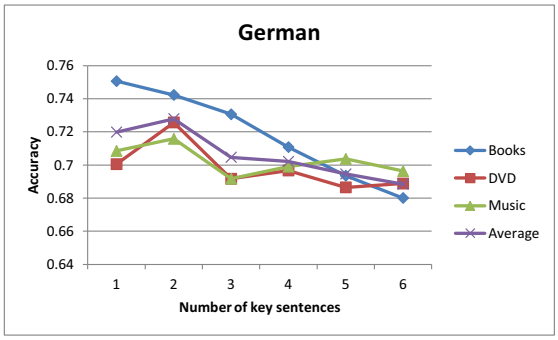
[5] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.



(a) English



(b) French



(c) German

Fig. 2. The accuracy of sentiment classification using the proposed SSL method in different domains and different languages and with different number of key sentences.

[6] I. Becker and V. Aharonson. Last but definitely not least: on the role of the last sentence in automatic polarity-classification. In *Proceedings of the acL 2010 conference Short Papers*, pages 331–335, 2010.

[7] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI tech report SS-04-07)*, 2004.

[8] J. Boyd-Graber and P. Resnik. Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 45–55, 2010.

[9] A. Esuli and F. Sebastiani. Pageranking wordnet synsets: An application to opinion mining. 45(1):424, 2007.

[10] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[11] M. Gamon. 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*, page 841. Association for Computational Linguistics, 2004.

[12] D. Gao, F. Wei, W. Li, X. Liu, and M. Zhou. Cotraining based bilingual sentiment lexicon learning. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[13] A. Hassan and D. Radev. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403, 2010.

[14] M. Koppel and J. Schler. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109, 2006.

[15] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.

[16] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 976, 2007.

[17] S. O. Orimaye, S. M. Alhashmi, and E.-G. Siew. Can predicate-argument structures be used for contextual opinion retrieval from blogs? *World Wide Web*, 16(5-6):763–791, 2013.

[18] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[19] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, 2010.

[20] G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.

[21] L. Qiu, W. Zhang, C. Hu, and K. Zhao. Selc: a self-supervised model for sentiment classification. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 929–936. ACM, 2009.

[22] L. Shi, R. Mihalcea, and M. Tian. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, 2010.

[23] V. K. Singh, M. Mukherjee, and G. K. Mehta. Combining collaborative filtering and sentiment classification for improved movie recommendations. In *Multi-disciplinary Trends in Artificial Intelligence*, pages 38–50. Springer, 2011.

[24] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424, 2002.

[25] X. Wan. Co-training for cross-lingual sentiment classification. 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: Volume 1-Volume 1*, pages 235–243, 2009.

[26] B. Wei and C. Pal. Cross lingual adaptation: an experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 258–262, 2010.

[27] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

[28] J. M. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the National Conference on Artificial Intelligence*, pages 735–741, 2000.

[29] A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1046–1056, 2010.

[30] T. Zagibalov and J. Carroll. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1073–1080. Association for Computational Linguistics, 2008.