International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Issue 11, November 2015
www.ijates.com
ijates
ISSN 2348 - 7550

DUAL SENTIMENT ANALYSIS WITH TERM COUNTING AND JOINT PREDICTION

Sharmila Zope

[1]*Computer Science, GESCOE Nashik, (India)*

ABSTRACT

Dual sentimental analysis is an important current research area. The sentiment found within comments, feedback or critiques provide useful indicators form any different purposes. These sentiments can be categorized either into two categories: positive or negative; or into an n -point scale, e.g. Very good, good, satisfactory, bad, very bad. In this respect, a sentimental analysis task can be interpreted as a classification task where each category represents a sentiment. We address unsupervised sentiment classification by creating reversed reviews for each testing example, integrate the dual prediction rule into a term counting methods and make a joint prediction based on two sides of one review. Dual Sentiment analysis provides companies with a means to estimate the extent of product acceptance and to determine strategies to improve product quality. It also facilitates policy makers or politicians to Analyze public sentiments with respect to policies, public services or political issues.

Keywords: Natural Language Processing, N-Point Scale, Opinion Mining, Public Service, Sentimental Analysis.

I. INTRODUCTION

In recent years, with the growing volume of online reviews available on the Internet, sentiment analysis and opinion mining, as a special text mining task for determining the subjective attitude (i.e., sentiment) expressed by the text, is becoming a hotspot in the field of data mining and natural language processing [5], [12], [13], [21], [22].

1.1 BOW Model

Sentiment classification is a basic task in sentiment analysis, with its aim to classify the sentiment (e.g., positive or negative) of a given text. The general practice in sentiment classification follows the techniques in traditional topic-based text classification, where the bag-of-words (BOW) model is typically used for text representation. In the BOW model, a review text is represented by a vector of independent words.

Although the BOW model is very simple and quite efficient in topic-based text classification, it is actually not very suitable for sentiment classification because it disrupts the word order, breaks the syntactic structures, and discards some semantic information. Consequently, a large number of researches in sentiment analysis aimed to enhance BOW by incorporating linguistic knowledge [6], [9], [14], [19]. However, due to the fundamental deficiencies in BOW, most of these efforts showed very slight effects in improving the classification accuracy. One of the most well-known difficulties is the polarity shift problem.

1.2 Polarity Shift

Polarity shift is a kind of linguistic phenomenon which can reverse the sentiment polarity of the text. Negation is the most important type of polarity shift. For example, by adding a negation word “don’t” to a positive text “I like this book” in front of the word “like”, the sentiment of the text will be reversed from positive to negative. However, the two sentiment-opposite texts are considered to be very similar by the BOW representation. This is the main reason why standard machine learning algorithms often fail under the circumstance of polarity shift.

Several approaches have been proposed in the literature to address the polarity shift problem [7], [9], [10], [11], [15]. However, most of them required either complex linguistic knowledge or extra human annotations. Such high-level dependency on external resources makes the systems difficult to be widely used in practice.

1.3 Dual Sentiment Analysis

Effective way to handle this is a simple yet efficient model, called dual sentiment analysis (DSA), to address the polarity shift problem in sentiment classification. By using the property that sentiment classification has two opposite class labels (i.e., positive and negative), initially a data expansion technique used for creating sentiment-reversed reviews. The original and reversed reviews are constructed in a one-to-one correspondence.

In DSA a dual training (DT) algorithm and a dual prediction (DP) algorithm respectively, to make use of the original and reversed samples in pairs for training a statistical classifier and make predictions. In DT, the classifier is learnt by maximizing a combination of likelihoods of the original and reversed training data set. In DP, predictions are made by considering two sides of one review. This method of analysis is used for supervised sentiment analysis. Also DSA framework is implemented for 3-class (positive-negative-neutral) sentiment classification, by taking the neutral reviews into consideration in both dual training and dual prediction [1].

A data mining approach to sentiment analysis translates an unstructured text problem to one that makes predictions on structured, quantitative data. The approach borrows several techniques from computational linguistics and information retrieval communities.

In this paper we propose a simple yet effective concept of term counting using subjective summarization.We address unsupervised sentiment classification by creating reversed reviews for each testing example, integrate the dual prediction rule into a term counting methods and make a joint prediction based on two sides of one review.

The organization of this paper is as follows. Section 2 reviews the related work. In Section 3, we present the existing data mining technique and natural language processing. In Section 4, we introduce the concept of subjective summarization in detail. Section 5 presents conclusions and outlines directions for the future work.

II. RELATED WORK

First summarize the work of sentiment analysis and polarity shift, and then review the technique of data expansion.

2.1 Sentiment Analysis

According to the levels of granularity, tasks in sentiment analysis can be divided into four categorizations: document-level, sentence-level, phrase-level, and aspect-level sentiment analysis.

For document- and sentence-level sentiment classification, there are two main types of methods in the literature: term-counting and machine learning methods. In term-counting methods, the overall orientation of a text is obtained by summing up the orientation scores of content words in the text, based on manually-collected or external lexical resources [17], [18]. In machine learning methods, sentiment classification is regarded as a statistical classification problem, where a text is represented by a bag-of-words; then, the supervised machine learning algorithms are applied as classifier [15]. Accordingly the way to handle polarity shift also differs in the two types of methods.

2.2 Polarity Shift

The term-counting methods can be easily modified to include polarity shift. One common way is to directly reverse the sentiment of polarity-shifted words, and then sum up the sentiment score word by word [8], [9], [16]. Compared with term counting methods, the machine learning methods are more widely discussed in the sentiment classification literatures.

There were also some attempts to model polarity shift by using more linguistic features or lexical resources. For example, Na et al. [14] proposed to model negation by looking for specific part-of-speech tag patterns. Kennedy and Inkpen [9] proposed to use syntactic parsing to capture three types of valence shifters (negative, intensifiers, and diminishes). Their results showed that handling polarity shift improves the performance of term-counting systems significantly, but the improvements upon the baselines of machine learning systems are very slight (less than 1 percent).

There were still some approaches that addressed polarity shift without complex linguistic analysis and extra annotations. For example, Li and Huang [10] pro-posed a method first to classify each sentence in a text into a polarity unshifted part and a polarity-shifted part according to certain rules, then to represent them as two bags-of-words for sentiment classification. Li et al. [11] further proposed a method to separate the shifted and unshifted text based on training a binary detector. Classification models are then trained based on each of the two parts.

2.3 Techniques of Sentiment Analysis

In Selecting Attributes for Sentiment Classification Using Feature Relation Networks (FRN) is a rule-based multivariate text feature selection method that considers semantic information and also leverages the syntactic relationship between n-gram features. FRN was able to select attributes resulting in significantly better classification accuracy irrespective of the feature subset sizes. FRN's use of syntactic relation and semantic information regarding n-grams enabled it to achieve improved results over various univariate, multivariate and hybrid feature selection methods. But other measurements, such as occurrence frequency and various positional/ distributional features were not considered in a multidimensional FRN. It needed the development of hybrid feature selection methods that incorporate FRN in conjunction with other multivariate selection techniques by Abbasi et al. [5].

C. Lin, et al. proposes a model for weakly supervised sentiment analysis called Joint Sentiment-Topic (JST) which is probabilistic modeling framework for detection of sentiment from Text and Reverse-JST based on latent Dirichlet allocation (LDA), which detects sentiment and topic simultaneously from text. Weakly supervised nature of JST makes it highly portable to other domains and JST model achieved either better or

ijates
ISSN 2348 - 7550

comparable performance compared to existing semi-supervised approaches. But JST model encodes the assumption that there is approximately a single sentiment for the entire document. The incremental learning of the JST parameters is not considered when facing with new data. Supervised information was not incorporated into JST model learning, such as some known topic knowledge for certain product reviews [12].

The concept of Dual Training and Dual Prediction for Polarity Classification is introduced by R. Xia et al. in this the focus was on the polarity shift problem, and given approach, called dual training and dual prediction (DTDP), to address it. The basic idea of DTDP is to first generate artificial samples that are polarity-opposite to the original samples by polarity reversion, and then leverage both the original and opposite samples for (dual) training and (dual) prediction. The limitation of current work is that the tuning of parameters in DTDP (such as a and t) is not well discussed [20].

Identifying features in opinion mining is the most basic process which can be perform via intrinsic and extrinsic domain relevance proposed by Z. Hai et al. where intrinsic and extrinsic domain relevance (IEDR) approach utilizes the fact that word distribution characteristics vary across different types of corpora. Here in particular domain-specific versus domain-independent, to derive powerful hints that help discriminate valid features from the invalid ones. But the opinion feature extraction performance of IEDR is poor on longer and complicated reviews, also on large number of noisy domain-irrelevant users. Neutral opinions are excluded. Currently only positive and negative opinions are considered [22].

Recently the technique used for sentiment analysis is Dual Sentiment Analysis (DSA) considering two sides of one review proposed by R. Xia et al. in which the DSA algorithm was strengthen by adding a selective data expansion procedure. Extend the DSA framework from sentiment polarity classification to positive-negative-neutral sentiment classification. Construct a pseudo-antonym dictionary that could remove DSA’s dependency on an external antonym dictionary. But this only focuses on supervised approaches to sentiment classification which often fail to produce satisfactory performance when shifting to other domains. Unsupervised, semi-supervised, class-imbalanced sentiment classifications are proposed for future implementation work [1].

III. EXISTING SYSTEM

In this section contain brief overview on Data Mining Approaches and Natural language processing concept along with its benefits and limitations.

3.1 Data Mining Approach

A data mining approach to sentiment analysis translates an unstructured text problem to one that makes predictions on structured, quantitative data. The approach borrows several techniques from computational linguistics and information retrieval communities to represent the text numerically, and then applies traditional data mining techniques to this numeric representation. In the end, a target variable is identified and a pattern is discovered from the training data for predicting sentiment polarity. This pattern can then be used to predict new observations.

The first step in creating the numeric representation is to convert the entire training collection into a document-by-term frequency matrix. Each document is parsed into individual terms, or term/part-of-speech pairs. Then the set of all terms becomes the variables on the data set so that documents are now represented as vectors of length equal to the number of distinct terms in the collection.

These vectors are very sparse, containing mostly zeroes – because any one document contains a very small percentage of the terms in the collection. Once the documents are represented as vectors, the frequencies in each cell can be weighted with a function that takes into account the distribution of the term across the collection and relative to the levels of the target variable.

After these document vectors are formed, a dimension reduction technique – such as the singular value decomposition is typically used to represent each document in a reduced-dimensional space of maybe 50 to 100 variables, where each variable is a linear combination of the weighted terms that originally represented each document.

Finally, these reduced-dimensional vectors, together with the sentiment variable, can be supplied to a predictive model. The model will attempt to learn from the training data by utilizing patterns in the reduced-dimensional vector. This predictive model will then create a function that will predict the sentiment for any document [13].

3.2 Benefits of Data Mining Approach

The data mining approach is appealing because it is based on learning patterns that are useful for making automated, efficient predictions. The algorithms are capable of discovering unimagined and complicated patterns that would be beyond what a human could anticipate.

3.3 Limitations of Data Mining Approach

The vector-based representation of a document, which is required for data mining techniques, does not maintain information that is potentially important to sentiment classification. For example, the vector representation does not capture when terms are close to one another in the document, if one term precedes another or any other contextual cues. The order of terms in a phrase can significantly affect meaning. Consider the phrases

"… night for a great movie"

and

"… great night for a movie"

These two phrases convey two different meanings; yet in a vector representation, the phrases have an identical representation.

Also forming the training and validation is an essential component of learning a predictive model, but it can be very time-consuming and challenging. A rating needs to be provided for every document, and if there are attributes of documents that you wish to use to measure sentiment, you will need to provide a rating for each of these as well [2].

3.4 Natural Language Processing

Natural language processing (NLP) is a field of artificial intelligence that deals with automatically extracting meaning from natural language text. As discussed in the introduction of this paper, it's very challenging to get machines to understand text at the same levels as humans. Doing this with the specific goal of extracting sentiment is even more challenging. For example, consider the text snippet below

"... with that out of the way, let me say this – this film is bad. This film is really, really bad. Yet somehow, it is strangely enjoyable. ..."

If interpreted by a human, the above text would imply a positive sentiment from the author toward the movie. However, it can be very challenging to get the same output from a computer because of the dense presence of the strongly negative words.

The rule-based NLP methods use certain entities and syntactic patterns in the text to understand its meaning. SAS Sentiment Analysis provides all the tools needed for this kind of disambiguation. You can use a combination of language dictionaries, linguistic constructs like parts of speech, and noun phrases along with a range of operators [3].

The operators fall into a few different categories as shown below

- Boolean operators. Used to include or exclude different entities (e.g., AND, OR, NOT).
- Frequency operators. Used to measure the specified number of occurrences of certain entities, (e.g., MIN, MINOC, and MAXOC).
- Context operators. Used to measure the context within which certain entities occur in the text (e.g., DIST, START, END, SENT, PARA).
- Sequence operators. Used to look for the entities in a specific sequence (e.g. ORD, ORDDIST).

3.5 Benefits of NLP Approach

The effective part of rule-based methods is the amount of control they give rule developers over how the analysis will be performed. Developers can use their knowledge of the domain and the language within it to develop rules that have high precision.

Unlike statistical analysis, the results of rule-based analysis are easily interpretable. This is very important for real-life applications where the analysts need to know exactly why a document or an attribute within a document was tagged as positive or negative. In other words, analysts need to know exactly what sentences, keywords or context within the document triggered the positive or negative sentiment.

Rule-based methods are completely unsupervised; that is, they do not require any training data. This is a big advantage in real-life applications where training data is scarce. The non-availability of training data is more pronounced when it comes to granular sentiment analysis (sentiment derived at the objects and attributes level) [4].

3.6 Limitations of NLP Approach

The limitations of rule-based method are that they require a lot of human involvement in developing the rules. These methods completely rely on the domain knowledge of rule developers. It might take a few weeks to come up with a strong rule-based model for a new domain. However, once you have a strong rule-based model for a domain, you can reuse that model with some minor modifications for different applications within the domain.

The importance of validation data is often underestimated while developing these models. The rules being written must be generic enough so that they are capable of handling all possible cases. Inexperienced rule developers tend to over-fit their rules to the sample data they are working with. Such rules might not work well when tested on

ijates
ISSN 2348 - 7550

different data sets. So, rule developers must make sure they validate the rules on different data sets before considering a model ready to deploy.

IV. PROPOSE SYSTEM

4.1 Dual Sentiment Analysis

In this chapter, let's study DSA framework in detail. Fig. 1 illustrates the process of a DSA algorithm [1]. It contains two main stages: 1) dual training and 2) dual prediction.

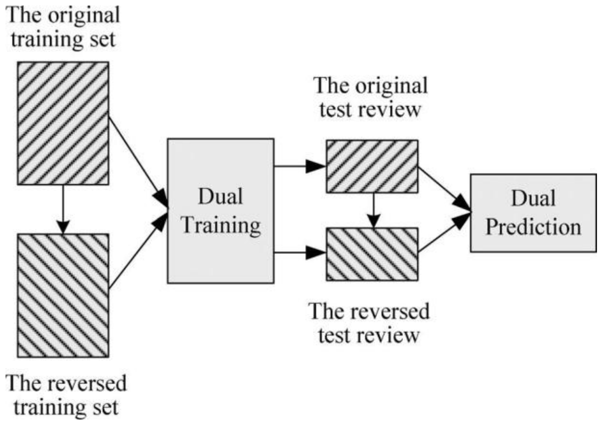

Fig. 4.1.1 The Process of Dual Sentiment Analysis

4.2 Dual Training

In the training stage, all of the original training samples are reversed to their opposites and refer to them as "original training set" and "reversed training set" respectively. In data expansion technique, there is a one-to-one correspondence between the original and reversed reviews. The classifier is trained by maximizing a combination of the likelihoods of the original and reversed training samples. This process is called dual training.

Table 4.2.1 An Example of Creating Reversed Training Reviews

	Review Text	Class
Original review	I don't like this book. It is boring.	Negative
Reversed review	I like this book. It is interesting.	Positive

Now let us use the example in Table 1 to explain the effectiveness of dual training in addressing the polarity shift problem. We assume "I don't like this book. It is boring. (class label: negative)" is the original training review. Hence, "I like this book. It is interesting. (class label: positive)" is reversed training review. Due to negation, the word "like" is (incorrectly) associated with the negative label in the original training sample. Hence, its weight will be added by a negative score in maximum likelihood estimation. There-fore, the weight of "like" will be falsely updated. While in DT, due to the removal of negation in the reversed review, "like" is (correctly) associated with the positive label, and its weight will be added by a positive score. Hence, the learning errors caused by negation can be partly compensated in the dual training process [1].

4.3 Dual Prediction

In the prediction stage, for each test sample, reversed test sample generated. Note that aim is not to predict the class of reverse sample. But instead, we use reverse sample to assist the prediction of x. This process is called dual prediction. Also consider posterior probabilities of each test sample and reverse sample respectively. A weighted combination of two component predictions is used as the dual prediction score. Hence the prediction with a higher posterior probability will be chosen as the final prediction.

Consider the example in Table 1 again to explain why dual prediction works in addressing the polarity shift problem. This time assume "I don't like this book. It is boring" is an original test review, and "I like this book. It is interesting" is the reversed test review. In traditional BOW, "like" will contribute a high positive score in predicting overall orientation of the test sample, despite of the negation structure "don't like". Hence, it is very likely that the original test review will be misclassified as Positive. While in DP, due to the removal of negation in the reversed review, "like" this time the plays a positive role. Therefore, the probability that the reversed review being classified into Positive must be high. In DP, a weighted combination of two component predictions is used as the dual prediction output. In this manner, the prediction error of the original test sample can also be compensated by the prediction of the reversed test sample. Apparently, this can reduce some prediction errors caused by polarity shift [1].

4.4 DSA With Selective Data Expansion

In the dual training procedure in which all of the training reviews are used in data expansion. However, in many cases, not all of the reviews have such distinct sentiment polarity. So here uses a selective data expansion procedure to select a part of training reviews for data expansion. Consequently, introduce a sentiment degree metric for selecting the most sentiment-distinct training reviews for data expansion. A threshold will be set to select a percentage of original reviews with higher sentiment degree for data reversion and dual training as a criterion for data expansion [1].

4.5 Our Approach to Sentiment Analysis

4.5.1 Algorithm

Dual Sentiment analysis for unsupervised sentiment classification using subjective summarisation considering reversed reviews for each testing example, integrate the dual prediction rule into a term counting methods and makes a joint prediction based on two sides of one review.

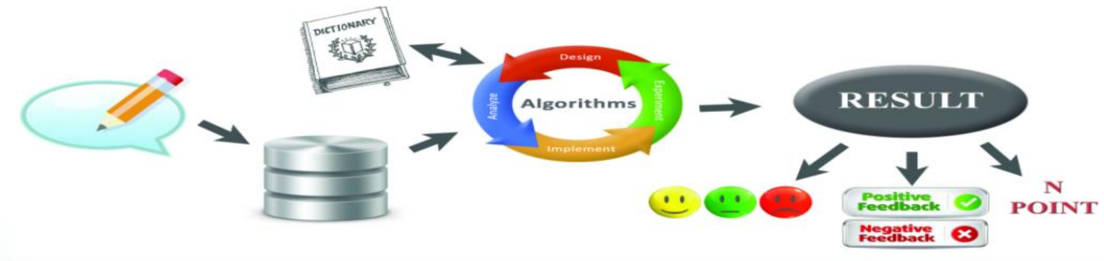

Fig.4.5.1.1 Our Propose Sentiment Analysis Process.

Steps of our propose Sentiment Analysis process

1. Fetch comments provided by the user for processing.

2. Using dictionary approaches to determine the product, user is speaking about.
3. Create dictionaries for weak and strong sentiment related patterns.
4. Apply strong negative sentiment patterns to the input in relative to the product.
5. If not found try searching for weak negative patterns.
6. Search for positive sentiment patterns in the comments with relative to the product.
7. If positive sentiment pattern is found make sure that it does not have negative pattern preceding it. If found just flip the polarity of the sentiment to negative.

Eg

i. Apple iphone 4 is not made me happy at all.
   - Here we need to flip the polarity, as positive pattern is preceded by a negative one.

ii. Samsung Ace is indeed a **smart** phone.
   - Here the term 'Smart' indicates that the sentiment is positive for Samsung Ace.

We are systematically going to evaluate our approach on two tasks including term counting method and positive-negative-neutral sentiment classification across four sentiment datasets, classification algorithm Dual Sentiment Classification using Subjective Summarization.

For this purpose we are working on English datasets contain product reviews taken from Amazon.com including four different domains: Book, DVD, Electronics and Kitchen. Each of the four datasets contains 1,000 positive and 1,000 negative reviews these reviews are used as test cases for our algorithm [23].

4.5.2 Features

- Analyse user reviews on different products in market.
- Helps any Organization for improving their products.
- Analyse thousands of feedbacks and provide generalized opinion for the product.
- Uses data mining concepts (Text Analytics) for sentimental analysis.

V. CONCLUSION

In real-life applications, to provide a completely automated solution is nowhere in sight. However, it is possible to devise effective semi automated solutions.

The key is to fully understand the whole range of issues and pitfalls, cleverly manage them, and determine what portions can be done automatically and what portions need human assistance. In the continuum between the fully manual solution and fully automated solution, we can push more and more toward automation.

REFERENCES

[1] R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi and T. Li, "Dual sentiment analysis: Considering two sides of one review", in IEEE Trans. Knowl. Data Eng., vol. 27, no. 8, pp. 2120-2133, Aug. 2015.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval2 (1-2), pp. 1–135, 2008.

[3] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning Subjective Language," Computational Linguistics, vol. 30, pp. 277–308, September 2004.

[4] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177, 2004.

[5] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol. 23, no. 3, pp. 447–462, Mar. 2011.

[6] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analy-sis," in Proc. Int. Conf. Comput. Linguistics, 2004, pp. 841–847.

[7] D. Ikeda, H. Takamura, L. Ratinov, and M. Okumura, "Learning to shift the polarity of words for sentiment classification," in Proc. Int. Joint Conf. Natural Language Process., 2008.

[8] S. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. Int. Conf. Comput. Linguistic, 2004, pp. 1367–1373.

[9] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Comput. Intell., vol. 22, pp 110–125, 2006.

[10] S. Li and C. Huang, "Sentiment classification considering negation and contrast transition," in Proc. Pacific Asia Conf. Lang., Inf. Com-put., 2009, pp. 307–316.

[11] S. Li, S. Lee, Y. Chen, C. Huang and G. Zhou, "Sentiment classifi-cation and polarity shifting," in Proc. Int. Conf. Comput. Linguistics, 2010, pp. 635–643.

[12] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," IEEE Trans. Knowl. Data Eng., vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[13] B. Liu, "Sentiment analysis and opinion mining," in Synthesis Lec-tures on Human Language Technologies, vol. 5, no. 1. San Rafael, CA, USA: Morgan & Claypool, 2012, pp. 1–165.

[14] J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, "Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews," in Proc. Conf. Int. Soc. Knowl. Org., 2004, pp.49–54.

[15] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in Proc. Conf. Empirical Methods Natural Language Process., 2002, pp. 79–86.

[16] L. Polanyi and A. Zaenen, "Contextual lexical valence shifters," in Proc. AAAI Spring Symp. Exploring Attitude Affect Text, 2004, pp 1–10.

[17] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proc. Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 417–424.

[18] P. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," ACM Trans. Inf. Syst., vol. 21, no. 4, pp. 315–346, 2003.

[19] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classifica-tion algorithms for sentiment classification," Inf. Sci., vol. 181, no. 6, pp. 1138–1152, 2011.

[20] R. Xia, T. Wang, X. Hu, S. Li, and C. Zong, "Dual training and dual prediction for polarity classification," in Proc. Annu. Meeting Assoc. Comput. Linguistics, 2013, pp. 521–525.

[21] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: A case study in the movie domain," IEEE Trans. Knowl. Data Eng., vol. 24, no. 4, pp. 720–734, Apr. 2012.

[22] Z. Hai, K. Chang, J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 447–462, Mar. 2014.

[23] http://www.cs.jhu.edu/ mdredze/datasets/sentiment/