entropy

MDPI

Article

CoFea: A Novel Approach to Spam Review Identification Based on Entropy and Co-Training

Wen Zhang [1,*], Chaoqi Bu [1], Taketoshi Yoshida [2] and Siguang Zhang [3]

1 Center for Reseach on Big Data Sciences, Beijing University of Chemical Technology, Beijing 100029, China; buchaoqi@mail.buct.edu.cn
2 School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Nomi City, Ishikawa 923-1292, Japan; yoshida@jaist.ac.jp
3 Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China; zhangsiguang@casipm.ac.cn
* Correspondence: zh6angwen@mail.buct.edu.cn; Tel./Fax: +86-10-6419-7040

Academic Editor: Kevin H. Knuth
Received: 20 October 2016; Accepted: 28 November 2016; Published: 30 November 2016

Abstract: With the rapid development of electronic commerce, spam reviews are rapidly growing on the Internet to manipulate online customers' opinions on goods being sold. This paper proposes a novel approach, called CoFea (Co-training by Features), to identify spam reviews, based on entropy and the co-training algorithm. After sorting all lexical terms of reviews by entropy, we produce two views on the reviews by dividing the lexical terms into two subsets. One subset contains odd-numbered terms and the other contains even-numbered terms. Using SVM (support vector machine) as the base classifier, we further propose two strategies, CoFea-T and CoFea-S, embedded with the CoFea approach. The CoFea-T strategy uses all terms in the subsets for spam review identification by SVM. The CoFea-S strategy uses a predefined number of terms with small entropy for spam review identification by SVM. The experiment results show that the CoFea-T strategy produces better accuracy than the CoFea-S strategy, while the CoFea-S strategy saves more computing time than the CoFea-T strategy with acceptable accuracy in spam review identification.

Keywords: spam review; co-training; CoFea

1. Introduction

User-created content (UCC, also called user-generated content) are appearing in the era of Web 2.0, encouraging individuals to publish their opinions or reviews on different types of subjects such as e-commerce [1], tourism industry [2], software engineering [3], and so forth. People can easily publish their opinions, comments, or views on online shopping platforms, blogs, and forums. For instance, users' comments are a must-have section on a shopping website. Potential customers typically view other people's reviews before making consumption decisions, and they may be strongly influenced by existing users' opinions. Negative reviews can cause poor reputation to a product or even a brand. On the contrary, positive reviews can impart a noble image to a product and bring about advantages over other products in the market competition.

Some companies purposely post positive reviews to promote their own products in order to make money. Even worse, they even hire an "Online Water Army" [4] to post negative comments to defame competitors' products. Since fake reviews can be generated at little cost, a large number of deceptive reviews have arisen with the development of Internet.

In order to get rid of fabricated reviews, researchers propose several approaches to differentiate false reviews from truthful ones. Jindal and Liu [5] point out that spam reviews are widespread after they analyzed 5.8 million reviews and 2.14 million reviewers from Amazon.com. They categorize spam

reviews into three types: untruthful reviews, only-one-brand ones, and irrelevant ones. They also named such fake opinions as "opinion spam", which is made with vicious incentives and purposes.

According to Jindal and Liu's work, it is not hard to manually identify only-one-brand reviews and irrelevant reviews. However, it is very difficult for human beings to differentiate purposely fabricated reviews from truthful ones; in other words, it is very difficult to identify deceptive reviews which are deliberately and elaborately crafted by spammers from truthful ones [5]. Thus, several machine learning techniques are proposed to fulfill the classification task of automatic deceptive review identification. Due to the lack of the golden set of spam reviews, Jindal and Liu assume duplicate reviews as spam information for research. However, Pang and Lee deem it inappropriate to merely regard duplicate reviews as spam reviews because the connotation of the spam review is much more than a duplicate review [6]. In this paper, we define spam reviews as those reviews that are purposely made to mislead ordinary consumers with fake information on the Internet.

Ott et al. [7] combine the n-gram feature and psychological feature to identify spam reviews. With support vector machine (SVM) as the base classifier, they claim that they obtain accuracy as high as 90% in the identification task, and this outcome is superior over human beings' manual decision, which is about 60% accuracy. Other researchers have also proposed several approaches to automatic spam identification, such as Feng et al. [8], Zhou et al. [9], and Li et al. [10].

In our prior work [11], we examined two base classifiers: SVM and Naïve Bayes. The experimental result indicates that the average accuracy is very similar for each of the two base-classifiers. This conclusion is also consistent with other researchers, like Huang et al. [12]. For this reason, we only use SVM as the base classifier in this research.

Labeled reviews are crucial to improve the performance of spam identification. However, labeled reviews are expensive to obtain in practice. It requires extensive human labor and a lot of time is spent to produce enough labeled reviews to train the base classifier. In contrast, there are vast amounts of unlabeled reviews available on the Internet. Thus, it is natural to attempt to make use of unlabeled reviews to solve the spam review identification problem by adopting particular assumptions like smoothness assumption, cluster assumption, and so forth [13].

This paper proposes a novel approach to identifying spam reviews based on entropy and the co-training algorithm by making use of unlabeled reviews. The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 proposes the CoFea (Co-training by Features) approach. Section 4 conducts the experiments on the spam dataset, and Section 5 concludes the paper.

2. Related Work

2.1. Entropy

The size of information for a message and its uncertainty has a direct relationship. Entropy H of a signal x is defined by Shannon [14] to measure the amount of information. The entropy can be written explicitly as follows in Equation (1).

$$H(x) = -\sum_{i=1}^{n} p(x_i)\log_b p(x_i) \tag{1}$$

where $H(x)$ is the entropy of a discrete variable $X = \{x_1, ..., x_n\}$. b is the base of the logarithm (generally $b = 2$, and the unit of entropy is bit). $p(x_i)$ is the probability of samples where x_i represents each sample of the data point i. The value of $p(x_i)\log_b p(x_i)$ is taken to be 0 in the case of $p(x_i) = 0$.

Based on the idea of entropy, in the proposed CoFea approach, we calculate the entropy $H(x)$ of each term x in Equation (2).

$$H(x) = -\sum_{i=1}^{2} \frac{|D_i|}{|D|}\left(\frac{x_i}{|D_i|}\log\frac{x_i}{|D_i|} + \frac{|D_i|-x_i}{|D_i|}\log\frac{|D_i|-x_i}{|D_i|}\right) \tag{2}$$

Here, $|D|$ is the total number of reviews. $|D_1|$ is the number of truthful reviews and $|D_2|$ is the number of deceptive reviews. x_1 is the number of truthful reviews which contain term x, and x_2 is the number of deceptive reviews which contain term x. The result of the summation is the entropy value of term x. If one term occurred with same frequency in both truthful and deceptive reviews, we cannot deduce useful information from it for deceptive review identification. However, if one term occurred merely in either truthful or deceptive reviews, then it can provide useful information for deceptive review identification due to the potential link between this term and the label of reviews. Further, as for the former terms, they have larger entropy values than the latter terms. From this point of view, the smaller the entropy is, the greater the amount of information the term contains for deceptive review identification.

We sort the lexical terms of all the reviews by its entropy scores in a descending order. Then we evenly divide the terms of the sequence into two distinct subsets: all odd-numbered terms groups and all even-numbered groups. Feature selection is conducted based on terms' entropy value—that is, we regard subset I as view X_1, and subset II as view X_2.

2.2. The Co-Training Algorithm

The Co-training algorithm is a semi-supervised method, invented to combine labeled and unlabeled samples when the data can be regarded as having two distinct views [15]. Pioneers using the co-training method obtain accuracy as high as 95% for the task of categorizing 788 webpages by using only 12 labeled ones [16]. Other researchers extend this study into three categories: co-training with multiple views, co-training with multiple classifiers, and co-training with multiple manifolds [17]. In this paper, we use co-training with multiple views on spam review identification.

We use two feature sets as X_1 and X_2 to describe a data sample$x$. In other words, we have two different "views" on the data sample x, and each view can be represented by a vector x_i $(i = 1, 2)$. We assume that the two views X_1 and X_2 are independent of each other and each view in itself can provide sufficient information for the classification task. When we utilize f_1 and f_2 to denote the classifiers derived from the two views, we will obtain $f_1(x_1) = f_2(x_2) = l$, where l is the true label of x. We also can describe this idea in another way: for a data point $x = (x_1, x_2)$, if we derive two classifiers as f_1 and f_2 from the training data, $f_1(x_1)$ and $f_2(x_2)$ will predict the same label l.

2.3. Support Vector Machine (SVM)

SVM is a type of supervised learning model in machine learning proposed by Vapnik et al. [18]. This method minimizes the structural risk and turns out to provide better performance than the traditional classifiers. Formally, an SVM constructs a hyperplane in a Hilbert space. The Hilbert space has higher dimensions than that of the original one using kernel methods [19]. A good hyperplane could make the largest distance to the nearest training data point of any class.

The optimum hyperplane can be expressed as a combination of support vectors (i.e., the data samples in the input space). Generally, the optimization problem of hyperplane can be written as follows [20].

$$\min_{\omega,\zeta_i} \frac{1}{2}\|\omega\|^2 + C\sum_{\mathrm{i}} \zeta_i \tag{3}$$

subject to

$$y_i(x_i \cdot \omega + b) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0, \forall i \tag{4}$$

Here, $(x_i, y_i)(1 \leq i \leq l)$ are the labeled data samples with l labels. For each $i \in \{1, ..., n\}$, we use the slack variable ζ_i to tolerate those outlier samples. Note that $\zeta_i = max(0, 1 - y_i(w \cdot x_i + b))$, if and only if ζ_i is the smallest nonnegative number satisfying $y_i(\omega \cdot x_i + b) \geq 1 - \zeta_i$. ω is the slope vector of the hyperplane. After the optimal hyperplane problem is solved, the decision function $f(x) = sgn((w, x) + b)$ can be used to classify the unlabeled data. Intuitively, the larger the distance of

a data point from the hyperplane is, the more confident we will be to classify the data sample using SVM. For this reason, we use the distance $\frac{|(w,x)+b|}{||\omega||}$ as the confidence of the classifying result.

3. CoFea—The Proposed Approach

The co-training algorithm needs two different views to combine labeled and unlabeled reviews. The state-of-the-art co-training techniques usually use different types of views. For instance, Liu et al. [17] propose combining textual content (lexical terms) of a webpage and the hyperlinks in the webpage as two distinct views for webpage classification using the co-training algorithm. Differing from their method, the CoFea algorithm is proposed to identify spam reviews using two views comprising purely lexical terms. The details of the CoFea algorithm are described in Algorithm 1 as follows.

Algorithm 1 The CoFea Algorithm

Input:

L: a n_L sized set of truthful or deceptive labeled reviews;
U: a n_U sized set of unlabeled reviews;
X_1: the feature 1 set of terms derived from the words in reviews;
X_2: the feature 2 set of terms derived from the words in reviews;
K: the iteration number;
$n_{U'}$: the number of reviews U' drawn from U;
n: the number of selected reviews which are classified as truthful;
p: the number of selected reviews which are classified as deceptive;
Y: the reviews' label, i.e. $Y = \{deceptive, truthful\}$.

Output:

Two trained classifiers $f_1^{(K)} : X_1 \to Y$ and $f_2^{(K)} : X_2 \to$ Y

Procedure:

1. Randomly sample a $n_{U'}$-sized reviews set U' from U;
2. Based on the L set, train a classifier $f_1^{(0)}$ on the view of X_1 and a classifier $f_2^{(0)}$ on the view of X_2;
3. For $t = 1, ..., K$ iterations:
4. Use $f_1^{(t-1)}$ to classify the reviews in U';
   Use $f_2^{(t-1)}$ to classify the reviews in U';
5. Select n reviews from U' classified by $f_1^{(t-1)}$ as truthful;
   Select p reviews from U' classified by $f_1^{(t-1)}$ as deceptive;
   Select n reviews from U' classified by $f_2^{(t-1)}$ as truthful;
   Select p reviews from U' classified by $f_2^{(t-1)}$ as deceptive;
6. Add all those selected $2n + 2p$ reviews to L, remove them from U';
7. Randomly choose $2n + 2p$ reviews from U and merge them to U';
8. Based on the new L set, retrain a classifier $f_1^{(t)}$ on the view of X_1 and a classifier $f_2^{(t)}$ on the view of X_2;
9. End for.

At the preparatory phase, we produce a lexicon including all those terms appearing in the reviews. Then, the lexicon is divided into two distinct subsets evenly. Here, we take one subset as I and the other as II. That is, the terms in subset I are different from the terms in subset II and vice versa. Note that the two subsets are of the same size. Further, we regard subset I as view X_1, and subset II as view X_2.

With lines 1 and 2, we use the labeled reviews L to train our base SVM classifier with two views, X_1 and X_2. With lines 4–6, we use the trained classifiers to classify the unlabeled reviews in the test set U' and select the number of $2n + 2p$ classified reviews from U' to augment the training set L. Then, with line 7, we fetch $2n + 2p$ unlabeled reviews from set U to complement the test set U'. Finally, with line 8, the base classifiers are retrained using the augmented training set L. With K times of iteration, the CoFea algorithm is trained completely for spam review identification. We adopt the 10-fold cross-validation method to evaluate the accuracy in testing the CoFea algorithm.

There are two different methods for feature selection (in this paper, we use the terms occurring in the reviews as our feature) and those features are used to represent reviews as data samples. That is, each individual review is represented by a numeric vector and each dimension of the vector corresponds to one feature.

The CoFea-T (T abbreviates "total") strategy is to use all terms in the lexicon for review representation. Under this strategy, we use the subset I and the subset II of views X_1 and X_2 as the features to represent each review as two numeric vectors. Then, the two numeric vectors are used to train the base classifier with the CoFea algorithm.

The details of the CoFea-T strategy are shown in Algorithm 2.

Algorithm 2 The CoFea-T Strategy

Input:

L: a n_L-sized set of truthful or deceptive labeled reviews;

U: a n_U-sized set of unlabeled reviews;

${r_L}^a$, ${r_L}^b$: a vector representing one review in the labeled reviews, a, b means the view of a or b;

${r_U}^a$, ${r_U}^b$: a vector representing one review in the unlabeled reviews, a, b means the view of a or b;

${R_L}^a$, ${R_L}^b$: a representation set of L, a, b means the view of a or b;

${R_U}^a$, ${R_U}^b$: a representation set of U, a, b means the view of a or b;

Output:

Four sets of vectors ${R_L}^a$, ${R_L}^b$, ${R_U}^a$ and ${R_U}^b$.

Procedure:

1. Traversing all the reviews without repetition both labeled and unlabeled, we can get a ρ-sized lexicon with the entropy of each term as the term additive attribute;
2. Sort the terms by entropy;
3. Represent every review with vector, use the odd–even order of the terms in the lexicon to divide them. Get ${r_L}^a$, ${r_L}^b$, ${r_U}^a$ and ${r_U}^b$. Each vector has the length of $\rho/2$, and each vector has nearly the same entropy words;
4. Return ${R_L}^a$, ${R_L}^b$, ${R_U}^a$ and ${R_U}^b$.

The CoFea-S (S abbreviates "sampling") strategy uses some of the terms in the subset I and the subset II for review representation. By the entropy score, we know that some lexical terms in the reviews are of very limited information for spam review identification. That is to say, these terms occur in deceptive and truthful reviews without any difference in frequency. If the terms with limited information were used for review representation, as conducted in the CoFea-T strategy, then it will induce much computation complexity because the length of each review vector would be very long.

Thus, it would be wise to reduce the length of each review vector for a time-saving benefit in the CoFea algorithm. This idea motivates the proposal of the CoFea-S strategy. Under this strategy, we rank all terms in the lexicon by its entropy score. Then, we predefine a threshold entropy score to remove the terms in the lexicon and use the new lexicon as a partition to produce the subset I and the subset II for review representation.

The details of the CoFea-S strategy are shown in Algorithm 3. The difference between the CoFea-T strategy and the CoFea-S strategy is at line 3 in Algorithm 3. Under the CoFea-S strategy, we stipulate that the length of the numeric vector of each review should be θ. That is to say, we use the top 2θ terms with minimum entropy scores as the lexicon. Then, the lexicon is further split into two subsets, the subset I and the subset II, for review representation to construct the views X_1 and X_2. In the experiments, we compare the two strategies embedded with the CoFea algorithm in spam review identification.

Algorithm 3 The CoFea-S Strategy

Input:

L: a n_L-size set of truthful or deceptive labeled reviews;
U: a n_U-sized set of unlabeled reviews;
${r_L}^a$, ${r_L}^b$: a vector representing one labeled review, a, b means the view of a or b;
${r_U}^a$, ${r_U}^b$: a vector representing one unlabeled review, a, b means the view of a or b;
${R_L}^a$, ${R_L}^b$: a representation set of L, a, b means the view of a or b;
${R_U}^a$, ${R_U}^b$: a representation set of U, a, b means the view of a or b;

Output:

Four sets of vectors ${R_L}^a$, ${R_L}^b$, ${R_U}^a$ and ${R_U}^b$.

Procedure:

1. Traversing all the reviews without repetition both labeled and unlabeled, we can get a ρ-sized lexicon with the entropy of each term as the term additive attribute;
2. Sort the terms by entropy;
3. Give a certain number θ to determine the length of a vector, i.e., the top minimum entropy terms will be in the vector;
4. Represent every review with vector, get ${r_L}^a$, ${r_L}^b$, ${r_U}^a$ and ${r_U}^b$. Each vector has the length of θ;
5. Return ${R_L}^a$, ${R_L}^b$, ${R_U}^a$ and ${R_U}^b$.

4. Experiments

4.1. The Data Set

The data set we use in the experiments is the spam dataset from Myle Ott et al. [7]. For each review, we conduct the stop-word elimination, stemming, and term frequency–inverse document frequency (TF-IDF) representation work [11]. The stop-word list is quoted from the USPTO (United States Patent and Trademark Office) patent full-text and image database [21]. The Porter stemming algorithm is employed to produce every individual word stem [22]. We extracted all those sentences from one of the reviews using the sentence boundary-determining method mentioned in Weiss et al.'s paper [23]. Representing all the terms in reviews by the TF-IDF method, which is a numerical statistical approach, reflects how important a word is in a document in collection or corpus [24]. We conduct entropy computation of all lexical terms in the reviews to divide them into two subsets as subset I and subset II under the CoFea-T strategy and the CoFea-S strategy, respectively. The basic information

about reviews in the spam data set is shown in Table 1. The data set has 7677 terms (i.e., words including numbers and abbreviations) after preprocessing.

Table 1. The basic information about the reviews in the data set from Ott et al.

Polarity	Category	Number of Hotels	Number of Reviews	Number of Sentences
Positive	Deceptive (from MTurk)	20	400	3043
	Truthful (from Web)	20	400	3480
Negative	Deceptive (from MTurk)	20	400	4149
	Truthful (from Web)	20	400	4483

MTurk: Mechanical Turk.

4.2. Experiment Setup

Based on the CoFea algorithm, three parameters, K, n, and p (i.e., the number of iterations, the number of unlabeled reviews classified as truthful, and the number of unlabeled reviews classified as deceptive) must be tuned in order to optimize the performance of the algorithm. On account of our previous work [11], we know that the value of n and p should be equal to each other. We track accuracies of spam review identifications when we tune the parameter $n(n = p)$ while fixing the parameter K. We also track accuracies of spam review identifications when we tune the parameter K while fixing the parameter $n(n = p)$. Because the CoFea-S strategy needs to predefine a threshold θ for feature selection, we use $\theta = 100$ in this paper. Results of the CoFea-T and CoFea-S strategies for parameter tuning are as shown in Figure 1a,b, respectively.

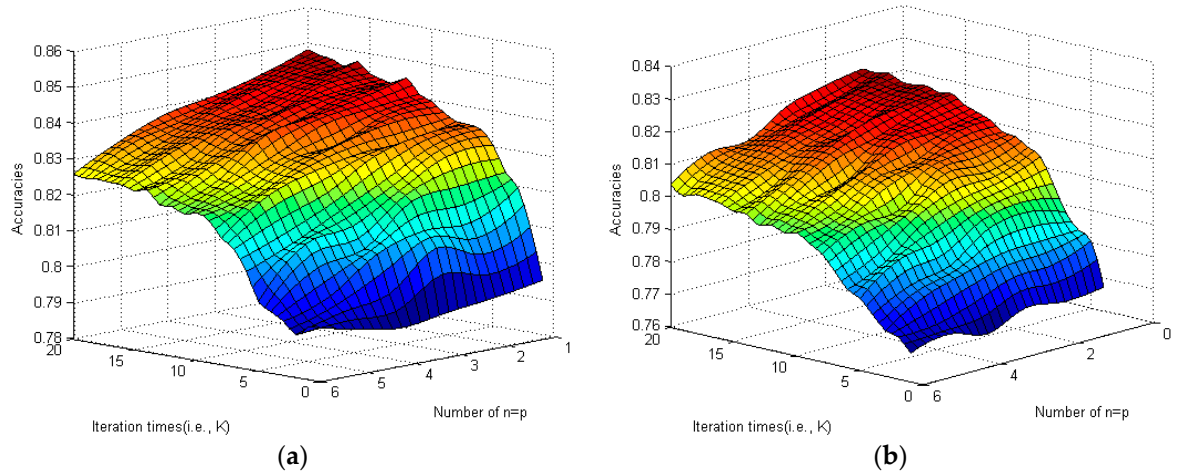

Figure 1. Experiment result of CoFea (Co-training by Features): (**a**) The result of CoFea-T; (**b**) The result of CoFea-S.

In the experiments, we set parameter $n = p$ from 1 to 6 and vary the parameter K from 1 to 20. To average the performance, we implement a 5-fold cross-validation. That means we use the training set containing 320 reviews and the test set containing 80 reviews 5 times. Each time we randomly choose 5% of the data (16 data samples) for classifier training, and 15% of the data (48 data samples) for the testing phase.

To compare the CoFea algorithm with other state-of-the-art techniques, we compare the CoFea algorithm and the CoSpa (Another approach for spam review identification based on Co-training algorithm) algorithm [11] in the experiments. The CoSpa algorithm use lexical terms as one view and probabilistic context-free grammar as another view in co-training for spam review identification. We

fix the parameters $n = p = 5$ in both the CoFea algorithm and the CoSpa algorithm when tuning the parameter K. We fix the parameter $K = 5$ in both algorithms when tuning the parameter $n = p$.

4.3. Experimental Results

Figure 1 shows the performances of the CoFea algorithm embedded with the CoFea-T strategy (left) and the CoFea-S strategy (right). In the figure, the warm-toned color lumps represent high accuracy, and the cool-toned color lumps represent low accuracy. We can see from Figure 1a that, when increasing the iterations, the accuracy of spam review identification has a certain degree of ascension. This approximate phenomenon also happened in the CoFea-S experiment. The accuracy quickly approaches the threshold at the $K = 6$ stage. That means the method (i.e., the CoFea algorithm) has certain limits. In general, the CoFea-T strategy performs better than the CoFea-S strategy.

Figure 2 shows the performances of different co-training algorithms—CoSpa, CoFea-T, and CoFea-S—in spam review identification. The CoSpa algorithm has an average accuracy of 0.8157, with 0.8375 as its highest accuracy. The CoFea-T algorithm has an average accuracy as 0.8202, with 0.8326 as its highest accuracy. The CoFea-S algorithm has an average accuracy as 0.7994, with 0.8105 as its highest accuracy.

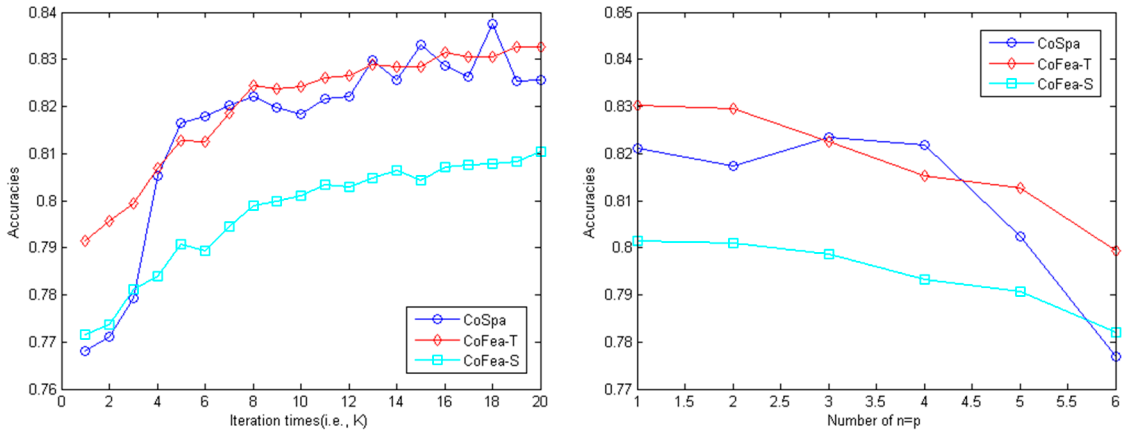

Figure 2. Performances of different co-training algorithms: (**a**) fixing $n = p = 5$; (**b**) fixing $K = 5$.

In order to better illustrate the effectiveness of CoFea-T, CoFea-S, and CoSpa, we employ Wilcoxon signed-rank test [25] to examine the statistical significance of experimental results when fixing $n = p = 5$. The Wilcoxon signed-rank test indicates the CoFea-T strategy outperforms the CoSpa algorithm with a two-tailed p-value of 0.0438, the CoFea-S strategy outperforms the CoSpa algorithm with a two-tailed p-value of 0.0002, and the CoFea-T strategy outperforms the CoFea-S strategy with a two-tailed p-value of 0.0000.

When fixing $K = 5$, the Wilcoxon signed-rank test indicates that the CoFea-T algorithm and the CoFea-S algorithm outperform the CoSpa algorithm, and the CoFea-T algorithm outperforms the CoFea-S algorithm.

The CoSpa algorithm only has its highest accuracy point without taking the stability into account. The CoFea-T algorithm has the highest mean accuracy among all the algorithms. The CoFea-S strategy has a good performance, very close to the CoFea-T strategy.

Figure 3 shows the time consumed in conducting the above experiments. The computer machine we use in the experiments had the following settings. CPU: Intel(R) Core(TM) i7-4700MQ @ 2.40 GHz; RAM: Kingston(R) DDR3 1600 MHz 4 × 2 GB; and Hard Drive: HGST(R) 500 GB @ 7200 r/min. The CoFea algorithm apparently has better speed performance in identifying the spam reviews, especially when dealing with a high number of iteration problems. Considering the motivation for proposing

the CoFea algorithm (i.e., using terms only without other views), we argue that it is one of the most advisable algorithm among the state-of-the-art techniques in spam review identification domain.

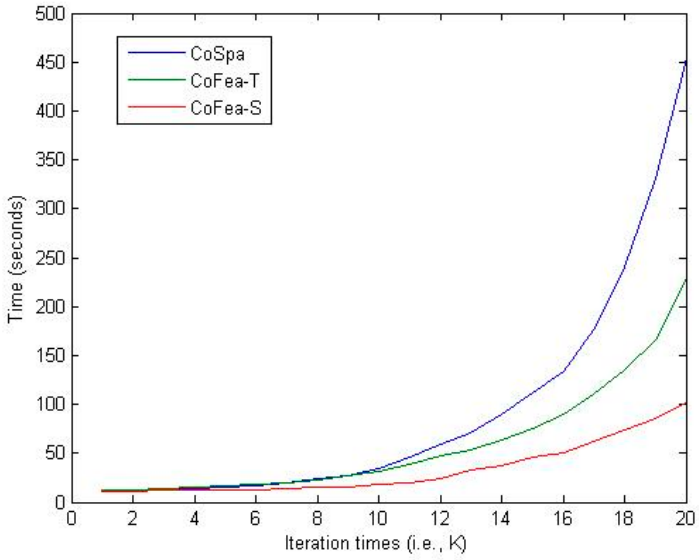

Figure 3. The time consumption result.

5. Concluding Remarks

In this paper, we propose a new approach, called CoFea, based on entropy and the co-training algorithm to identify spam reviews by making use of unlabeled reviews. The experimental results show some promising aspects of the proposed approach in the paper. The contribution of the paper can be summarized as follows.

First, we sort terms by means of entropy measures. This allows feature selection and can be conducted based on the amount of information a term contains.

Second, we propose two strategies, the CoFea-T strategy and the CoFea-S strategy, with different lengths of vectors of each view to be embedded with the CoFea algorithm for spam review identification.

Third, we conduct experiments on the spam review set to compare the proposed approach with the state-of-the-art techniques in spam review identification. Experimental results show that both the CoFea-T and CoFea-S strategies have produced good performances on spam review identification. The CoFea-T strategy has produced better accuracies than the CoFea-S strategy, while the CoFea-S strategy needs less time for computation than the CoFea-T strategy. Under the condition of lacking other views to implement the co-training algorithm, the CoFea algorithm is a good alternative for spam review identification using textual content only.

Although the paper has shown some promising aspects of using co-training on spam review identification, we admit that this paper is merely the initial step. In the future, on the one hand, we will use more data sets to examine the effectiveness of the proposed CoFea algorithm in spam review identification. On the other hand, we will also extend the co-training algorithm to more research areas such as sentiment analysis [26], image recognition [27], and text classification [28] to explore more fields. In fact, text classification is a basic technique for deceptive review identification. All the techniques mentioned in the paper can be extended to text classification. With the prosperity of e-commerce and online shopping, we regard the deceptive review identification to be a more practical application for users than pure text classification. As for sentiment analysis and image recognition, we cannot apply simple lexical terms as its feature for co-training, and there is no golden theory on how to build its feature sets. For this reason, we will firstly conduct research on feature extraction of the two tasks and further apply the proposed co-training approach on them.

Acknowledgments: This work is supported by the National Natural Science Foundation of China under Grant No. 91218302, 91318301, 61379046 and 61432001; the Fundamental Research Funds for the Central Universities (buctrc201504).

Author Contributions: Wen Zhang and Taketoshi Yoshida conceived and designed the experiments; Wen Zhang performed the experiments; Wen Zhang and Chaoqi Bu analyzed the data; Siguang Zhang contributed analysis tools; Wen Zhang and Chaoqi Bu wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aljukhadar, M.; Senecal, S. The user multifaceted expertise: Divergent effects of the website versuse-commerce expertise. *Int. J. Inf. Manag.* **2016**, *36*, 322–332. [CrossRef]
2. Xiang, Z.; Magnini, V.P.; Fesenmaier, D.R. Information technology and consumer behavior in travel andtourism: Insights from travel planning using the Internet. *J. Retail. Consum. Serv.* **2015**, *22*, 244–249. [CrossRef]
3. Zhang, W.; Wang, S.; Wang, Q. KSAP: An approach to bug report assignment using KNN search and heterogeneous proximity. *Inf. Softw.* **2016**, *70*, 68–84. [CrossRef]
4. Sui, D.Z. Mapping and Modeling Strategic Manipulation and Adversarial Propaganda in Social Media: Towards a tipping point/critical mass model. In Proceedings of the Workshop on Mapping Ideas: Discovering and Information Landscapes, San Diego, CA, USA, 29–30 June 2011.
5. Jindal, N.; Liu, B. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, CA, USA, 11–12 February 2008; pp. 219–230.
6. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–134. [CrossRef]
7. Ott, M.; Choi, Y.; Cardie, C.; Hancock, J.T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 309–319.
8. Feng, V.W.; Hirst, G. Detecting deceptive opinions with profile compatibility. In Proceedings of the International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October 2013.
9. Zhou, L.; Shi, Y.; Zhang, D. A Statistical Language Modeling Approach to Online Deception Detection. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 1077–1081. [CrossRef]
10. Li, H.; Chen, Z.; Mukherjee, A.; Liu, B.; Shao, J. Analyzing and Detecting Opinion Spam on a Large scale Dataset via Temporal and Spatial Patterns. In Proceedings of the 9th International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
11. Zhang, W.; Bu, C.; Yoshida, T.; Zhang, S. CoSpa: A Co-training Approach for Spam Review Identification with Support Vector Machine. *Information* **2016**, *7*, 12. [CrossRef]
12. Huang, J.; Lu, J.; Ling, C.X. Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. In Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 November 2003; p. 553.
13. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT Press: Cambridge, MA, USA, 2006.
14. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
15. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Workshop on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.
16. Committee on the Fundamentals of Computer Science—Challenges and Opportunities. *Computer Science: Reflections on the Field, Reflections from the Field*; ISBN: 0-309-09301-5. The National Academies Press: Washington, DC, USA, 2004.
17. Liu, W.; Li, Y.; Tao, D.; Wang, Y. A general framework for co-training and its applications. *Neurocomputing* **2015**, *167*, 112–121. [CrossRef]
18. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
19. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
20. Joachims, T. Transductive Inference for Text Classification using Support Vector Machines. In Proceedings of the 1999 International Conference on Machine Learning, Bled, Slovenia, 27–30 June 1999; pp. 200–209.

21. USPTO Stop Words. Available online: http://ftp.uspto.gov/patft/help/stopword.htm (accessed on 1 March 2016).
22. Porter Stemming Algorithm. Available online: http://tartarus.org/martin/PorterStemmer/ (accessed on 29 November 2016).
23. Weiss, S.M.; Indurkhya, N.; Zhang, T.; Damerau, F. *Text Mining: Predictive Methods for Analyzing Unstructured Information*; Springer: New York, NY, USA, 2004; pp. 36–37.
24. Rajaraman, A.; Ullman, J.D. Data Mining. In *Mining of Massive Datasets*; Cambridge University Press: London, UK, 2011; pp. 1–17.
25. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
26. Ravi, K.; Ravi, V. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowl. Based Syst.* **2015**, *89*, 14–46. [CrossRef]
27. Jain, A.K.; Duin, R.P.W.; Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 4–37. [CrossRef]
28. Zhang, W.; Yoshida, T.; Tang, X. Text classification based on multi-word with support vector machine. *Knowl. Based Syst.* **2008**, *21*, 879–886. [CrossRef]