문서 감정 분류를 위한 개선된 공동 학습 접근

자와드 칸, 아프타 발람, 무함마드 누만 칸, 이르판 울라, 무하마드 우매르, 구두스 우매르, 타리크 하비브 아프리디, 박성수, 이영구 [1]
경희대학교 컴퓨터공학과
{jkhanbk1, aftab, numan, irfan, umair, umair.qudus, afridi, cssp, yklee}@khu.ac.kr

An improved co-training approach for document Sentiment classification

Jawad Khan, Aftab Alam, Muhammad Numan Khan, Irfan Ullah, Muhammad Umair, Umair Qudus, Tariq Habib Afridi, Sung Soo Park, Young-Koo Lee[1]
Department of Computer Science and Engineering, Kyung Hee University

ABSTRACT

Sentiment Analysis (SA) is an active research area that is used to automatically extract useful information from the user-generated content (UGC) to classify into positive and negative classes. Recently, various machine-learning techniques, such as supervised machine learning, semi-supervised learning, and lexicon scoring for SA have been proposed. A high-quality training data is vital to learn a sentiment classifier for textual sentiment classification, but due to various domains, the labeled data for each domain is scarce or unavailable. The manual construction of labeled corpora is a time-consuming and laborious task because of the unstructured and unorganized nature of data. In order to address this issue, in this paper, we propose an improved co-training approach based on the n-gram and word2vec model for sentiment classification. Co-training is a semi-supervised learning approach that has effective applications in textual sentiment classification. The empirical evaluation of movie review datasets shows that the proposed approach outperforms existing techniques in terms of classification accuracy.

1. INTRODUCTION

Large size online user-generated data in the form of user reviews and comments are usually in an unstructured format. SA using natural language processing and machine learning (ML) techniques can provide help to researchers to analyze automatically unstructured UGC and extract valuable information for decision making. SA, also known as opinion mining, is the process that analyzes people's opinions, sentiments, and emotions about entities, like products, services, events, and topics. The main aim of SA is to categorize textual reviews into either positive or negative categories. Most of the methods for SA work on supervised ML. The Supervised ML methods require a large amount of labeled data to train a classifier (such as Support vector machines or Naive Bayes) for sentiment classification. However, there is a shortage of labeled data that could be used to train a classifier for various domains. Besides, it is difficult, expensive, and time-consuming to label a large amount of data manually for every domain, because it needs domain experts and experienced human annotators.

In the absence of enough labeled data, lexicon-based scoring and semi-supervised learning approaches for sentiment classification have been considered by the research community. The lexicon-based approaches need human efforts for lexicon construction and also suffer from low coverage and domain dependency. Semi-supervised Learning (SSL) is an ML technique that utilizes a large amount of unlabeled data along with a small amount of labeled data to learn classifiers and achieve better performance [1].

In literature, some SSL techniques, such as active learning, self-training, and co-training, have been used for textual sentiment classification. Dasgupta et al. [2] proposed an SSL approach to automatic sentiment classification. First, they mined unambiguous reviews using spectral techniques, then using them and classified the ambiguous reviews by a combination of active learning, transudative learning, and ensemble learning. Yang et al. [3] presented the SSL model (LCCT) for sentiment classification, where they integrated the lexicon-based and corpus-based learning in a co-training framework. Wang et al. [4] addressed an SSL approach, self-training, for sentence subjectivity classification. Xia et al. [5] proposed co-training for semi-supervised sentiment classification based on a dual view bag of words (BOW) representation. The existing SSL for sentiment classification mostly relies on the

traditional BOW approaches. Such approaches are beneficial due to its simplicity, but the feature space of word vector is high and also ignore the semantics aspects among words.

In this paper, we propose an improved co-training approach based on the n-gram model (unigram and bigram) and neural network-based pre-trained word2vec model for textual representation and sentiment classification. The proposed approach can efficiently utilize unlabeled instances to increase the classification performance. The main contributions of our proposed work are the following.

First, we propose an improved co-training approach based on two popular models: n-gram model (unigram and bigram) and word2vec model. Then, we extract and select the most appropriate sentiment features for classifier learning. Finally, we conducted experiments on movie review datasets and found that the proposed approach improves the classification performance.

2. Proposed Improved co-training approach

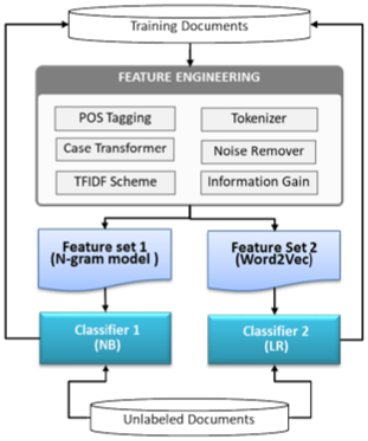

Figure 1: The architecture of the proposed improved co-training approach for SA

The architecture of the proposed improved co-training approach is shown in figure 1. Each review is tokenized, and noisy text is removed. We assigned POS tags to adjectives, adverbs, verbs, and nouns. We employed the TF-IDF scheme and Information Gain (IG) for feature ranking and selection. The process of the proposed method is shown in algorithm 1. Initially, the document is transformed into two feature vectors using n-gram (unigram and bigram) model and pre-trained word2vec model. We consider two independent and sufficient views, e.g., view 1 and view 2. The text features based on the n-gram model (unigram and bigram) are considered as one view, and the text features based on the pre-trained word2vec model (skip-gram) is considered as another view. The two classifiers are trained over the label set L, each working exclusively on one view, view 1 and view 2, respectively. Then the two classifiers (S1 and S2) predict the labels for the unlabeled documents according to a specific threshold λ. This process is continuing for several iterations. In each iteration, the high confidence level prediction of each classifier on the unlabeled instances is used to expand the labeled training set of the other classifier. In this way, the two classifiers are learning from each other, and the labeled training data is increased. This process is stopped when the unlabeled data is finished or until a particular criterion is met. After the complete learning process, the labeled training data of two co-training components are combined into a final classifier. Further, the test data is presented to the final classifier for evaluation.

Algorithm 1: Improved Co-training for textual sentiment classification

```
Begin
    /*unlabeled data of n instances*/
    inputs: X - {x1,x2,....xn};
    /*labeled data of n instances*/
     Y - {y1,y2,....yn};
     S - {s1,s2}; /*classifiers*/
     λ- {0.8}; /*Threshold*/
    output: Y0 {y1,y2,....yn}: /*labeled data*/
  While X contain instances
    /* divides labeled data into two parts*/
    Y1, Y2 ← Split Y
    /*s1 is train on view 1 using n-gram
    model, and s2 is trained on view 2 using
    word2vec model */
    s1, s2 ← train classifiers on Y1, Y2
    X ← pick a portion of the unlabeled data
    Xs1 ← classify X by s1 and pick high
         confident instances, use λ
    Xs2← classify X by s2 and pick high
         confident instance, using λ
    Add Xs1, Xs2 to Y
    Remove X1 from X
  Return Y0
End
```

3. EXPERIMENTS

We evaluated our proposed approach on the Cornell movie review dataset [6]. The Cornell movie review dataset consists of 2000 reviews contain 1000 positive and 1000 negative reviews. We used 80 percent instances as the training set and the remaining 20 percent instances as the test set. We conducted experiments with two models, (i) N-gram Model and (ii) Neural network-basedWord2Vec Model. We used Naive Bayes and Logistic Regression classifiers for the N-gram and Word2Vec Model, respectively. We adapted the high confidence threshold value 0.8, and the dimension for the appropriated feature selection is 1000. We used accuracy as an evaluation criterion for the sentiment classification performance.

Results: Figure 2, along with the table, shows the results for movie review sentiment classification based on n-

gram model (unigram and bi-gram), word2vec model, and proposed co-training approach. The classification performance obtained by co-training is high than the n-gram and word2vec model. The high performance of the proposed approach is due to the co-operative strategy of co-training components.

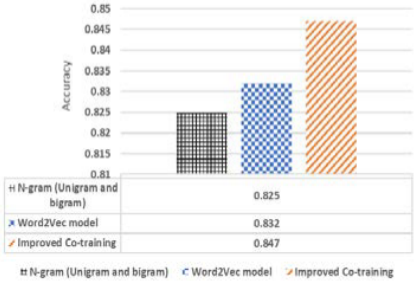

Figure 2: Classification performance on movie review dataset

Comparative experiment on movie review dataset

We compared the proposed approach with state-of-the-art approaches [2, 3, 7, 8] for textual sentiment classification. From figure 3 it is clear that the proposed approach obtained the best performance on the movie review dataset.

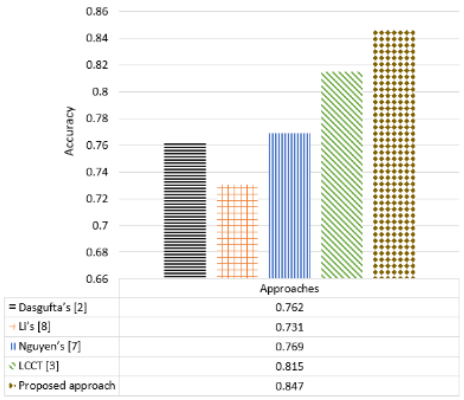

Figure 3: Classification performance of the proposed approach with state-of-the-art- approaches

4. CONCLUSION

With the rise of Web 2.0, a large amount of unstructured data is generated regularly. This unstructured data contains useful information for business analysis to know people's perceptions regarding online product reviews, services, events, topics. SA is a candidate solution that automatically extracts useful information from the unstructured UGC intending to classify into positive and negative classes. In this context, reliable training data is essential to learn a sentiment classifier for textual sentiment classification. Still, due to scarce label information, and the unstructured nature of data, the manual construction of labeled corpora is a time consuming and laborious task. In order to solve these issues, in this paper, we propose an improved co-training approach based on two different techniques (n-gram model and word2vec model) for textual sentiment classification. The experimental results show that the proposed approach outperforms existing approaches in terms of classification accuracy. In the future, we will incorporate Neural Network-based classification models and other techniques such as BERT, Glove, and ELMO in the proposed approach for SA, besides we will also apply the proposed approach on multiple domain datasets.

ACKNOWLEDGEMENTS

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2020-2015-0-00742) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

REFERENCES

[1] Z. Xiaojin, "Semi-supervised learning literature survey," *Computer Sciences TR*, vol. 1530, 2008.

[2] S. Dasgupta and V. Ng, "Mine the easy, classify the hard: a semi-supervised approach to automatic 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 2-Volume 2*, 2009, pp. 701-709.

[3] M. Yang, W. Tu, Z. Lu, W. Yin, and K.-P. Chow, "LCCT: A semi-supervised model for sentiment classification," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 546-555.

[4] B. Wang, B. Spencer, C. X. Ling, and H. Zhang, "Semi-supervised self-training for sentence subjectivity classification," in *Conference of the Canadian Society for Computational Studies of Intelligence*, 2008, pp. 344-355.

[5] R. Xia, C. Wang, X.-Y. Dai, and T. Li, "Co-training for semi-supervised sentiment classification based on dual-view bags-of-words representation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1054-1063.

[6] 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*, 2002, pp. 79-86.

[7] D. Q. Nguyen, D. Q. Nguyen, T. Vu, and S. B. Pham, "Sentiment classification on polarity reviews: an empirical study using rating-based features," 2014.

[8] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, "Employing personal/impersonal views in supervised and semi-supervised sentiment classification," *in Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 414-423.