



19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Leveraging User Ratings for Resource-Poor Sentiment Classification

Ngo Xuan Bach^{a,b,*}, Tu Minh Phuong^a

^aDepartment of Computer Science, Posts and Telecommunications Institute of Technology, Vietnam

^bFPT Software Research Lab, Vietnam

Abstract

This paper presents a general, simple, yet effective method for weakly supervised sentiment classification in resource-poor languages. Given as input weak training signals in forms of textual reviews and associated ratings, which are available in many e-commerce websites, our method computes class distributions for sentences using the statistical information of n-grams in the reviews. These distributions can then be used directly to build sentiment classifiers in unsupervised settings, or they can be used as extra features to boost the classification accuracy in semi-supervised settings. We empirically verified the effectiveness of the proposed method on two datasets in Japanese and Vietnamese languages. The results are promising, showing that the method is able to make relatively accurate predictions even when no labeled data are given. In the semi-supervised settings, the method achieved from 1.8% to 4.7% relative improvement over the pure supervised baseline method, depending on the amount of labeled data.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of KES International

Keywords: User Ratings; Sentiment Classification; Resource-Poor Languages; Unsupervised Learning; Semi-Supervised Learning.

1. Introduction

Sentiment analysis and opinion mining have become an attractive research topic in natural language processing and data mining communities in recent years. A sentiment analysis system analyzes opinionated texts, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations¹². Such a system can provide useful information not only for customers in choosing products and services but also for companies and vendors in marketing and market studies.

Sentiment classification is an important task in sentiment analysis and opinion mining. The goal of this task is to classify a given text as either *positive* or *negative* (or *neutral*). For example, the sentence “*It was a wonderful trip.*” can be labeled as positive, while the sentence “*That hotel provides very bad services.*” can be labeled as negative. Sentiment classification can be done at the sentence level as in these examples, where each sentence is classified as

* Corresponding author. Tel.: +84-126-317-2610.
E-mail address: bachnx@ptit.edu.vn

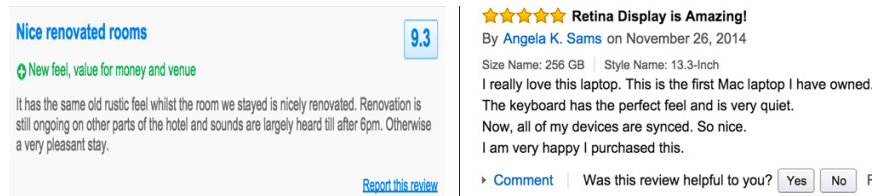


Fig. 1. Examples of reviews and associated ratings. Ratings are given as a score on the left, and stars on the right.

positive or negative or neutral, or at the document level, where a whole document is given one of the sentiment labels. In this work, we will focus on sentence level sentiment classification.

Many existing sentiment classification methods use machine learning techniques to predict sentiment labels. Various learning-based methods have been proposed and achieved good results, including supervised models^{2,15,17,21}, unsupervised models^{19,23,26}, and semi-supervised models^{11,13,20,22}. Most findings from these studies show that supervised methods give the most accurate results, and the success of those methods depends, to a large extent, on the size and quality of training data that are corpora annotated with sentiment labels. For English, a number of such corpora have been developed and made available for both commercial and academic purposes. For other (resource-poor) languages, such as Chinese, German, Spanish, Japanese, Vietnamese, etc. however, little work has been done to create similar annotated corpora, making it difficult to apply machine learning based sentiment analysis. Given the fact that building annotated data is a costly and time-consuming task, and sentiment analysis is highly domain-dependent³, it is tempted to develop methods that leverage other (unannotated) data sources in constructing sentiment classifiers, especially for resource-poor languages.

In this paper, we propose using overall ratings associated with user reviews and feedback in some websites as an additional source of information in constructing sentiment classification models. Many e-commerce websites allow users to give feedback about products or services they purchased in forms of both textual reviews and ratings as illustrated in Fig. 1. Although here we give examples of reviews in English, a plenty of reviews in other languages are available and can be used as resources for sentiment analysis.

In previous studies, some authors have found correlations between the sentiment of the sentences in a review and the associated rating^{8,22}. In this paper, we make a step further by considering ratings as weak training signals, and using them to compute the sentiment class distributions of sentences using the statistical information of n-grams found in the reviews. We show how these distributions can be used to build an unsupervised sentiment classifier, or can be used as extra features in a semi-supervised sentiment classification system. Our method is general in the sense that it is independent of languages. It is widely applicable because reviews with ratings are plenty for many languages and for many domains. Experimental results on two datasets in Japanese and Vietnamese languages show the effectiveness of the proposed method in both unsupervised and semi-supervised scenarios.

The contributions of this paper are two-fold. First, we introduce a method for generating class distributions of review sentences from raw data with overall rating information, and describe how to exploit class distributions in unsupervised and semi-supervised sentiment classification. Second, we conduct a series of experiments on two datasets to show the effectiveness of the proposed method.

The rest of this paper is structured as follows. Section 2 discusses related work. In Section 3 we present our proposed method, including how to infer class distributions from raw data and how to exploit them in unsupervised and semi-supervised scenarios. Section 4 describes our experiments on Japanese and Vietnamese datasets. Finally, Section 5 concludes the paper.

2. Related Work

Many supervised models have been proposed to deal with sentiment classification for English. Pang et al.¹⁷ conduct experiments to compare several supervised learning methods, including Naive Bayes¹⁸, Maximum Entropy Models¹, and Support Vector Machines^{6,24} for the task of sentiment classification. They obtained 82.9% accuracy on a corpus of English movie reviews. A survey on techniques and approaches applied to sentiment classification is presented in

Download English Version:

<https://daneshyari.com/en/article/489558>

Download Persian Version:

<https://daneshyari.com/article/489558>

[Daneshyari.com](https://daneshyari.com)