KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 3 (2017) 128–131

www.elsevier.com/locate/icte

Cross domain analyzer to acquire review proficiency in big data☆

Deepali Virmani[a], Preeti Arora[a,*], Pradnya Satish Kulkarni[b]

[a] *CSE, BPIT, New Delhi, India*
[b] *Mathematics, COER, Roorkee, India*

Received 2 October 2016; received in revised form 2 March 2017; accepted 13 April 2017
Available online 10 May 2017

Abstract

Sentiment analysis is the pre-eminent technology for extracting relevant information in the data domain. In this paper, a cross-domain sentimental classification approach, the cross-domain analyzer (CDA), is proposed, which will extract positive words and replace their synonyms to escalate polarity. Additionally, the approach blends two different domains and detects all self-sufficient words. This is executed on Amazon datasets, in which two different domains are trained to analyze the sentiments of the reviews in the other domain. The proposed approach contributes promising results in the cross-domain analysis, and an accuracy of 92% is achieved. In BOMEST, the CDA improves precision and recall by 16% and 7%, respectively.

Keywords: Feature extraction; Opinion mining; Sentimental classification

1. Introduction

Sentimental analysis is processing technique that can be used with blogs, reviews (movie, beauty, online shopping sites, etc.) to assess their sentiments as positive or negative. Customers express their views related to the product or facility they use. By analyzing those views, consumers can effectively improve their decision making. The classification of sentiment has been applied in various areas, such as market analysis, opinion mining, and opinion summarization [1,2]. In a single-domain analysis, all the sentiments are related to a single specific domain, which might not produce an ample scope for different domains. So, a solution is needed for the proposed cross-domain analyzer (CDA) [3,4], which trains the classifier for one or more domains and uses the trained classifier in a different domain known for better performance.

In this paper, an approach is proposed for revealing domain-dependent words and inferring independent words. A virtuous approach and CDA algorithm for cross-domain sentiment classification are proposed to improve the cross-domain data and minimize the gaps between domains. This algorithm, BOMEST, is adopted from the modified version of Jain et al. [5]. The algorithm works efficiently on a single domain with an accuracy of 78%. BOMEST is based on a bigram model that uses a bag of words (BOW) approach and tags all of the part of speech (POS) in an efficient way, to create a trained dictionary that stores all combinations of words (nouns + adjectives, adverbs + adjectives, etc.). Then, the score is calculated by multiplying positive polarity incremental value by 0.45 whereas negative polarity decremental value by 0.35 for the indexed data, and summing all lines in the range of 0.0 to 1.0.

The paper is organized as follows. Section 2 summarizes relevant literature and describes the study. Section 3 describes our proposed approach and the CDA algorithm used with the Amazon dataset. Finally, the results of our experiments are validated in Section 4. Section 5 presents conclusions.

2. Literature review

Blitzer et al. [4] focus on cross-domain classification and the challenges of training a classifier using source domains. They apply the trained classifier in a target domain for the

* Corresponding author.
E-mail addresses: deepalivirmani@gmail.com (D. Virmani), erpreetiarora07@gmail.com (P. Arora), syk_iitr@yahoo.com (P.S. Kulkarni).
Peer review under responsibility of The Korean Institute of Communications Information Sciences.
☆ This paper has been handled by Prof. Jun Heo.

http://dx.doi.org/10.1016/j.icte.2017.04.004

identification of features, and use the learning framework to find the significance of source and target domain features. Pan and Ni [6] propose a method for sentiment classification that bridges the gaps between the domains, using a spectral feature alignment (SFA) algorithm to group domain-specific words from different domains into unified clusters. These clusters can be used to reduce the gaps between domain-specific words from the two domains, thus enhancing the sentiment classifier.

3. CDA

3.1. Flow analysis of proposed approach

In this paper we propose a CDA approach for cross-domain analysis. To date, existing cross-domain approaches use a single-source domain and a classifier for predicting the target domain. The flow diagram for the CDA is shown in Fig. 1, with a detailed explanation as follows:

Raw data from dataset: The first step in our proposed CDA approach is to gather all the reviews of interest. The dataset used for implementation is an Amazon dataset available at http://jmcauley.ucsd.edu/data/amazon/. This dataset, which was used in our analysis, contains 160,792 reviews of baby products, 198,502 reviews of beauty products, 346,355 reviews of health products, and 1,689,188 reviews of electronics products from May 1996 to July 2014. After data gathering, the data are cleaned and prepared for classification by removing all HTML tags, extra spaces, white spaces, repetitive words, stop words, images, URLs, videos, and audio, which do not contribute to the meaning of the text. Then, a Porter Stemming Algorithm is used to remove suffixes from the words, reducing them to their roots to compact the dimensionality of the dataset.

Intensification of count: After stemming, reviews are tokenized using a BOW approach. Then BOMEST is used for POS tagging, which effectively identifies nouns, verbs, adverbs, and adjectives in order to create indexed data, assign scores, and store it in the trained dictionary [5,7].

Then, we identified all synonyms of a word available in the reviews, using MS Word Intro. The synonyms were replaced with the word, and the total word occurrence is calculated. For example, a dataset identifies “even” and “bad” as having positive and negative polarity, respectively, with their number of occurrences shown in Table 1.

“Even” is matched with its synonyms, and all the synonyms are replaced with the word “even”. The total count is evaluated as shown in Fig. 2A. Similarly, “bad” is compared to its synonyms, and its count is generated as shown in Fig. 2B.

Lexical_Boms_Dictionary: Lexical_Boms_Dictionary is the output that contains the positive and negative polarity list with the occurrence count of each word, as shown in Figs. 3A and 3B, which are generated from all the reviews in different categories.

Feature classifier: In sentimental classification, targets are the frequent words, POS, phrases, or terms that significantly affect an opinion to show positive or negative polarity. The selection of proper targets yields higher classification accuracy by reducing the extent of a text. Synonym replacement (WSR) and CDA (CBM) target-replace all synonyms for a given word. Searching and replacing text with word increases the total

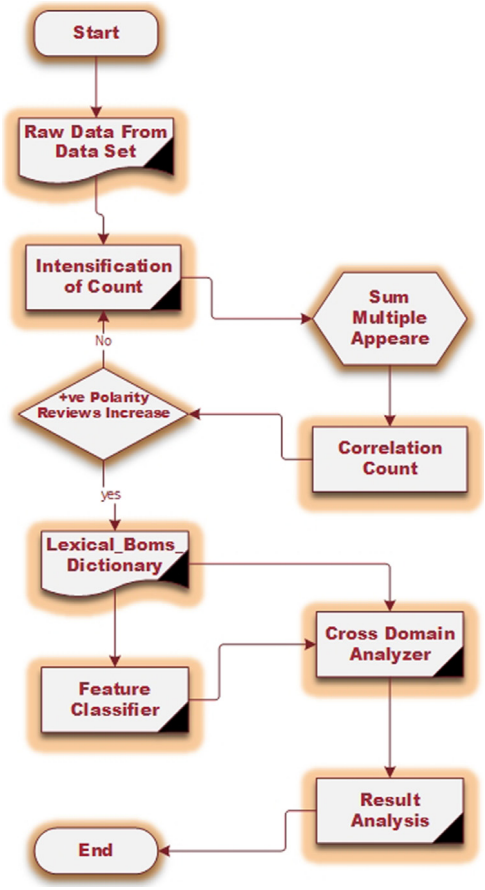

Fig. 1. CDA process sequence.

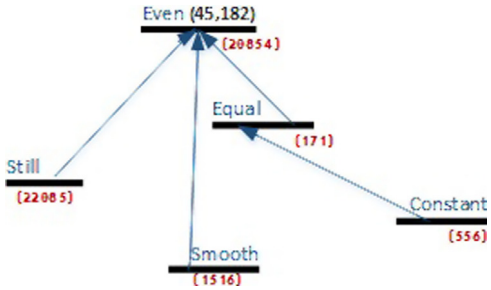

Fig. 2A. +ve synonym replacement.

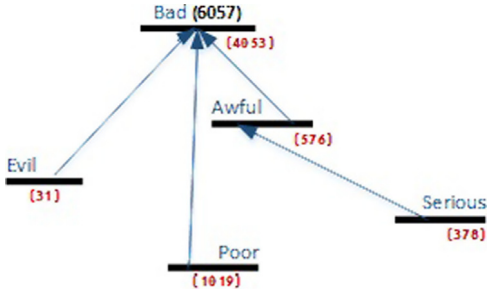

Fig. 2B. Negative synonym replacement.