Applied Computing and Informatics 14 (2018) 55–64

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

Original Article

Entropy based classifier for cross-domain opinion mining

Jyoti S. Deshmukh [a], Amiya Kumar Tripathy [b,*]

[a] Department of Computer Engineering, PAHER University, Udaipur, India
[b] Department of Computer Engineering, Don Bosco Institute of Technology, Mumbai, India

ARTICLE INFO

Article history:
Received 30 August 2016
Revised 11 February 2017
Accepted 20 March 2017
Available online 22 March 2017

Keywords:
Data mining
Opinion mining
Knowledge discovery
Expert systems
Information systems
Machine learning

ABSTRACT

In recent years, the growth of social network has increased the interest of people in analyzing reviews and opinions for products before they buy them. Consequently, this has given rise to the domain adaptation as a prominent area of research in sentiment analysis. A classifier trained from one domain often gives poor results on data from another domain. Expression of sentiment is different in every domain. The labeling cost of each domain separately is very high as well as time consuming. Therefore, this study has proposed an approach that extracts and classifies opinion words from one domain called source domain and predicts opinion words of another domain called target domain using a semi-supervised approach, which combines modified maximum entropy and bipartite graph clustering. A comparison of opinion classification on reviews on four different product domains is presented. The results demonstrate that the proposed method performs relatively well in comparison to the other methods. Comparison of SentiWordNet of domain-specific and domain-independent words reveals that on an average 72.6% and 88.4% words, respectively, are correctly classified.

1. Introduction

Opinionated text has created a new area of research in text analysis. Traditionally, fact and information-centric view of text was expanded to enable sentiment-aware applications. Nowadays, increased use of the Internet and online activities like ticket booking, online transactions, e-commerce, social media communications, blogging, etc. has led to the need for the extraction, transformation and analysis of huge amount of information. Therefore, new approaches need to applied to analyze and summarize the information [14].

Organizations take the review of product given by users seriously, as it adversely affects the sales of the product. Consequently, organizations take the effort to respond to the reviews, as well as monitor the effectiveness of its advertising campaigns. In this regard, sentiment analysis, a popular method, is used to extract and analyze sentiments [5,4].

Opinion mining is constantly growing due to the availability of views, opinions and experiences about a product/service online, as people are shedding their inhibition to express their opinions online. However, automatic detection and analysis of opinions about products, brands, political issues, etc. is a daunting task. Opinion mining involves three chief elements: feature and feature-of relations, opinion expressions and the related opinion attributes (e.g., polarity), and feature-opinion relations. An opinion lexicon is a list of opinion expressions or a set of adjectives, which are used to indicate opinion/sentiment polarity like positive, negative and neutral. This lexicon arises from synonyms in the WordNet, while antonyms are used to expand lexicon in the form of graphs. Such a dictionary-based approach has been used to partially disambiguate the results of parts of speech tagger. Further, fuzzy logic is used to determine opinion boundaries and to adopt syntactic parsing to learn and infer propagation rules between opinions and features [24,13].

Medhat et al. [18] conducted a survey on sentiment algorithms and its applications and found that sentiment classification and feature selection are more prominent areas in recent research. They also reported that Support vector machine and Naïve Bayes algorithms are the generally used algorithms to classify sentiments, and English is the language used in many resources like WordNet. Opinions and reviews given on social networking sites are used to generate datasets for the experiments.

The WordNet is a generalized lexicon and cannot be used for sentiment analysis; therefore, a need arose for the development

* Corresponding author at: School of Science, Edith Cowan University, Perth, Australia.
E-mail addresses: jyoja2007@gmail.com (J.S. Deshmukh), amiya@dbit.in (A.K. Tripathy).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2017.03.001
2210-8327/

of sentiment lexicon. SentiWordNet evolved out of WordNet was created as a lexical resource for opinion mining. It assigns to each synset of WordNet three sentiment scores: positive, negative and neutral [19,11].

Manufacturers, as well as consumers, require opinion mining tools to collect opinions about a certain product. The opinion analysis tools can be used by manufacturers to decide a marketing strategy for estimating production rate. On the other hand, consumers can use these tools to make decision on buying a new product or take a trip to vacation locations, or select hotel, etc.

Labeled opinions are used to analyze the classifier. Practically, labeled opinions for every domain is not possible, as it delimited by time and cost, while domain adaptation or transfer learning could be used to circumvent this limitation. In this paper, we propose the approach of domain adaptable lexicon which predicts the polarity of lexicon of one domain using a set of labeled lexicon of another domain using a modified entropy algorithm. This algorithm uses enhanced entropy with modified increment quantity instead of traditional entropy algorithm. Dataset of different types of products containing textual reviews has been used for evaluation. Multiple experiments were carried out to analyze the algorithm using accuracy and F-measure. We designed the approach in two phases: (i) preprocessing of dataset and (ii) applying classifier and clustering on dataset.

The rest of the paper is structured as follows. In Section 2, we describe the related work on domain adaptation approaches. In Section 3, we introduce our new improved entropy based semi-supervised approach. In Section 4, we evaluate our approach using cross-domain sentiment classification tasks, and compare it with other baseline methods. Finally, in Section 5 we draw conclusions on the proposed approach and set directions for future work.

2. Related work

The text documents containing opinions or sentiments were classified based on their polarity, i.e. whether a document is written with a positive approach or a negative approach. Although machine learning approach uses a word's polarity as a feature, the polarity of some words cannot be determined without domain knowledge. Hence, the reusability of learned result of a domain is essential. Transfer learning, also known as domain adaptation, can be used to address this challenge. Transfer learning utilizes the results learned in a source domain to solve a similar problem in another target domain [22]. Approaches used to classify single and cross-domain polarity opinions are usually a bag of words, n-grams or lexical resource-based classifiers.

The main aim of domain adaptation is to transfer knowledge across domains or tasks. Tagging the opinion word and building a classifier is time consuming and expensive, as opinions are domain dependent. Normally, users express their opinions specific to a particular domain. An opinion classifier trained in one domain may not work well when directly applied to another domain due to mismatch between domain-specific words. Thus, domain adaptation algorithms are extremely desirable to reduce domain dependency and labeling costs. Sentiment classification problem are considered as a feature expansion problem, in which related features are appended to reduce mismatch of features between the two domains. To overcome this problem, sentiment-sensitive thesaurus, which contains different words and their orientation in different domains, has been created. Bollegala et al. [7] used labeled, as well as unlabeled data, for evaluation. The results suggested that method performs significantly well compared to baseline.

To overcome domain adaptation issue, various adaptation methods have been proposed in the past, e.g., ensemble of classifiers. Combination of various feature sets and classification techniques yielded in the ensemble framework was proposed by Xia et al. [26]. They used two types of feature sets, namely, Parts-of-speech information and Word-relations and Naïve Bayes, Maximum Entropy and Support Vector Machines classifiers. For better accuracy, ensemble approaches like fixed combination, weighted combination and Meta-classifier combination, were applied. Li et al. [29] proposed active learning in which source and target classifiers were trained separately. Using Query By Committee (QBC) selection strategy, informative samples were selected, and classification decision were made by combining classifiers. Label propagation was used to train both classifiers. The result demonstrated that significantly outperformed the baseline methods.

Most often, opinions are given in the natural language. One major issue with natural language is the ambiguity of words. Fersini et al. [10] applied Bayesian ensemble model in which uncertainty and reliability was taken care. Greedy approach was used for classifier selection, while gold standard datasets were used for experimental analysis. However, classification performance is frequently affected by the polarity shift problem. Polarity shifters are words and phrases that can change sentiment orientation of texts. Xia et al. [28] addressed this issue using three-stage models which include detection of polarity shift, removal of polarity shifts and sentiment classification. Onan et al. [2] proposed the weight based ensemble classifier, in which weighted voting scheme was used to assign weight to classifier. As a base learner Bayesian logistic regression, Naïve Bayes, linear discriminant analysis, logistic regression and Support vector machine are used. A different type of experimental analysis shows better result than conventional ensemble learning. Da Silva et al. [8] used classifier ensembles formed by different classifier which is applicable to find products on the web. Augustyniak et al. [16] demonstrated Twitter dataset to have good accuracy only for positive and negative queries. They found that Bag of Words with ensemble classifier performs better than supervised approach.

Identification of feature and weighting is an important step in opinion mining. Khan et al. [12] proposed a new approach that identified features and assigned term label using SentiWordNet. In this method, point wise mutual information and chi square approaches were used to select features to SentiWordNet that were weighted. Support vector machine was used as classifier. Experimental evaluation on benchmark dataset shows effectiveness of approach.

Social networking sites contains text data in long format as well as short messages with symbols, emoticons etc. Opinion detection in long reviews is easy than short reviews, as short reviews contain fewer features, and more symbols, idioms etc. hence difficult to extract opinion. Lochter et al. [15] proposed ensemble approach to tackle this issue. This approach used text normalization methods to improve the quality of features. The features thus filtered and enhanced served as the input for machine learning algorithms. Proposed framework was evaluated using real and non-coded datasets and concluded that this approach was superior to other methods with a 99.9% confidence level. However, this approach was suggested to be expensive for offline processes due to higher cost of computing power. Hence, parallelization of this process has been stated as future work by the authors.

Sparseness is another issue in short text data. Word co-occurrence and context information approaches are generally used for solving sparseness issue. These approaches are less efficient. To address this problem, Chutao et al. [32] considered probability distribution of terms as the weight of terms.

Similar to ensemble classifiers, graph-based methodology are also used for domain adaptation. Dhillon et al. [25] proposed the graph-based domain adaptation method. Similarity graphs were constructed between features from all domains, if these features were similar then it demonstrated the presence of edge between