Computers and Electrical Engineering 000 (2017) 1–13

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

A novel sentiment aware dictionary for multi-domain sentiment classification☆

Vandana Jha[a,*], Savitha R[a], P Deepa Shenoy[a], Venugopal K R[a], Arun Kumar Sangaiah[b]

[a] Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India
[b] School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 19 July 2017
Revised 14 October 2017
Accepted 17 October 2017
Available online xxx

Keywords:
Sentiment analysis
Hindi language
Hindi Sentiwordnet
Multi-domain
Sentiment aware dictionary

ABSTRACT

Sentiment Analysis is a sub area of Natural Language Processing (NLP) which extracts user's opinion and classifies it according to its polarity. This task has many applications but it is domain dependent and a costly task to annotate the corpora in every possible domain of interest before training the classifier. We are making an attempt to solve this problem by creating a sentiment aware dictionary using multiple domain data. This dictionary is created using labeled data from the source domain and unlabeled data from both source and target domains. Next, this dictionary is used to classify the unlabeled reviews of the target domain. The work is carried out in Hindi, the official language of India. The web pages in Hindi language is booming after the introduction of UTF-8 encoding style. When compared with labeling done by Hindi Sentiwordnet (HSWN), a general lexicon for word polarity, the proposed method is able to label 23–24% more number of words of target domain. The labels assigned by our method and the labels given by HSWN, for the available words, are compared and found matching with 76% accuracy.

1. Introduction

Now a days, the opinions about movies, products or services are available in abundance on review sites, blogs and product sites. In products also, reviews are available for every type of products like kitchen appliances, books, DVDs, electronics etc.. Some of the always watched review sites are amazon.com, imdb.com, tripadvisor.com, caranddriver.com. These reviews are useful for both consumers and producers. The consumers can understand the performance of the product by reading other's views whereas the producers can get the information for improvement in the products or services. These advantages of reviews are the reason for the popularity of areas like opinion mining, opinion summarization, contextual advertising and market analysis. However, the words used to write reviews are different in different domains. For example, the words "energy saving" and "high quality" are used to write positive reviews about kitchen appliances, whereas "minimum in warranties" and "expensive" indicate negative reviews. In another way, the words "entertaining" and "enjoyable" are used to write positive reviews about DVDs, whereas "unfunny" and "boorish" indicate negative sentiments. It is expensive to train data in every new domain in which we want to test and classify the reviews. A supervised classifier trained in one domain

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Shuai Liu.

* Corresponding author.
E-mail address: vjvandanajha@gmail.com (V. Jha).

https://doi.org/10.1016/j.compeleceng.2017.10.015
0045-7906/

may not perform well in other domain test data because of the inability to learn unseen sentiment words. Hence, there is a need of sentiment aware dictionary from multiple domains to train the classifier for sentiment classification.

Sentiment classification is an important area of text classification whose goal is to classify a review based on the sentimental opinions conveyed by the reviewer in it. Sentiments can be classified into positive, negative, neutral or mixed category. A review with strong or more positive sentiment words in it, is treated as positive review whereas a review with strong or more negative sentiment words in it, is treated as negative review. A review with neither positive nor negative sentiment words is considered as neutral review whereas a review with both positive and negative sentiment words is considered as mixed review. Sentiment classification can be carried out at word level, sentence level or document level. Classifiers can be categorized, based on the domains in which they are trained and tested, into single-domain classifiers and multiple-domain classifiers. Single-domain classifiers are trained by the labeled data available in the domain and later tested on the same domain data whereas multiple-domain classifiers are trained by one or more domains, labeled or unlabeled data (source domains) and tested on another domain data (target domain). Our dictionary is useful for multiple-domain sentiment classification at document level. By creating a sentiment aware dictionary, the proposed method is able to label, unlabeled reviews from the target domain, into positive and negative classes with considerable accuracy. However, the proposed method can be easily extended to address multi-category sentiment classification problems.

1.1. Motivation

The multiple-domain sentiment classification is a challenging task and has recently received attention of the researchers. The main challenges involved are as follows:

1. It should be identified correctly that which features of the source domain are similar to which features of the target domain.
2. It should have a learning structure like a dictionary to accommodate the knowledge about the relatedness of the features from the source and target domains for the classification of target domain features.

Here, we are trying to overcome all these challenges by creating and using a multi-domain dictionary for labeling the target domain reviews into positive and negative classes. Our dictionary is in Hindi language. Hindi, the 4th largest language in the world, is the official language of India and has 310 million speakers across the world consisting of 4.46% of the world population.[1] Hindi content consumption on Internet is growing at whopping 94%.[2,3,4] But it is an uphill task for a resource scarce language like Hindi. Good Hindi language tagger and annotated corpus is not available. This problem is solved by using translation[5] of the reviews available in English language.

1.2. Contribution

In this paper, a fully automated Hindi Multi-Domain Sentiment Aware Dictionary, HMDSAD is proposed and used for the classification of unlabeled reviews of target domain. HMDSAD is created using labeled source domain data, unlabeled source domain data and unlabeled target domain data. It is based on the words those are co-occurring together in a review, also known as distributional context of the words. It keep, in multiple columns, different words from different domains, which express the same sentiment in the reviews.

A small part of our work describing HMDSAD creation is published in [1]. Here, we extend on that work in several ways:

1. HMDSAD dictionary is used to classify the unlabeled reviews from target domain into positive and negative classes. This can be done by labeling most of the words in the dictionary and using these labels for classification.
2. The dictionary is compared to HSWN, a general lexicon for word polarity, in which each sentiment associated word has a positive and a negative score. Our dictionary is able to label approximately 24% more number of words as compared to HSWN (shown in Table 10).
3. For the words available in HSWN, the label assigned by our methods is compared to the label given in HSWN. The achieved accuracy of available words with similar labeling is approximately 76% (shown in Table 11).

1.3. Resources

Hindi translation using translator[5] of the sentiment classification data set[6] for multiple-domain is used for our work. This is a benchmark data set, generated by Blitzer et al. [2]. It consists of product reviews from Amazon.com for four different product types: kitchen appliances, DVDs, electronics and books. The statistics of this data set is given in Table 1. The dataset contains user rating, review text and some other details. The reviews are labeled on the basis of user ratings. From now onwards, we refer this data set as review documents in this paper.

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.
[2] http://www.news18.com/news/tech/hindi-content-consumption-on-internet-growing-at-94-1-in-5-indian-users-prefer-hindi-google-1047247.html.
[3] http://www.internetworldstats.com/top20.htm.
[4] http://trak.in/tags/business/2015/08/19/hindi-content-content-consumption-growth-india-google/.
[5] https://translate.google.co.in/.
[6] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

Please cite this article as: V. Jha et al., A novel sentiment aware dictionary for multi-domain sentiment classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.10.015