Computers and Electrical Engineering 57 (2017) 311–323

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A comparison study on active learning integrated ensemble approaches in sentiment analysis☆

Deniz Aldoğan*, Yusuf Yaslan

Department of Computer Engineering, Istanbul Technical University, Maslak, Istanbul 34469, Turkey

ARTICLE INFO

Article history:
Received 29 January 2016
Revised 10 November 2016
Accepted 10 November 2016
Available online 23 November 2016

Keywords:
Active learning
Ensemble learning
Sentiment analysis
Machine learning
Artificial intelligence

ABSTRACT

One of the most challenging problems of sentiment analysis on social media is that labelling huge amounts of instances can be very expensive. Active learning has been proposed to overcome this problem and to provide means for choosing the most useful training instances. In this study, we introduce active learning to a framework which is comprised of most popular base and ensemble approaches for sentiment analysis. In addition, the implemented framework contains two ensemble approaches, i.e. a probabilistic algorithm and a derived version of Behavior Knowledge Space (BKS) algorithm. The Shannon Entropy approach was utilized for choosing among training data during active learning process and it was compared with maximum disagreement method and random selection of instances. It was observed that the former method causes better accuracies in less number of iterations. The above methods were tested on Cornell movie review dataset and a popular multi-domain product review dataset.

1. Introduction

As the data gathered by the social media increases in vast amounts on a daily basis, the analysis of such data has become extremely important for most companies, governmental or private institutions, universities, research centers, business entrepreneurs, etc. The main reason for these analyses is to retrieve feed-back information about people, products, services, events and so on. Therefore, sentiment analysis has been receiving an increasing amount of interest as mentioned in [1,2]. Sentiment analysis is basically a natural language processing application for identifying text sentiment, typically as positive, neutral or negative.

Since sentiment classification systems are generally dependant on the domain, there is a need for annotating large amounts of input data for each domain, which arises as a bottleneck for building more complex systems providing multi-domain solutions. In addition, in most of the modern machine learning problems, the number of data may reach up to enormous amounts whereas the ratio of labelled instances remain rather smaller. Therefore, choosing a method that can provide results with appropriate accuracies by using less number of training data becomes necessary.

Active learning has attracted attention to be used for sentiment analysis since it aims to minimize the necessary number of samples to input to the classifier model in order to reduce the annotation costs of large amounts of unlabelled data. In active learning, we let the system to choose among the data samples to train its classifier models. The algorithm chooses the training instances which have more significant roles in refining the model. Active learning has been applied to data sets

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. M. S. Kumar.

* Corresponding author.
E-mail address: deniz.aldogan@gmail.com (D. Aldoğan).

http://dx.doi.org/10.1016/j.compeleceng.2016.11.015
0045-7906/

in various languages. In the work [3], the authors attempt the problem of sentiment analysis in online reviews, blogs and forum texts written in English, French and Dutch by using machine learning techniques, namely Support Vector Machines (SVMs), Multinomial Naive Bayes (MNB) and Maximum Entropy (ME), while using Active Learning to reduce the number of examples to be annotated manually. Turkish sentiment analysis has been attempted as well with active learning [4]. Besides, there are also studies that consider domain knowledge. In the study of [5], six different active learning stategies incoorporating classifier uncertainty and sentiment dictionaries are experimented with in order to increase the training corpus quality.

There have been various amounts of research on instance selection mechanisms for active learning. In their paper, [6] perform a compherensive study on the current literature on active learning from an instance-selection perspective and classify them into two categories with (1) active learning merely based on uncertainty of independent instances, and (2) active learning by further taking into account instance correlations. They state that these correlations can further be divided into 4 groups, namely feature correlations, label correlations, both feature and label correlations and finally structure correlations. They also analyze and compare the algorithms in terms of computational complexities. As they have clearly demostrated the more information the algorithm takes into consideration, the more its time complexity increases.

One of the important study is given in [7], where the authors propose an active learning method that directly optimizes the learner's expected error on future test examples. They also emphasize that uncertainty sampling and Query By Committee (QBC) methods are not immune to being misled by the outliers in the data set. Their strategy utilizes a Monte Carlo approach to estimate the expected reduction in error due to the labeling of a query. Unlike existing algorithms which typically consider each unlabelled example in isolation, they use the entire pool of unlabelled data to estimate the expected error of the current learner, and they determine the impact on the expected error of each potential labelling request. They state that they reach high accuracy with four times fewer labelled examples than competing methods in experimental results on three real-world data sets. Instead of utilizing single classifiers, the study [8] presents a novel active learning approach, namely co-selecting, which takes into account both the imbalanced distribution of dataset and uncertainty by the utilization of two complementary classifiers. The uncertainty sampling they have applied depends on an uncertainty measure, while they also make use of a certainty measure since they believe that the quality and the quantity of the samples from the minority class are crucial. ME and SVM are used as machine learning classifiers on the multi-domain review dataset consisting of reviews for DVD, Electronics, Books and Kitchen, which had been introduced in [9]. Active learning has also been combined with ensemble learning and transductive learning to classify the ambiguous samples which have been revealed by spectral techniques [10]. In another study [11], a dynamic approach, called DUAL, is proposed where the selection strategy parameters are adaptively updated based on estimated future residual error reduction after each actively sampled point.

In addition to pool-based active learning strategies, there are also stream-based mechanisms for active learning. As stated in [12], the benefits of active learning are crucial especially for the analysis of online data streams in real-time. In their paper, a cloud-based scientific workflow platform which is able to perform on-line dynamic adaptive sentiment analysis of microblogging post, namely the ClowdFlow platform, is introduced. The study [13] also focuses on stream-based active learning by proposing a new uncertainty measure based on instance weighting instead of using popular uncertainty strategies. The authors claim that the proposed query strategy greatly improves the performance of the active learner compared to the commonly used active learning query strategies that are based on uncertainty. They also propose an efficient adaptive threshold for the stream-based active learning which can be used with any uncertainty measure. In the study [14], the authors propose an online active learning method by proposing a novel approach, namely Iterative Decreased Threshold (IDT), which is a threshold update algorithm for the margin-based selection criterion utilized in the active learning, and by combining it with the classical Stochastic Gradient Descent (SGD) updating rule. They test their algorithm on binary classification in six different data sets and include other SGD including algorithms for benchmarking. They claim that their algorithm is superior to the other algorithms due to its efficiency in computation and classification performance.

Active learning has also been integrated with semi-supervised methodoloy. The work in [15] proposes two novel approaches called active deep network and information density combined active deep network and state that these methods outperform the classical semi-supervised algorithms and deep learning techniques. In [16], active learning has also been integrated with semi-supervised learning by utilizing teams of walking particles.

In this study, it is intended to integrate active learning into a previously built benchmark framework to compare different common base and ensemble classifiers for sentiment classification of reviews [17].

The rest of the paper is organized as follows: the next chapter formalizes the sentiment analysis problem for review data and summarizes the methods used in the study, which is followed by the experimental study section discussing the results for the comparisons of algorithms. Conclusion and future work are the two final sections.

2. Roblem statement and explanation of methods used in the study

2.1. Problem formulation for the sentiment analysis of review data

The problem in sentiment analysis is classifying the polarity of a given text. In this study, the dataset is comprised of online reviews, which initially undergo a preprocessing step. Since unigram representation is utilized in the study, each review can be displayed as a vector of numbers, where each number corresponds to the frequency of a certain word of the vocabulary in the current review. Therefore, the whole dataset ($\mathbf{X}$) can be represented as a matrix where each row ($\mathbf{x}_i$)