ARTICLE IN PRESS
JID: KNOSYS [m5G;August 13, 2016;5:21]

Knowledge-Based Systems 000 (2016) 1–10

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Sentiment and emotion classification over noisy labels

Yaowei Wang, Yanghui Rao*, Xueying Zhan, Huijun Chen, Maoquan Luo, Jian Yin

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

A R T I C L E I N F O

Article history:
Received 30 March 2016
Revised 18 July 2016
Accepted 10 August 2016
Available online xxx

Keywords:
Hidden de-noising classification model
Sentiment analysis
Emotion detection
Noisy label

A B S T R A C T

With the rapid development of social media, online users are allowed to share their opinions conveniently. However, the ground truth for sentiments and emotions in social media is often constructed through surveys, hashtags or emoticons, where the labels may contain many errors. There are also amateurs and malicious users expressing offensive opinions or spreading fraudulent reviews, which has been identified as a growing threat to the trustworthiness of online comments. Thus, it is valuable for us to reconcile this noise in the ground truth when training sentiment and emotion classifiers. In this paper, we propose a hidden de-noising classification model (HDCM) that does not need any outsourcing systems or lexicons to estimate the actual sentimental or emotional category of each instance from corpora with noisy labels. The simplicity of assigning the category to a document by users under any contexts, and the authority of a user in assigning categories to documents with various domains are modeled as the unobserved hidden constraints in HDCM. Extensive evaluations using datasets with different scales of noisy labels validate the effectiveness of the proposed model for both sentiment and emotion classification tasks.

1. Introduction

The Web has generated product reviews, sentimental and emotional text about brand perception and political issues continuously [1], which can be utilized for capturing the opinions of consumers and the general public on product preferences, company strategies, marketing campaigns, and political movements [2]. Sentiment analysis refers to the inference of users' views, positions and attitudes in their written or spoken documents [3]. Both lexical and learning-based approaches have been utilized for this task [4,5]. Lexical-based methods detect sentiments by exploiting a predefined list of words or phrases, where each word or phrase is associated with a specific sentiment [6]. Learning-based methods often use labeled data to train supervised algorithms, which could adapt and create trained models for specific purposes and contexts [7]. With the rapid development of Web 2.0, another stream of work detects emotions (*e.g.*, joy, anger, fear, shame, and sadness) by fuzzy logic models [8], joint emotion-topic models [9] and a constrained optimization framework [10]. Although sentiment analysis and emotion detection methods were widely adopted in promoting sales [11], predicting stock prices [12], detecting the polarity of emerging political topics [13], and other areas, they face an additional challenge in noisy domains, *e.g.*, social media. Typically, the ground truth for sentiments and emotions in social media is constructed through surveys, hashtags or emoticons, in which the labels may contain many errors [10,14]. It can be hard for a model to reconcile this noise in the ground truth, because the quality of the available training data is critical to most natural language solutions, especially for supervised learning approaches [15].

To alleviate the detrimental effect of noisy labels on the sentiment analysis models, Barbosa and Feng [16] proposed to obtain the best label by combining the outcome of different sentimental data sources, *i.e.*, Twendz, Twitter Sentiment and TweetFeel. In their study, the tweets that are disagreed between these data sources in terms of subjectivity are removed. The similar strategy was also used to improve the quality of training labels for the broader classification problems. For instance, a repeated-labeling method was evaluated on the decision tree classifier by obtaining multiple noisy labels from online outsourcing systems such as Rent-A-Coder and Amazon's Mechanical Turk [17]. However, the above method assumed that all the labelers have the same importance, thus it could not always improve label quality and model quality. Another work is concerned with the detection of emotions over noisy labels via non-negative matrix factorization [10]. The limitation of such a constrained optimization approach is that it requires the word-emotion lexicon, topics, and various sentence features including ngrams, smiles, question mark, curse and greeting words as inputs.

* Corresponding author.
E-mail address: raoyangh@mail.sysu.edu.cn (Y. Rao).

http://dx.doi.org/10.1016/j.knosys.2016.08.012

In light of these considerations, we propose a hidden de-noising classification model (HDCM) that does not need any outsourcing systems or lexicons for both sentiment and emotion classification over noisy labels. We consider the following two unobserved hidden constraints in our model: (i) the simplicity of assigning the sentimental or emotion category to a document by users under any contexts, and (ii) the authority of a user in assigning the sentimental or emotion categories to documents with various domains. The main contributions are summarized as follows:

(1) We develop a unified probabilistic framework to address the challenging issue of noisy labels on sentiment and emotion classification, which does not rely on well-established lexicons or any outsourcing systems.
(2) Our model exploits the connection between users and labels only, that is, the model quality can be improved by the proposed HDCM when the additional information (*e.g.*, user behavior and link) is unavailable across multiple sites or incomplete in social networks.
(3) The proposed method performs well and is robustness on both sentiment and emotion classification when using documents with noisy labels for training.

The experiment is conducted on estimating the actual sentimental or emotional category of documents with noisy labels. The results show that HDCM achieves better and more stable performance than state-of-the-art baseline models. The potential applications of our model are opinion spam detection [18], quality control in crowdsourcing [11], sarcasm and nastiness detection [19]. The remainder of this paper is organized as follows. We describe related work in Section 2. We present HDCM in Section 3, which contains problem definition, objective function, parameter estimation and sentiment/emotion classification. We detail the dataset, experimental design, results, and discussions in Section 4. Finally, we present conclusions and future directions in Section 5.

2. Related work

In this section, we review works on sentiment analysis, opinion spam detection and emotion categorization, which will shed light on the background and current state of research in this area.

2.1. Sentiment analysis

In the past years, sentiment analysis has attracted an increasing amount of attention from researchers of natural language processing [20] and human-machine interaction [21]. Turney [22] applied an unsupervised learning algorithm to classify the sentiment orientation of reviews. Pang et al. [7] applied state-of-the-art supervised learning algorithms to classify movie reviews as positive or negative. Different from the above studies with single-domain data, Pan et al. [23], Bollegala et al. [24], and Xia et al. [25] investigated cross-domain sentiment analysis by spectral feature alignment, binary classification with a sentiment sensitive thesaurus, and feature ensemble plus sample selection, respectively.

In spite of the major efforts of sentiment analysis on various datasets, opinion spam detection is an important subtask for practical considerations, because useful information is often mixed with noisy data that makes sentiment analysis more difficult [18].

2.2. Opinion spam detection

Opinion spam is one of the various types of spams in reviews, web forums, and many other materials [18]. Jindal and Liu [26] addressed the importance of opinion spam detection and proposed machine learning models to detect three types of spam reviews, that is, untruthful opinions, reviews on brands only, and non-reviews. The method was based on features derived from textual contents, in which the percentage of opinion-bearing words, brand name mentions, numerals, capitals and other features were counted. Since then, the studies were extended to detect opinion spams using a wide range of information such as content, user behavior, link, and hashtag [27,28].

Several preliminary works used features derived from part-of-speech tags, n-grams, and the sentiment of reviews to detect opinion spams [29,30]. Another way discarded singleton malicious reviewers based on the behavioral patterns of users [31–33]. The method used graph propagation to identify spammers who write reviews in short bursts. Recently, a framework of opinion spam detection was proposed by exploiting features of metadata (text, timestamp and rating) and relational data (review network) [34]. The relations in social networks were also incorporated in several other opinion spam detection systems [11,35]. Experimental results have shown that the proposed framework is effectiveness and applicable to relational social networks.

Although opinion spam detection provides a possible way of refining labels for sentiment analysis, *e.g.*, discard labels that are annotated by the identified malicious reviewers before training, we tackle the issue of noisy labels on sentiment analysis with a unified probabilistic framework. Unlike the above methods which employ content, user behavior, link or other information to obtain guaranteed performance, our approach exploits the connection between users and labels only. Thus, the model quality can be improved by our method when the additional information is unavailable across multiple sites or incomplete in social networks.

2.3. Emotion categorization

Natural language text not only contains informative contents and its users' positive and negative attitudes, but also emotions including joy, anger, fear, disgust, sadness, and surprise [36]. To represent the meaning of emotion words, Kazemzadeh, Lee, and Narayanan [8] presented two fuzzy logic models based on interval type-2 fuzzy sets. Loia and Senatore [37] also modeled emotions as fuzzy sets, in which fuzzy modifiers tuned the intensity of emotions. With the incorporation of social annotation services such as online crowdsourcing [10], a stream of work aims to tag the emotions of news articles and social media materials. For example, Bao et al. [9] and Rao et al. [38–40] developed joint emotion-topic models to detect emotions from news articles. Zhang et al. [41] studied the emotion-tagging problem for news comments. Wang et al. [10] proposed a constrained optimization framework with constraints such as topic correlations, emotion bindings, and bias factors to discover emotions in noisy domains.

Different from the previous method of detecting emotions over noisy labels based on non-negative matrix factorization [10], we do not rely on well-established lexicons, topics, and specialized features for emotion classification.

3. Hidden de-noising classification model

In this section, we propose a hidden de-noising classification model-the HDCM-for sentiment and emotion classification over noisy labels. Documents with noisy labels are common in social networks, since there are often actual positive (negative) documents being annotated as negative (positive) by an amateur or spam labelers. For each document, we firstly model the probabilities of visible noisy labels equal to a hidden actual category by the logistic function. Then, the above objective function is estimated by the expectation maximization algorithm. Finally, the optimized hidden variables are used to train support vector machine classifiers. The proposed HDCM is applicable to both binary (sentiment) and multi-label (emotion) classification over noisy labels.