Future Generation Computer Systems 37 (2014) 438–448

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Towards building a social emotion detection system for online news

Jingsheng Lei [a], Yanghui Rao [b,*], Qing Li [b], Xiaojun Quan [c], Liu Wenyin [a]

[a] College of Computer Science and Technology, Shanghai University of Electric Power, China
[b] Department of Computer Science, City University of Hong Kong, Hong Kong
[c] Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong

HIGHLIGHTS

- We study the problem of social emotion detection of online users in the news domain.
- A novel algorithm is proposed and devised for generating emotion lexicon and to predict emotions.
- Based on the above-mentioned algorithm, a hybrid approach combining document selection and POS is developed.
- A social emotion detection system is developed for online news.
- Extensive evaluations have been conducted to validate the performance of our approach vis-a-vis other existing ones.

ARTICLE INFO

Article history:
Received 29 December 2012
Received in revised form
12 June 2013
Accepted 26 September 2013
Available online 21 October 2013

Keywords:
Social emotion detection
Emotion lexicon
Part of speech

ABSTRACT

Social emotion detection of online users has become an important task for mining public opinions. Social emotion detection aims at predicting the readers' emotions evoked by news articles, tweets, etc. In this article, we focus on building a social emotion detection system for online news. The system is built based on the modules of document selection, Part-of-speech (POS) tagging, and social emotion lexicon generation. Empirical studies are extensively conducted on a large scale real-world collection of news articles. Experiments show that the document selection algorithm has a positive effect on the social emotion detection. The system performs better with the words and POS combination compared to a feature set consisting only of words. POS is also useful to detect emotion ambiguity of words and the context dependence of their sentiment orientations. Furthermore, the proposed method of generating the lexicon outperforms the baselines in terms of social emotion prediction.

1. Introduction

News websites publish news reports describing events that happen around the world. News is not only a medium to state facts, but also to express an emotional state, such as touching, happiness, fear, anger, and so on [1]. Nowadays, an interactive emotion rating has become an important service on various channels of several news websites, including the official news portals of Chinese government and the popular news portal in China (see Table 1). This special service allows news readers to express their feelings by clicking the emotion labels (see Fig. 1). Facing these increasing electronic social datasets, it becomes useful and necessary to detect social emotions evoked by online news automatically. Leveraging the social emotion data, marketers are able to get and evaluate opinions of online customers, before promoting a new product or consolidating their brand; politicians, public relations officials and reporters can track public emotions towards current events.

The computational study of sentiments, emotions, and opinions expressed in text is known as Sentiment Analysis or Opinion Mining [2–6]. However, most works of Opinion Mining have focused on analyzing reviews [7–12]. The research on Emotion Detection of news began from the SemEval tasks in 2007 [13], which aimed at annotating news headlines by emotions. The underlying assumption is that all words, even those neutral ones, can effectively communicate information about the writer's positive or negative sentiment underlying an opinion [1], and provoke reader's pleasant or painful experiences. For the tasks of reader's emotion detection [2,3,6] (or the "social emotion detection" to be used throughout this paper), a lexicon is constructed firstly, in which each word is scored according to multiple emotion labels such as joy, anger, fear, surprise, etc. Then, the lexicon is used to detect social emotions of news headlines [14]. However, due to the limited information, annotating news headlines by emotions is a hard task. It is usually intractable to annotate sentence-level news headlines consistently, or to agree on emotions or emotion components of a word, even for annotators [15,16]. As a result, document-level online news is used to detect social emotions. In this paper, we adopt the lexicon-based approach in building our social emotion

* Corresponding author. Tel.: +852 65258082.
E-mail addresses: jshlei@126.com (J. Lei), raoroland@gmail.com (Y. Rao), itqli@cityu.edu.hk (Q. Li), xiaoquan@cityu.edu.hk (X. Quan), liuwenyin@gmail.com (L. Wenyin).

0167-739X/$ – see front matter
http://dx.doi.org/10.1016/j.future.2013.09.024

Table 1
The news websites with emotion rating service.

News websites	Channel	# of emotions	Emotion labels
www.People.com.cn	Opinion	8	Shocked, anger, sadness, touching, pleased, happiness, boredom, amusement
www.Sina.com.cn	Society	8	Touching, empathy, boredom, anger, amusement, sadness, surprise, warmness
www.Chinanews.com	Culture	8	Touching, empathy, boredom, anger, amusement, sadness, pleased, transient
www.Huanqiu.com	Technology	8	Shocked, anger, sadness, touching, pleased, happiness, boredom, amusement
www.QQ.com	Entertainment	6	Pleased, touching, empathy, anger, amusement, sadness
www.Sohu.com	Education	5	Pleased, anger, sadness, cool, disdainful

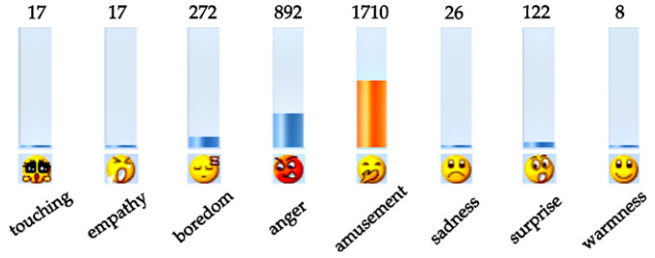

Fig. 1. An example of social emotions and user ratings.

detection system, since the lexicon cannot only be used to predict the emotion distributions of previous unseen documents, so as to assist authors in foreseeing how their work will influence online users emotionally, but also benefit a number of other social media applications, such as emotion-based document retrieval and emotion-aware news recommendation. With the help of lexical analysis, we can further identify organizations, cities, products, brands and other entities evoking different social emotions, so as to find what aspects the society liked and disliked, and know the targets of each emotional state.

The new social emotion detection system consists of three modules: document selection, Part-of-speech (POS) tagging, and a social emotion lexicon generation algorithm. We evaluate the system on an online news collection containing 40,897 articles gathered from the Sina society channel. It has 2,083,818 ratings distributed over 8 kinds of social emotions as shown in Fig. 1. Experimental results show that the proposed system can effectively choose a well-formed training set, and generate meaningful social emotion lexicon with POS information. For social emotion prediction, the Emotion-Topic Model (ETM) [2,3], SWAT system [14,15] and Emotion-Term model (ET) [2,3] are the state-of-the-art algorithms. As to be shown, our proposed social emotion lexicon generation algorithm outperforms SWAT and ET significantly, and yields competitive results with ETM. In addition, based on the best algorithm of generating the lexicon, choosing the training set randomly would yield an average accuracy of 55.29% and using the whole corpus for training would yield an accuracy of 57.32%. Our system with document selection algorithm achieves an average accuracy of 63.57%. We also conduct qualitative investigation on samples of the social emotion lexicon. The result shows that the lexicon not only reflects explicit emotive words, but also implicit words that convey emotions potentially. The POS of each word is useful to detect the emotional ambiguity of words and the context dependence of their sentiment orientations.

The rest of this article is organized as follows. Related works are given in Section 2. The problem formulation, existing methods of generating the social emotion lexicon, the framework of our social emotion detection system, and a demonstrative example are presented in Section 3. The dataset, results and discussions are illustrated in Section 4. Finally, we draw conclusions in Section 5.

2. Related work

In this section, we first review some of the related work on emotion classification, prediction and analysis, which arose from the year 2001 and shed light on the recent research of social emotion detection. Then, several works on social emotion detection from 2007 are summarized. Finally, we describe the research that is most related to our study, i.e., the emotion lexicon-based approach towards social emotion detection.

2.1. Emotion classification, prediction and analysis

In previous studies, most systems and algorithms focus on classification of emotions (opinions) from the perspective of the writers/authors, and the corresponding data sources are reviews or comments. During the incipient stage of research on sentiment analysis of reviews, the first way is using classification techniques, which include Naïve Bayes, Maximum Entropy and Support Vector Machines, or unsupervised learning technique to classify reviews into positive and negative [8,9]; the second way focuses on emotion categorization by constructing discriminate word dictionaries manually or semi-manually [7]. Other works focus on using linguistic heuristics or a set of seed words preselected, to classify the emotional orientation of words or phrases [10]. However, previous experiments show that the intuition of selecting discriminating words may not always be the best for humans [9]. Besides classifying emotions into positive or negative, predicting the rating scores of reviews has also been conducted [11,12]. As the rating scores are ordinal (e.g., 1–5 stars), the problem is tackled by regression. For sentiment analysis of reviews or comments, as review/comment holders are usually anonymous, useful information is often mixed with noisy data that makes sentiment analysis more difficult [17]. There are, for example, malicious users expressing offensive opinions, using their comments for the purpose of advertising, or even spreading rumors and fraudulent reviews. Considering this issue, Opinion Spam Detection [18] is essential to detect and filter out irrelevant information in reviews, which is an important subtask when performing sentiment analysis.

2.2. Social emotion detection

Recently, preliminary works focus on analyzing online users' sentiment responses when they are exposed to news articles [6]. This kind of sentiment response is called social emotions [2], and the studies conducted are social emotion detection [2,3,19,20].

The first line of research related to this direction is from the task of "Affective Text" in SemEval-2007 [14], where a corpus of news headlines extracted from Google news and CNN was provided. The task aimed at exploring the connection between news headlines and the evoked emotions of readers, in which three systems were introduced: SWAT, UA and UPAR7. The system SWAT [14,15] adopted a supervised approach by developing a word-emotion mapping dictionary. The dictionary was then used to score each word of a headline to have an average score for the headline and to decide its emotion. UA [14] gathered statistics from three search engines and computed the Pointwise Mutual Information (PMI) scores, to determine the emotion labels of headlines. UPAR7 [13,14] was a rule-based system which particularly relies on syntactic parser and lexicons. It utilized a linguistic and rule-based approach to tag news headlines for predefined emotions,