Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Using non-lexical features for identifying factual and opinionative threads in online forums

Prakhar Biyani [a],[*], Sumit Bhatia [b], Cornelia Caragea [c], Prasenjit Mitra [a]

[a] *College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA*
[b] *Xerox Research Center Webster, NY, USA*
[c] *Computer Science and Engineering, University of North Texas, Denton, TX, USA*

ARTICLE INFO

Article history:
Received 29 November 2013
Received in revised form 28 April 2014
Accepted 29 April 2014
Available online xxxx

Keywords:
Subjectivity analysis
Online forums
Classification

ABSTRACT

Subjectivity analysis essentially deals with separating factual information and opinionative information. It has been actively used in various applications such as opinion mining of customer reviews in online review sites, improving answering of opinion questions in community question–answering (CQA) sites, multi-document summarization, etc. However, there has not been much focus on subjectivity analysis in the domain of online forums. Online forums contain huge amounts of user-generated data in the form of discussions between forum members on specific topics and are a valuable source of information. In this work, we perform subjectivity analysis of online forum threads. We model the task as a binary classification of threads in one of the two classes: subjective (seeking opinions, emotions, other private states) and non-subjective (seeking factual information). Unlike previous works on subjectivity analysis, we use several non-lexical thread-specific features for identifying subjectivity orientation of threads. We evaluate our methods by comparing them with several state-of-the-art subjectivity analysis techniques. Experimental results on two popular online forums demonstrate that our methods outperform strong baselines in most of the cases.

1. Introduction

A large number of online forums in various domains (e.g., health, sports, travel, camera, laptops, etc.) exists today that enable internet users to discuss topics of mutual interest with other users, often separated by large geographical distances. The topics discussed in the threads of these forums are very unique in nature as they are often related to practical aspects of life (e.g., *How much to tip after bad service?*). Since such information is not readily available in other webpages, online forums are increasingly becoming very popular among internet users for discussing real life problems. Also, the interactive nature of online discussion forums enable users to discuss their problems in finer details and obtain customized solutions from their peers.

As a result of the ever increasing popularity and adoption of online discussion forums, hundreds of thousands of such forums exist today with a large number of discussions going on in each forum. Consequently, management and analysis of online forum data is a classical Big Data problem with complexities arising along the three dimensions of Velocity, Volume and Variety. To understand this, let us take the example of the official forum of the Ubuntu operating system (http://ubuntuforums.org). This forum boasts of close to 2 million threads created by more than 1.8 million users (*volume*). Further, the community has an active user population of more than 14,000 users actively participating in various discussions and thus, continuously creating new content (*velocity*). The user population that creates the content in these forums also has diverse characteristics. Users come from different social, educational and economic backgrounds and they may have varying level of expertise related to the topics of discussion. While some users may be information seekers, some might be information providers [1]. Thus, the content created by this diverse user population also had varied properties (*variety*) that makes the analysis of the content a non-trivial task. Thus, traditional text analysis and data management techniques cannot be directly applied to the online discussion data and thus, need to be adopted to address the peculiarities of this new data.

In this work, we *analyze subjectivity orientation of online forum threads*. We identify two types of threads in an online forum: *subjective* and *non-subjective* and we model the subjectivity analysis

[*] Corresponding author. Address: College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA. Tel.: +1 8149542802.

E-mail addresses: pxb5080@ist.psu.edu (P. Biyani), Sumit.Bhatia@xerox.com (S. Bhatia), ccaragea@unt.edu (C. Caragea), pmitra@ist.psu.edu (P. Mitra).

http://dx.doi.org/10.1016/j.knosys.2014.04.048
0950-7051/© 2014 Elsevier B.V. All rights reserved.

task as a binary classification problem. We define subjective threads as the threads discussing subjective topics that seek opinions, viewpoints, evaluations, and other private states of people and non-subjective threads as the threads discussing non-subjective topics that seek factual information. Table 1 shows a subjective thread from an online forum, Trip-Advisor New York. Table 2 shows a non-subjective thread from the same forum. In the former, the topic of discussion is *whether to tip or not after bad service?*, which seeks opinions, whereas the latter seeks factual information about *bands/artists playing in December in Madison Square Gardens*.

Even though there exist many previous works on subjectivity analysis of text, to the best of our knowledge, we are the first to address the problem of identifying subjectivity orientation of online forum threads in these works [2–4]. In the current work, we build upon our these previous works. Specifically, our main contributions are in terms of comprehensively evaluating our subjectivity classification model against strong baselines, using the classification models to predict and analyze subjectivity of threads started by top posting users in online forums, and analyze sources of error in subjectivity classification.

Previous works on subjectivity classification have extensively used lexical features such as bag-of-words, *n*-grams, combinations of *n*-grams and parts of speech tags, etc. [5–7]. A major issue with these features is their high dimensional feature space and hence there is a risk of model overfitting especially with small training data. Further, a large feature space (typically hundreds of thousands of features) results in higher resource requirements and longer times to train standard machine learning algorithms. The huge volume of data in online discussion forums further worsens this problem. In order to address the scalability issues, in this work we explore the possibility of using non-lexical and thread specific features for the subjectivity classification of threads. Specifically, we explore the following research question: *Can non-lexical thread specific features (e.g., number of users in a thread, number of posts in a thread, etc.) help in inferring the subjectivity of online forum threads?* To address the question, we propose and evaluate several thread specific features for subjectivity classification. While developing our features for the classification task, we design features to capture the diverse behavior of content creators (i.e., the participating users in a discussion). This is strikingly different from previous works on subjectivity classification, where no attention is given to the content creators. We compare the performance of our classification model with various state-of-the-art techniques and show that our model outperforms the baselines in most of the cases.

1.1. Why subjectivity analysis of online forum threads?

- ***Improving forum search***: Internet users search online forums, generally, for two types of information. Some of them search the forums for subjective information such as different viewpoints, opinions, emotions, evaluations, etc., on specific problems instead of a single correct answer. Other users want short factual (objective) answers. Previous works on online forum search have focused on improving the lexical match between searcher's query keywords and thread content [8,1,9]. However, these works do not take into account a searcher's intent, i.e., the *type of* information a searcher wants. Let us consider the following two example queries issued by a searcher to some camera forum: (1) How is the resolution of Canon 7D, 2) What is the resolution of Canon 7D. The two queries look similar, but they differ in their intents. In the first query, the searcher wants to know what other camera users think about the resolution of the Canon 7D, how are their experiences (good, bad, okay, excellent, etc.) with the camera as far as its resolution is concerned and other such types of *subjective* information. The second query, however, is *objective* in nature in which the searcher wants a factual answer, which, in this case, is the value of the resolution of the camera. Hence, queries having similar keywords may differ in their intents. Search algorithms based only on keyword search would perform badly for these types of queries. We believe that by knowing the type of information (subjective or objective) contained in a forum thread, these types of queries can be addressed in a better way. A forum search model can then match the searcher's intent with the type of information a thread contains in addition to the keyword match between the two and thus, handle the queries more intelligently.
- ***Spam detection***: Online forums are informal in nature. Often, there are trolls posting spam, extraneous, inflammatory and off-topic messages in discussion threads [10,11]. Forum administrators continuously monitor forums for such contents and remove them as they are against the community rules. The content of such messages is generally subjective in nature and hence can potentially be detected by analyzing threads for subjectivity.

The rest of the paper is organized as follows: The next section overviews the related work in the field of subjectivity analysis. Section 3 describes the problem and the features used for subjectivity classification. In Section 4, we describe our dataset, experimental settings and present and analyze the results of the classification. Section 5 concludes the paper and discusses the future work.

2. Related work

Subjectivity analysis has been an active field of research due to its important applications in opinion mining [12–16], question–answering [17–19,5,20], summarization [21], etc. Here, we first provide a brief survey of works on subjectivity analysis in general and then we review the works that performed subjectivity analysis in different domains (online review sites, community answers, etc.) and used it in different applications (opinion mining, question–answering, etc.).

Table 1
An example subjective thread.

Initiator	After looking for restaurants options for my trip to NY in September (Trip Advisor, Menu Pages, etc.) I can see that most of the complains are on the bad service received in the restaurant, but not the food quality. So as I am not used much to tip in restaurants as you do in the States (since I am not American and not living there), what do you do when you suffer bad service in a restaurant, even if the food i good? Do you still tip 15%? Thanks in advance for your comments on this
User1	I would tip 10%
User2	Actually, these days tipping 20% is more the norm for good service. If you get bad service, depending on how bad it is either (1) leave a smaller tip; or (2) do not leave a tip at all. However, in all my years of dining out, there have been only two occasions where we had such bad service that we did not leave a tip. Needless to say, we did not return to those places either!
User3	I lower the tip if the service is not good (once lowered it to under a $$). However, if you are not tipping because of bad service it is important to let someone in the restaurant know WHY you are not tipping!

Q1 Please cite this article in press as: P. Biyani et al., Using non-lexical features for identifying factual and opinionative threads in online forums, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.04.048