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 Dialog act Classification Supervised learning

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 analysis techniques. Experimental results on two popular online forums demonstrate that our methods outperform strong baselines in most of the cases.

© 2014 Elsevier B.V. All rights reserved.

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

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

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 plaving 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

Table 1

An example subjective thread.

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.).

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!

P. Biyani et al. / Knowledge-Based Systems xxx (2014) xxx-xxx

Table 2An example objective thread.

User1 Hi guys, We are coming over to catch Oasis at Madiso Sqaure Gardens in December. What other quality bands/artists are playing from 6 December onwards? Cheers User2 Have a look at www.pollstart.com and, in the weeks leading up to your trip, at www.timeout.com/newyork/

2.1. Subjectivity analysis

Wiebe et al. [22] did a seminal work on generating and using a gold standard dataset for subjectivity classification. They performed subjectivity classification of sentences using basic features such as presence of a pronoun, an adjective, a modal in the sentence. Bruce et al. [23] performed a case study of manual subjectivity tagging. Wiebe et al. [24] performed subjectivity classification of sentences in World Press articles using unannotated data. They used high precision rule-based classifiers for generating an initial training data and then used semi-supervised learning to iteratively learn subjectivity patterns and augment the training data. Su et al. [25] performed word sense subjectivity classification using the training data generated from the existing opinion mining resources and showed that the performance is comparable with that of the classifier trained on a dedicated training set. Other works have performed subjectivity classification across different languages [26,27]. They discussed and evaluated methods to develop subjectivity analysis tools for selected languages by applying machine translation on the available subjectivity analysis tools and resources for English language.

2.2. Opinion mining

An integral part of opinion mining and sentiment analysis is to separate subjective sentences from objective ones and then to identify the polarity (negative, neutral or positive) of the opinions expressed in the subjective sentences [12,14]. Works in this area have mainly focused on online review sites for summarizing product reviews given by different users of those products [28–30]. Our work, in contrast, deals with online forum threads. A review in a review site is a continuous piece of text written by a person with additional information such as ratings, date and time. On the other hand, a thread in an online forum has a distinctive structure due to the presence of messages posted by multiple users. Also, a review, usually, has a single role of providing user's feedback on a product whereas posts in a thread have multiple roles, e.g., a post can be a question, solution, feedback, junk, etc. [31]. These differences make subjectivity analysis of online forum threads different from that in review sites in both nature and the approaches that can be used for the analysis. For example, thread structure, role of posts and other thread-specific information can be used as features for subjectivity analysis (as will be described later in the paper).

2.3. Question-answering

There has been a great amount of research in developing automatic question-answering systems such as IBM Watson. Subjectivity analysis has been used to improve question-answering in online communities and social media [17–19,5,20]. Yu et al. [5] classify documents and sentences from news data into facts and opinions with the aim of improving answering of complex opinion questions. Stoyanov et al. [19] separate opinion (subjective) answers from factual (objective) answers and then filter out factual answers for opinion questions to improve answering of opinion questions in multi-perspective question answering. Somasundaran et al. [20] identify different types of attitudes in questions and answers and then use it to improve opinion question answering on web-based discussions and news data by matching the attitude types of questions and answers. Li et al. [6] classify questions in Yahoo QA as subjective or objective using semi-supervised learning by utilizing the text of labeled questions and their unlabeled answers for learning subjectivity patterns. These works did subjectivity analysis of questions and answers given by single authors in community sites. In contrast, we analyze the subjectivity of online forum threads that contain replies from multiple authors. These differences have implications described in the previous paragraph.

2.4. Online forums

In the domain of online forums, there have been two recent works that are close to our work. Hassan et al. [32] performed sentence-level attitude classification in online discussions to model user interaction that may be helpful in facilitating collaborations. Zhai et al. [33] classified sentences in online discussions as evaluative or non-evaluative for getting relevant opinion sentences. In contrast, our work does thread-level subjectivity classification as we are interested in knowing the subjectivity of the overall topic of discussion of a thread and plan to use it for improving online forum search in the future. There have been works analyzing dialogic structure of posts in online debates to find on which side of the debate (FOR or AGAINST) the posts are [34] and identify disagreements between posts [35]. However, the current work is very different from these works. We identify eight types of dialog acts expressed in a thread posts and use them to infer subjectivity of the thread's topic.

3. Problem formulation and feature engineering

In this section, we state our problem and describe various features used in the subjectivity classification task.

3.1. Problem formulation

An online forum thread discusses a topic specified by thread starter in the title and the initial post. The topics of discussion in the threads can either be subjective or non-subjective (see Figs. 1 and 2 for examples of subjective and non-subjective threads, respectively). Based on the definitions of subjective and objective sentences given by [23], we define a subjective topic of discussion as a topic that seeks people's opinions, viewpoints, evaluations, speculations, and other private states and a non-subjective topic as a topic that seeks factual information. We call a thread subjective if its topic of discussion is subjective and non-subjective if it discusses a non-subjective topic. We assume that in online forum threads subjective topics have discussions in subjective language (i.e., expressing different private states) and non-subjective topics have discussions in objective language (i.e., expressing facts and verifiable information). We note that there may be some cases where the assumption does not hold good, however, analysis of such exceptional cases is not the focus of this paper and is left for future work.

3.1.1. Problem statement

Given an online forum thread *T*, our task is to classify it into one of the two classes: *Subjective* (denoted by *s*) or *Non-Subjective* (denoted by *ns*).

3

P. Biyani et al./Knowledge-Based Systems xxx (2014) xxx-xxx

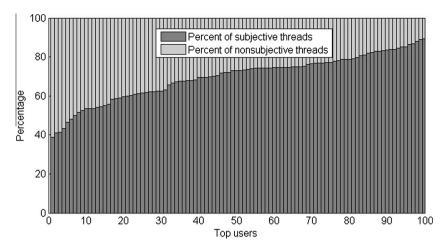


Fig. 1. Figure showing distribution of threads from top 100 users in subjective and non-subjective classes for Trip Advisor - New York forum.

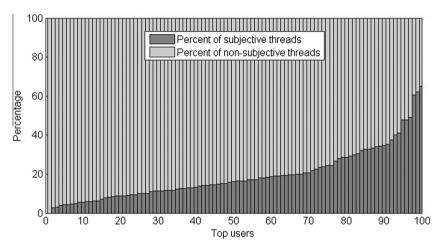


Fig. 2. Figure showing distribution of threads from top 100 users in subjective and non-subjective classes for Ubuntu forum.

In this work, we assume that a thread has a single topic of discussion which is specified by the thread starter in the title and the initial post. Analyzing subjectivity of threads with multiple topics is a separate research problem that is out of scope of this work.

3.2. Feature engineering

As discussed before, we wanted to explore the effect of using various thread specific features for subjectivity analysis of online forum threads and compare them with the state-of-the-art subjectivity analysis techniques. In this section, we describe the features used and intuition behind using them.

3.2.1. Structural features

We posit that subjective threads have different structural properties than non-subjective threads. Since subjective topics have more scope of discussion, we expect the subjective threads to be longer and invoke more participation of users than non-subjective threads. We use the length of a thread and the participation of users in a thread as features. For the length, we use the length of the initial post, the length of the thread and the average of the length of all the reply posts in the thread as features. All the lengths are measured in terms of the number of words. For the participation, we use the number of users that participated in the given thread, the number of posts and the average number of posts by a user in a thread as features.

3.2.2. Dialog act features

Online forum threads have conversational nature and hence there are different types of dialog acts (question, solution, feedback, etc.) expressed in thread posts [31,36,37]. For example, a thread starts with a *question* posted by the thread starter. Then, there are posts (by other users) that ask for some *clarifying* details about the question and the thread starter provides further details to make the question clearer. After getting the details, users suggest solutions and finally there are *feedbacks* (by the thread starter or other users) to the suggested solutions that can be *positive* or *negative*. Also, there may be posts that ask the same question (as asked in previous posts) and posts that are junk and not related to thread discussion. We posit that dialog acts expressed in the posts maybe helpful in identifying thread's subjectivity. In a subjective thread, there could be multiple solutions suggested for a question (e.g. Sony or Nikon which is better?) as there is no single correct answer to subjective questions and hence multiple feedbacks would be given. In contrast, in non-subjective threads, since questions seek factual materials (e.g., what do the numbers on camera lens mean?), there is little scope of discussion or disagreement among solution providers and hence there would be less solutions suggested and less number of feedbacks. Also, in subjective threads, the discussions can get heated due to disagreements with users posting inappropriate content such as abuses which are *junk* as they are not related to the discussion whereas in non-subjective threads, these situations are unlikely to happen. To explore the impact of dialog acts on a thread's subjectivity, we used eight dialog acts in thread posts as

proposed by Bhatia et al. [31] and used their presence in a thread as features for the subjectivity classification. The dialog acts are as follows: 1. Question, 2. Repeat Question, 3. Clarification, 4. Further Details, 5. Solution, 6. Negative Feedback, 7. Positive Feedback, 8. Junk. We implemented their classification model to identify the dialog acts in thread posts. We designed 8 features corresponding to the 8 dialog acts for a thread. Each feature represents the number of posts in a thread that belong to a given dialog act class.

3.2.3. Subjectivity lexicon based features

Subjective threads discuss subjective topics seeking private states such as opinions, emotions, evaluations whereas non-subjective threads seek factual information. This difference should result in differences in the vocabularies of these two types of threads. Subjective threads should contain words that are used to express subjectivity whereas non-subjective threads should either not have these words or have less number of these words. We call these words subjectivity clues in this paper. Hence, the frequency or term counts of subjectivity clues in a thread should be a good indicator of its subjectivity. We use a publicly available subjectivity lexicon compiled from MPQA corpus by [38] to get the subjectivity clues. The lexicon contains 8221 subjectivity clues. Some of the examples of subjectivity clues from the lexicon are abhor, abuse, bother, champion. We count the number of subjectivity clues in the title, initial post and all reply posts of a thread, normalize the subjectivity clue counts with the number of words in the corresponding element (title, initial post, reply posts) and use them as features. For a thread, we computed three lexicon features: Num-SubTitle, NumSubInit and NumSubReply. We calculated NumSub-Title and NumSubInit by normalizing the frequency counts of subjectivity clues in the title and the initial post, respectively, by their total number of words. For computing NumSubReply, we first calculated the normalized frequency counts of subjectivity clues for all the reply posts and then added all the normalized counts.

3.2.4. Sentiment features

¹ http://sentistrength.wlv.ac.uk/.

These features take into account the sentiment/emotion of a thread. We expect subjective threads to have posts with higher sentiments (as they expose private states) than the posts in nonsubjective threads. To calculate sentiment features for a thread. we compute sentiment strength of its individual posts. There are several resources available for calculating sentiment of text such as sentiment lexicons (e.g., SentiWordNet [39] and WordNet-Affect [40]) and sentiment analysis tools that are specifically developed for online social media text (e.g., SenticNet [41] and SentiStrength [42]). We use the SentiStrength tool¹ to compute strength of the sentiment expressed in posts. The tool is developed specifically to compute sentiment strength scores for short informal pieces of text common in social media such as forum posts, blog comments, etc. It uses lexical knowledge along with several heuristics (e.g., repeated letters, repeated punctuations, etc.) to calculate both positive as well as negative sentiment scores for a piece of text. This feature is desirable as the posts can express sentiments of multiple polarity and a single sentiment score (positive, negative or neutral) will not be able to capture the individual sentiments. For both polarities, the algorithm gives two types of scores for a piece of text (i) using the strongest sentiment-indicative word patterns and (ii) using all the sentiment-indicative word patterns and taking their average. Thus, we get four different sentiment strength scores for each post. We use the four sentiment strength scores for the initial post and averages of the four sentiment scores for all the reply posts as features, thus getting eight sentiment features for a thread.

Table 3

Statistics of the dataset.

Statistic	Trip-Advisor	Ubuntu
Total # threads	609	621
Total # posts	6591	3603
Total # users	1206	1786
Average thread length (in terms of # posts)	10.82	5.80
Average thread length (in terms of # words)	907	387.57
Average # users in a thread	1.98	3.41

4. Experiments

4.1. Data

To conduct our experiments, we used threads from two popular online forums: 1. **Trip Advisor–New York** that contains travel related discussions mainly for New York city² and 2. **Ubuntu Forums** that contains discussions related to the Ubuntu operating system.³ We used a publicly available dataset [1]. We randomly sampled 700 threads from both the datasets to conduct our experiments. Table 3 provides various statistics of the data. We selected these two forums because we wanted to evaluate our methods on two different genres of online forums. Ubuntu forums generally have technical discussions that tend to be non-subjective in nature whereas Trip Advisor is a travel related forum having discussions on topics like transport, hotels, restaurants, tourism, etc. that are generally non-technical in nature and hence tend to be subjective.

We hired two human annotators for tagging the threads. The annotators were asked to tag a thread as subjective if its topic of discussion is subjective or non-subjective if the topic of discussion is non-subjective. The annotators were provided with a set of instructions for annotations. The set contained definitions of subjective and non-subjective topics with examples and guidelines for doing annotations. The annotators were asked to annotate a sample of 20 threads from the dataset using the instruction set. Second, separate discussions were held between the authors and each annotator. Each annotator was asked to provide his arguments (for his annotations) and, in case of inconsistent arguments, he was educated through discussions. Next, he was given the full dataset for annotation.

The overall percentage agreement between the annotators and Kappa value for the Trip Advisor dataset were 87% and 0.713 respectively and for the Ubuntu dataset were 88.7% and 0.732 respectively, indicating substantial agreement between the taggers in both the cases. For our experiments, we used the data on which the annotators agreed. There were 412 subjective and 197 non-subjective threads in Trip Advisor dataset and 231 subjective and 390 non-subjective threads in Ubuntu dataset. The tagged dataset can be downloaded from the authors' website.⁴

4.2. Baseline

Lexical features such as *n*-grams and parts-of-speech tags have been shown to perform well for subjectivity analysis tasks. Many works have used these features for subjectivity classification [6,5,7]. In this work, we use classifiers built on these features as our baselines. We used the *Lingua-en-tagger* package from CPAN⁵ for part-of-speech tagging. The extracted features and their description is given in Table 4. The table describes feature generation on a sentence containing three words W_i , W_{i+1} and W_{i+2} . POS_i, POS_{i+1}

² http://www.tripadvisor.com/ShowForum-g60763-i5-

New_York_City_New_York.html.

³ http://ubuntuforums.org.

⁴ http://personal.psu.edu/pxb5080/dataSubj.html.

⁵ http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm.

6

ARTICLE IN PRESS

P. Biyani et al. / Knowledge-Based Systems xxx (2014) xxx-xxx

Table	4

Feature generation for sentence W_i W_{i+1} W_{i+2}. Uni, Bi, Tri and POS denote unigrams, bigrams, trigrams and parts-of-speech tags respectively.

Feature type	Generated feature
Uni	W_{i}, W_{i+1}, W_{i+2}
Uni + Bi	$W_{i}, W_{i+1}, W_{i+1}, W_{i}W_{i+1}, W_{i+1}W_{i+2}$
Uni + Bi + Tri	$W_{i}, W_{i+1}, W_{i+1}, W_{i}W_{i+1}, W_{i+1}W_{i+2}, W_{i}W_{i+1}W_{i+2}$
Uni + POS	$W_i, POS_i, W_{i+1}, POS_{i+1}, W_{i+2}, POS_{i+2}$
Uni + Bi + POS	W_i , POS_i , W_{i+1} , POS_{i+1} , W_{i+2} , POS_{i+2} , W_iW_{i+1} , W_iPOS_{i+1} , POS_iW_{i+1} , $W_{i+1}W_{i+2}$, $W_{i+1}POS_{i+2}$, $POS_{i+1}W_{i+2}$
Uni + Bi + Tri + POS	$ \begin{split} & W_{i}, \ POS_{i}, \ W_{i+1}, \ POS_{i+1}, \ W_{i+2}, \ POS_{i+2}, \ W_{i}W_{i+1}, \ W_{i}POS_{i+1}, \ POS_{i}W_{i+1}, \ W_{i+1}W_{i+2}, \ W_{i+1}POS_{i+2}, \ POS_{i+1}W_{i+2}, \ W_{i}W_{i+1}W_{i+2}, $

and POS_{i+2} are the parts-of-speech tags for the words W_i , W_{i+1} and W_{i+2} , respectively. For feature representation, we used term frequency as the weighting scheme (we empirically found it to be more effective than *tf-idf* and *binary* representations), and used minimum document frequency for a term to be included in the vocabulary as 3 (we experimented with minimum document frequency 3, 5 and 10 and 3 gave the best results).

We extracted the above features (Table 4) from the textual content of different structural units (title, initial post, reply posts) of the threads. First, we built a basic model where we used only the text of the titles (denoted by t) for classification. Then, we used the text of initial posts and reply posts. We experimented with the following four settings: title (t), initial post (I), title and initial post (t + I), entire thread (t + I + R).

4.3. Experimental setting

We used various supervised learning algorithms to perform our classification experiments. We experimented with Multinomial NaiveBayes, Support Vector Machines, Logistic regression, Bagging, Boosting and Decision Trees. Logistic regression gave the best results with our features whereas in case of the baseline lexical features, Multinomial NaiveBayes outperformed all the other classifiers. We used Weka data mining toolkit with default settings to conduct our experiments [43]. To evaluate the performance of our classifiers, we used macro-averaged precision, recall and F-1 measure. For a metric *M*, macro-average M_{mav} is calculated by taking weighted average of *M* for both subjective and non-subjective classes for each fold and then taking mean of the weighted averages across all the folds. For *n*-fold cross validation, M_{mav} is mathematically defined as follows:

$$M_{mav} = \frac{1}{n} \sum_{i=1}^{n} \frac{n_{s_i} M_{s_i} + n_{ns_i} M_{ns_i}}{n_{s_i} + n_{ns_i}}$$
(1)

where n_{s_i} and n_{ns_i} are the number of subjective and non-subjective threads in the test set in the *i*th fold. M_{s_i} and M_{ns_i} are the values of metric *M* for the subjective and the non-subjective classes, respectively, in the *i*th fold. We used n = 10 in our experiments. We use F-1 measure to compare performances of two classifiers. A naive baseline that classifies all the threads in the majority class will have a macro-averaged precision, recall and F-1 measure of 0.457, 0.676 and 0.545 respectively for Trip-Advisor and 0.394, 0.628 and 0.485 respectively for Ubuntu. We consider these values to be the lower bounds for any of our methods.

4.4. Classification results

4.4.1. Baseline results

Table 5 reports the results of the subjectivity classification obtained from different baselines. A total of 24 experiments (using the six types of features for the four settings (t, I, t+I, t+I+R)were conducted for both the datasets. From the table, we note that titles give fair estimate of thread's subjectivity and initial posts (I) provide a better estimate. Incorporating text from initial post and title (t + I) improves the performance slightly over the initial post (*I*) setting. Further, adding the text of reply posts (t + I + R) gives the best performance. This is expected as titles only contain some keywords related to the discussion topic whereas initial posts contain the entire problem of discussion and reply posts constitute a major portion of the discussion in the thread. We also note that unigrams + bigrams + POS and unigrams + bigrams consistently perform better than the other features for all the settings except for the title (t) setting where unigrams and unigrams + POS performed the best.

4.4.2. Performance of the proposed classification model

Table 6 reports the results of our classification model. We achieve an overall accuracy of 77.01%, a precision of 0.763 and

Table 5

Classification performance of different baseline features (Table 4) extracted from different structural components of the forum threads. *t*, *I* and *R* are title, initial post and set of all reply posts of a thread respectively. *U*, *B*, *T* and POS are unigrams, bigrams, trigrams and parts-of-speech tags respectively. Numbers in bold correspond to the best performing methods.

Trip-Advisor	t			Ι			t + I		t + I + R	t + I + R		
	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1s
U	0.618	0.644	0.625	0.662	0.665	0.664	0.671	0.673	0.672	0.703	0.716	0.706
U + B	0.56	0.586	0.565	0.713	0.718	0.715	0.700	0.704	0.702	0.738	0.747	0.723
U + B + T	0.627	0.55	0.564	0.703	0.658	0.669	0.697	0.655	0.666	0.721	0.732	0.723
U + POS	0.56	0.586	0.565	0.669	0.673	0.671	0.686	0.69	0.688	0.701	0.713	0.704
U + B + POS	0.606	0.616	0.610	0.704	0.711	0.704	0.701	0.709	0.704	0.733	0.741	0.71
U + B + T + POS	0.614	0.522	0.566	0.709	0.67	0.68	0.706	0.675	0.684	0.722	0.736	0.716
Ubuntu												
U	0.546	0.578	0.553	0.652	0.646	0.648	0.649	0.643	0.645	0.694	0.689	0.691
U + B	0.551	0.58	0.557	0.662	0.655	0.658	0.659	0.654	0.656	0.688	0.67	0.675
U + B + T	0.548	0.576	0.554	0.656	0.646	0.649	0.657	0.647	0.651	0.696	0.663	0.669
U + POS	0.626	0.647	0.633	0.644	0.638	0.64	0.649	0.641	0.644	0.694	0.688	0.69
U + B + POS	0.552	0.564	0.556	0.659	0.652	0.655	0.659	0.652	0.655	0.72	0.696	0.701
U + B + T + POS	0.551	0.557	0.554	0.646	0.631	0.636	0.64	0.63	0.633	0.705	0.657	0.662

P. Biyani et al. / Knowledge-Based Systems xxx (2014) xxx-xxx

Table 6

Classification	results.
----------------	----------

Metric	Trip-Advisor	Ubuntu
Classification accuracy	77.01%	70.05%
Precision	0.763	0.692
F1-Measure	0.764	0.684

an F-1 measure of 0.764 on the Trip-Advisor dataset and an overall accuracy of 70.05%, a precision of 0.692 and an F-1 measure of 0.684 on the Ubuntu dataset. We further analyze the classification performance of our classifier by analyzing its performance for the two classes. Table 7 reports precision, recall and F-1 measure for subjective and non-subjective classes for both the datasets. We observe that the classification performance for the subjective class is better than the non-subjective class for the Trip-Advisor dataset. This can be attributed to the significantly more number of subjective threads than non-subjective threads (refer to Section 4.1) in the Trip-Advisor dataset and hence more patterns for the classifier to learn for the majority (subjective) class leading to the better performance for that class. Similarly, for the Ubuntu dataset, we see a better performance for the non-subjective class whose number of threads are significantly more than that of the subjective class.

Next, we compare the performance of our classification model with the baselines. As can be seen from Table 7, our classification model outperforms the best performing baseline (U+B) for the t+I+R setting, refer to Table 5), thus outperforming all the 24 baselines, for the Trip-Advisor dataset. For the Ubuntu dataset, our model achieves an F-1 measure of 0.684 whereas the best performing baseline (U + B + POS for the t + I + R setting, refer to)Table 5) achieves an F-1 measure of 0.701. In this case, our model outperforms 21 out of the 24 baselines. The other two baselines that achieved a better performance than our model are unigrams (U) for the t + I + R setting and unigrams + POS (U + POS) for the t + I + R setting with an F-1 measure of 0.691 and 0.69 respectively. Thus, we see that we achieve classification performance which is similar to, and, in most cases, better than that obtained from the baseline features by using thread specific features which are much less in number (No. of baseline features is of the order of the size of the vocabulary whereas No. of features in our model = 25).

4.4.3. Relative performance of different types of features

In this subsection, we investigate the effect of different types of features used for the subjectivity classification task. We perform the classification experiment using only one type of feature at a time. Table 8 shows the relative performance of different types of features. We see that, for both the datasets, structural features gave the best performance which confirms our hypothesis that thread structure is a strong indicator of its subjectivity orientation. Lexicon-based and Sentiment features are the second best performing features, outperforming the dialog act features, for the Trip-Advisor forum whereas for the Ubuntu forum, dialog act features outperform the two types of features with sentiment features being the worst performing and Lexicon-based features being the third best performing features. This difference in the relative performance of Sentiment and Lexicon-based features across the two forums may be attributed to the difference in the nature of the two forums. Trip-Advisor is a non-technical forum where majority of discussions are subjective in nature and hence there are more number of subjectivity clues and sentiment indication patterns for the classifier to learn, whereas discussions in Ubuntu forum are technical and hence, usually, non-subjective in nature. Further, we use ensemble methods and a stacked classification approach. We make ensembles of the four classifiers corresponding to the four types of features. For a test instance, we calculate the final prediction of the ensemble using two methods: (1) averaging the confidences of the four classifiers (denoted by EnsembleAvg), (2) taking prediction of the most confident classifier out of the four classifiers (denoted by EnsembleMostConf). Next, we used a stacked classifier where the confidences of the four classifiers were provided as features for the second stage classifier. Finally, we build classifier using the combined feature set. The model is denoted by FeatureAll. We see that combined performance of all the features (FeatureAll model) is better than the performances of all the individual types of features. However, ensemble models perform worse than the combined feature model and models built on individual feature types. This suggests that the predictions of the four classifiers are quite different from each other. The conflicts between the classifiers in terms of their predictions result in a lower performance of the ensemble models. Similarly, the stacked classifier performed worse than the classifier built using all the features.

Table 7

Classification performance of the proposed model for subjective and non-subjective classes on the two datasets.

	Trip-Advisor			Ubuntu			
	Precision	Recall	F-1	Precision	Recall	F-1	
Subjective class	0.805	0.871	0.837	0.647	0.429	0.516	
Non-subjective class	0.675	0.558	0.611	0.718	0.862	0.783	
Overall	0.763	0.77	0.764	0.692	0.7	0.684	
Best performing baseline	0.738	0.747	0.723	0.72	0.696	0.701	

Table 8

Classification results for NYC and Ubuntu datasets obtained using different types of features. Numbers in bold correspond to the best performing methods.

Class	Trip-Advisor			Ubuntu		
	Precision	Recall	F-1	Precision	Recall	F-1
Structural	0.741	0.75	0.742	0.692	0.697	0.67
Dialog act	0.683	0.703	0.683	0.639	0.654	0.598
Subjectivity lexicon based	0.713	0.727	0.699	0.622	0.643	0.569
Sentiment	0.71	0.726	0.699	0.534	0.602	0.525
EnsembleAvg	0.644	0.681	0.662	0.646	0.65	0.648
EnsembleMostConf	0.631	0.678	0.65	0.6	0.627	0.613
Stacked classifier	0.74	0.749	0.741	0.678	0.688	0.663
AllFeatures	0.762	0.768	0.763	0.692	0.7	0.684

8

P. Biyani et al. / Knowledge-Based Systems xxx (2014) xxx-xxx

4.4.4. Most informative features

We study the importance of individual features by measuring the chi-squared statistic with respect to the class variable. Table 9 shows top 10 features, ranked by their chi-square values. From the table, we note that, for both the datasets, five out of six structural features (ThreadLength, NumPost, AvgPostLength, NumAuthor, InitPostLength) are among the top 10 most informative features which again confirms that a thread's structure is a strong indicator of its subjectivity. We note that the lexicon-based features and the sentiment features have relatively higher ranks in Trip Advisor dataset as compared to the Ubuntu dataset. We also note that, for Trip-Advisor, two of the three lexicon-based features (NumSub-Reply, NumSubInit) are among the top 10 features whereas for Ubuntu, only one lexicon-based feature (NumSubReply) is ranked among the top 10 features. This observation is consistent with our previous observation where we noted that sentiment and lexicon-based features performed relatively better in Trip-Advisor as compared to Ubuntu and can be attributed to the difference in the nature of the two forums as explained in the previous subsection. Among the lexicon-based features, NumSubReply is the most informative feature which suggests that, for a thread, reply posts are more helpful than initial post and title of the thread in identifying the thread's subjectivity. This is also manifested in case of sentiment features where features corresponding to reply posts (ReplySentiStrngPos, ReplySentiAvgNeg, etc.) are ranked higher than the corresponding features for the initial post (which are not in the top 10 list). These observations are consistent with the results we got from our baselines where we found that incorporating text from reply posts gave the best performance across all the features. We note that, for Ubuntu, there is one dialog act feature (NumSol) in the top 10 list, whereas for Trip-Advisor, none of the dialog act features are in the list.

Next, we analyze the behavior of users in the two forums in terms of starting a subjective or non-subjective thread. We used our subjectivity classifiers to predict labels of all the threads in the two datasets. Since most of the users have started very few threads, we take into account top 100 users for the two forums. We ranked the users according to the number of threads they have started and selected top 100 users from the ranked list for both the forums. For Trip-Advisor New York forum, the top 100 users have started 43 or more threads and the top user has started 397 threads. For Ubuntu forum, the top 100 users have started 28 or more threads and the top user has started 140 threads. Figs. 1 and 2 show the distribution of threads started by top 100 users in the subjective and non-subjective classes for Trip-Advisor New York and Ubuntu forums respectively. Each vertical bar corresponds to one of the top 100 users. Yellow and red portions in a bar represent the percentage of subjective and non-subjective threads started by the user represented by the bar. We note that most of the users started more subjective threads (than non-subjective) in Trip Advisor forum whereas in Ubuntu forum, most users started more number of non-subjective threads. We also

Table 9

Top 10 features ranked by chi-square values for the two datasets.

Trip-Advisor	Ubuntu
ThreadLength	ThreadLength
NumSubReply	NumPost
AvgPostLength	NumSubReply
NumPost	NumUser
NumUser	AvgPostLength
ReplySentiStrngPos	InitPostLength
ReplySentiAvgNeg	NumSol
InitPostLength	ReplySentiAvgNeg
ReplySentiAvgPos	ReplySentiStrngPos
NumSubInit	ReplySentiStrngNeg

observe that in Trip-Advisor forum, more users have higher percentage of non-subjective threads (than subjective threads) as compared to Ubuntu forum where a very few users have started more subjective threads than non-subjective threads.

4.4.5. Sources of error

Next, we conduct error analysis to better understand the results. We found that one of the main causes of errors in both the forums is related to thread structure. There are cases where thread starters initiate subjective topics of discussion and the threads either do not get responses from other users in the forum or the discussions are left incomplete resulting into smaller threads. In such cases, the threads are wrongly classified as nonsubjective as their structure is similar to that of non-subjective threads in terms of number of posts, number of users participating the discussion, thread length, etc. On the other hand, we found that in some cases, due to topic drift, non-subjective threads get more participation of users which changes their structure and tend to make them similar to subjective threads in terms of their structure. In such cases, the threads are wrongly classified as subjective.

5. Conclusions and future work

In this work, we proposed a supervised machine learning model for subjectivity classification of online forum threads. We used various novel thread-specific features in addition to lexicon-based and sentiment features for the classification task. We evaluated our model by comparing it with various state-of-the-art techniques used for subjectivity classification and showed that our model outperformed them in most of the cases. A major contribution of this work is the introduction of thread-specific features for subjectivity classification of online forum threads which significantly reduces the complexity of the learning model compared to that of the models built on lexical features without compromising the performance of the model. In future, we plan to investigate semisupervised and unsupervised learning for subjectivity classification of online forum threads. We also plan to use the subjectivity analysis to improve the search in online forums.

References

- S. Bhatia, P. Mitra, Adopting inference networks for online thread retrieval, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010, pp. 1300–1305.
- [2] P. Biyani, S. Bhatia, C. Caragea, P. Mitra, Thread specific features are helpful for identifying subjectivity orientation of online forum threads., in: Proceedings of the 24th International Conference on Computational Linguistics, 2012a, pp. 295–310.
- [3] P. Biyani, C. Caragea, A. Singh, P. Mitra, I want what i need!: analyzing subjectivity of online forum threads, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012b, pp. 2495–2498.
- [4] P. Biyani, C. Caragea, P. Mitra, Predicting subjectivity orientation of online forum threads, in: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics, 2013, pp. 109–120.
- [5] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2003, pp. 129–136.
- [6] B. Li, Y. Liu, E. Agichtein, Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2008, pp. 937–946.
- [7] N. Aikawa, T. Sakai, H. Yamana, Community qa question classification: is the asker looking for subjective answers or not?, IPSJ Online Trans 4 (2011) 160– 168.
- [8] J. Seo, W.B. Croft, D.A. Smith, Online community search using thread structure, in: Proceedings of ACM International Conference on Information and Knowledge Management, 2009, pp. 1907–1910.
- [9] H. Duan, C. Zhai, Exploiting thread structures to improve smoothing of language models for forum post retrieval, Advances in Information Retrieval, vol. 6611, Springer, Berlin Heidelberg, 2011, pp. 350–361.

P. Biyani et al. / Knowledge-Based Systems xxx (2014) xxx-xxx

- [10] N. Jindal, B. Liu, Opinion spam and analysis, in: Proceedings of the First ACM International Conference on Web Search and Data Mining, ACM, 2008, pp. 219–230.
- [11] E. Cambria, P. Chandra, A. Sharma, A. Hussain, Do not feel the trolls, in: Proceedings of the 3rd International Workshop on Social Data on the Web, ISWC, 2010.
- [12] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. 28 (2) (2013) 15–21.
- [13] R. Xia, C. Zong, X. Hu, E. Cambria, Feature ensemble plus sample selection: domain adaptation for sentiment classification, IEEE Intell. Syst. 28 (3) (2013) 10–18.
- [14] B. Liu, Sentiment analysis and subjectivity, Handbook Natural Language Process. 2 (2010) 627–666.
- [15] E. Cambria, A. Hussain, Sentic Computing: Techniques, Tools, and Applications, Springer, Dordrecht, Netherlands, 2012. ISBN: 978-94-007-5069-2.
- [16] E. Cambria, B. White, Jumping nlp curves: a review of natural language processing research, IEEE Comput. Intell. Mag. 9 (2) (2014) 48–57.
- [17] B. Li, Y. Liu, A. Ram, E. Garcia, E. Agichtein, Exploring question subjectivity prediction in community qa, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, pp. 735–736.
- [18] I. Gurevych, D. Bernhard, K. Ignatova, C. Toprak, Educational question answering based on social media content, in: Proceedings of the 14th International Conference on Artificial Intelligence in Education, 2009, pp. 133–140.
- [19] V. Stoyanov, C. Cardie, J. Wiebe, Multi-perspective question answering using the opqa corpus, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 923–930.
- [20] S. Somasundaran, T. Wilson, J. Wiebe, V. Stoyanov, Qa with attitude: exploiting opinion type analysis for improving question answering in on-line discussions and the news, in: Proceedings of the 1st International Conference on Weblogs and Social Media, 2007.
- [21] A. Balahur, E. Lloret, E. Boldrini, A. Montoyo, M. Palomar, P. Martínez-Barco, Summarizing threads in blogs using opinion polarity, in: Proceedings of the Workshop on Events in Emerging Text Types, Association for Computational Linguistics, 2009, pp. 23–31.
- [22] J. Wiebe, R. Bruce, T. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999, pp. 246–253.
- [23] R. Bruce, J. Wiebe, Recognizing subjectivity: a case study in manual tagging, Natural Language Eng. 5 (2) (1999) 187–205.
- [24] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, 2005, pp. 486–497.
- [25] F. Su, K. Markert, From words to senses: a case study of subjectivity recognition, in: Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 825–832.
- [26] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp. 976–983.

- [27] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: Proceedings of the Conference Empirical Methods in Natural Language Processing, 2008, pp. 127–135.
- [28] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
- [29] D. Ly, K. Sugiyama, Z. Lin, M. Kan, Product review summarization from a deeper perspective, in: Proceedings of the Joint Conference on Digital Libraries, 2011, pp. 311–314.
- [30] G. Di Fabbrizio, A. Aker, R. Gaizauskas, Summarizing online reviews using aspect rating distributions and language modeling, IEEE Intell. Syst. 28 (3) (2013) 28–37.
- [31] S. Bhatia, P. Biyani, P. Mitra, Classifying user messages for managing web forum data, in: Proceedings of the 15th International Workshop on the Web and Databases, 2012, pp. 13–18.
- [32] A. Hassan, V. Qazvinian, D.R. Radev, What's with the attitude? identifying sentences with attitude in online discussions, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2010, pp. 1245–1255.
- [33] Z. Zhai, B. Liu, L. Zhang, H. Xu, P. Jia, Identifying evaluative sentences in online discussions, in: Proceedings of the 25th AAAI Conference on Artificial Intelligence, 2011, pp. 933–938.
- [34] M.A. Walker, P. Anand, R. Abbott, J.E.F. Tree, C. Martell, J. King, That is your evidence?: classifying stance in online political debate, Decision Support Syst. 53 (4) (2012) 719–729.
- [35] R. Abbott, M. Walker, P. Anand, J.E. Fox Tree, R. Bowmani, J. King, How can you say such things?!?: Recognizing disagreement in informal political argument, in: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, 2011, pp. 2–11.
- [36] M. Jeong, C.-Y. Lin, G.G. Lee, Semi-supervised speech act recognition in emails and forums, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, vol. 3, 2009, pp. 1250–1259.
- [37] S. Joty, G. Carenini, C.-Y. Lin, Unsupervised modeling of dialog acts in asynchronous conversations, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011, pp. 1807–1813.
- [38] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, Language Resour. Evaluat. 39 (2) (2005) 165–210.
- [39] A. Esuli, F. Sebastiani, Sentiwordnet: a publicly available lexical resource for opinion mining, in: Proceedings of the Language Resources and Evaluation Conference, 2006, pp. 417–422.
- [40] C. Strapparava, A. Valitutti, Wordnet affect: an affective extension of wordnet., in: Proceedings of the Language Resources and Evaluation Conference, 2004, pp. 1083–1086.
- [41] E. Cambria, D. Olsher, D. Rajagopal, Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence (to appear), 2014.
- [42] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Assoc. Inform. Sci. Technol. 63 (1) (2012) 163–173.
- [43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The weka data mining software: an update, ACM SIGKDD Explorat. Newslett. 11 (1) (2009) 10–18.