Thread Specific Features Are Helpful For Identifying Subjectivity Orientation of Online Forum Threads

Prakhar Biyani[1] *Sumit Bhatia*[2] *Cornelia Caragea*[1] *Prasenjit Mitra*[1]
(1) College of Information Sciences and Technology, The Pennsylvania State University, USA
(2) Department of Computer Science, The Pennsylvania State University, USA
prakharbiyani@gmail.com, sumit@cse.psu.edu, ccaragea@ist.psu.edu,
pmitra@ist.psu.edu

ABSTRACT
Subjectivity analysis has been actively used in various applications such as opinion mining of customer reviews in online review sites, question-answering in CQA sites, multi-document summarization, etc. However, there has been very little 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 and non-subjective. 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.

KEYWORDS: Online Forums, subjecitivity, dialogue act.

1 Introduction

A large number of online forums in various domains (e.g., health, sports, travel, camera, laptops, etc.) exists today, containing huge volumes of user-generated data in the form of discussions between members. 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 available in other webpages, online forums are increasingly becoming very popular among internet users for discussing real life problems.

In this work, we *analyze subjectivity of online forum threads*. We identify two types of threads in an online forum: *subjective* and *non-subjective* and we model the subjectivity analysis task as a binary classification problem. Subjective threads discuss subjective topics that seek opinions, viewpoints, evaluations, and other private states of people, whereas non-subjective threads discuss non-subjective topics that seek factual information. Figure 1 shows a subjective thread from an online forum, Trip–Advisor New York. Figure 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*. To the best of our knowledge, previous work on subjectivity analysis has not tackled the problem of identifying subjectivity orientation of online threads.

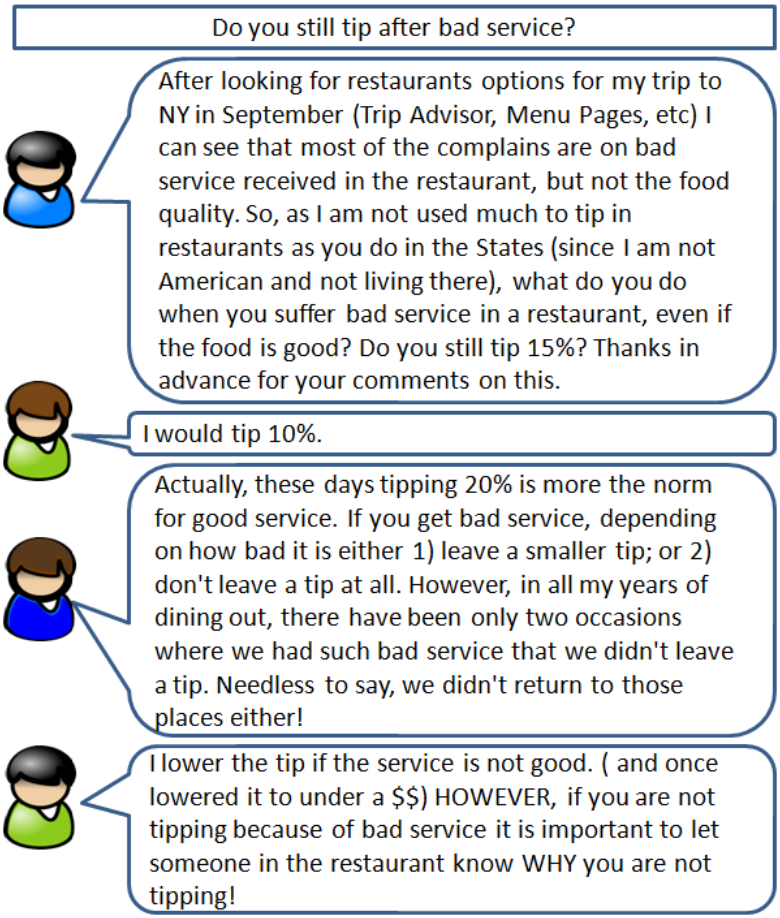

Figure 1: An example thread with subjective topic.

Figure 2: An example thread with non-subjective topic.

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 (Seo et al., 2009; Bhatia and Mitra, 2010; Duan and Zhai, 2011). 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.
- ***Abuse detection***: Online forums are informal in nature. Often, discussions in threads get heated with users getting engaged in abusive conversations. Forum administrators continuously monitor forums for such contents and remove them as they are against the community rules. These conversations are subjective in nature and hence can potentially be detected by analyzing threads for subjectivity.

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 (Yu and Hatzivassiloglou, 2003; Li et al., 2008a; Aikawa et al., 2011). A major issue with these features is their high dimensionality feature space and hence there is a risk of model overfitting especially with small training data. 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. 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.

1.2 Contributions

Our work has the following contributions:

1. We are the first to perform subjectivity analysis of online forum threads automatically.
2. We propose two new types of non-lexical features for subjectivity analysis of online fo-

rum threads: *structural* features and *dialogue act* features. Previous works on subjectivity analysis have mainly used lexical and syntactic features like n-grams, POS tags, subjectivity clues, etc. In this work, we empirically show that, for online forum threads, in addition to the traditionally used features, thread's structure and information about dialogue acts of its posts also help in analyzing thread's subjectivity.

3. We extensively evaluate our methods by comparing with various state-of-the-art baselines.
4. The dataset used in this paper for subjectivity analysis of online forum threads is being made publicly available for the research community [1].

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, sentiment analysis, question-answering, summarization, 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.).

2.1 Subjectivity Analysis

Wiebe et al. (1999) 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, etc. in the sentence. Bruce and Wiebe (1999) performed a case study of manual subjectivity tagging. Wiebe and Riloff (2005) 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 and Markert (2008) 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 (Mihalcea et al., 2007; Banea et al., 2008). 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. Banea et al. (2010) performed subjectivity classification in six different languages and showed that including multilingual information improves the classification performance across all the languages. Mukund and Srihari (2010) proposed a vector-space classification algorithm boosted by co-training for subjectivity classification of sentences in Urdu Language.

[1] http://www.personal.psu.edu/pxb5080/dataSubj.html

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 (Liu, 2010). Works in this area have mainly focused on online review sites for summarizing product reviews given by different users of those products (Hu and Liu, 2004; Ly et al., 2011). 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 (Bhatia et al., 2012). 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

Subjectivity analysis has also been used to improve question-answering in online communities and social media (Li et al., 2008b; Gurevych et al., 2009; Stoyanov et al., 2005; Yu and Hatzivassiloglou, 2003; Somasundaran et al., 2007). Yu and Hatzivassiloglou (2003) classify documents and sentences from news data into facts and opinions with the aim of improving answering of complex opinion questions. Stoyanov et al. (2005) 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. (2007) 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. (2008a) classify questions in Yahoo QA as subjective or objective using semi-supervised learning by utilizing the text of labeled questions and their unlabelled answers for learning subjectivity patterns. Gurevych et al. (2009) used an unsupervised lexicon based approach to classify questions as subjective or factoid (non-subjective). They manually build a lexicon of subjective words and word patterns from annotated questions and classify test questions based on a score calculated using the number of patterns present in them. 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. (2010) performed sentence-level attitude classification in online discussions to model user interaction that may be helpful in facilitating collaborations. Zhai et al. (2011) 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.

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 Figures 1 and 2 for examples of subjective and non-subjective threads, respectively). Based on the definitions of subjective and objective sentences given by Bruce and Wiebe (1999), 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 (ie., 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.

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*).

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. Table 1 lists the features used.

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.

Feature Name	Description
	Structural Features
InitPostLength	Total number of words in the initial post.
ThreadLength	Total number of words in the thread.
NumPost	Total number of posts in the thread.
NumUser	Total number of users in the thread.
AvgPostAuthor	Average number of posts by a user in the thread.
AvgLengthPost	Average number of words in a post in the thread.
	Dialogue Act Features
numQues	No. of *question* posts in the thread.
numRepeat	No. of *repeat question* posts in the thread.
numClar	No. of *clarification* posts in the thread.
numDetails	No. of *further details* posts in the thread.
numSol	No. of *solution* posts in the thread.
numNegFB	No. of *negative feedback* posts in the thread.
numPosFB	No. of *positive feedback* posts in the thread.
numJunk	No. of *junk* posts in the thread
	Subjectivity Lexicon-based Features
NumSubjTitle	No. of subjectivity clues in the title of the thread.
NumSubjInit	No. of subjectivity clues in the initial post of the thread.
NumSubjReply	No. of subjectivity clues in all the reply posts of the thread.
	Sentiment Features
InitSentiAvgPos	Positive sentiment score of initial post based on all the indicative word patterns in it.
InitSentiAvgNeg	Negative sentiment score of initial post based on all the indicative word patterns in it.
InitSentiStrngPos	positive sentiment score of initial post based on the strongest indicative word patterns in it.
InitSentiStrngNeg	Negative sentiment score of initial post based on the strongest indicative word patterns in it.
ReplySentiAvgPos	Average of positive sentiment scores of all the reply posts based on all the indicative word patterns in them.
ReplySentiAvgNeg	Average of Negative sentiment scores of all the reply posts based on all the indicative word patterns in them.
ReplySentiStrngPos	Average of positive Sentiment scores of all the reply posts based on the strongest word patterns in them.
ReplySentiStrngNeg	Average of Negative Sentiment scores of all the reply posts based on the strongest word patterns in them.

Table 1: Description of various features used for subjectivity classification.

3.2.2 Dialogue Act Features

Online forum threads have conversational nature and hence there are different types of dialogue acts (question, solution, feedback, etc.) expressed in thread posts (Bhatia et al., 2012;

Jeong et al., 2009; Joty et al., 2011). 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 dialogue 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 discusion whereas in non-subjective threads, these situations are unlikely to happen. To explore the impact of dialogue acts on a thread's subjectivity, we used eight dialogue acts in thread posts as proposed by Bhatia et al. (2012) and used their presence in a thread as features for the subjectivity classification. The dialogue 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 dialogue acts in thread posts. We designed 8 features corresponding to the 8 dialogue acts for a thread. Each feature represents the number of posts in a thread that belong to a given dialogue act class.

3.2.3 Subjectivity Lexicon Based Features

Subjective threads discuss subjective topics seeking private states such as opinions, emotions, evaluations, etc. 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 Wiebe et al. (2005) 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: NumSubTitle, NumSubInit and NumSubReply. We calculated NumSubTitle 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

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

non-subjective threads. To calculate sentiment features for a thread, we compute sentiment strength of its individual posts using the SentiStrength algorithm (Thelwall et al., 2012). We use the implementation of the algorithm available at http://sentistrength.wlv.ac.uk/. The algorithm 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. SentiStrength calculates 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 (see Table 1).

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 city [2] and 2. **Ubuntu Forums** that contains discussions related to the Ubuntu operating system [3]. We used a publicly available dataset [4] (Bhatia and Mitra, 2010). We randomly sampled 700 threads from both the datasets to conduct our experiments. Table 2 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.

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

Table 2: Statistics of the Dataset

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 annotations for each dataset were conducted in three stages. First, the annotators were asked to annotate a sample of 20 threads from the

[2] http://www.tripadvisor.com/ShowForum-g60763-i5-New_York_City_New_York.html
[3] http://ubuntuforums.org
[4] http://www.cse.psu.edu/ sub194/datasets/ForumData.tar.gz

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. [5].

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 (Li et al., 2008a; Yu and Hatzivassiloglou, 2003; Aikawa et al., 2011). In this work, we use classifiers built on these features as our baselines. We used the *Lingua-en-tagger* package from CPAN[6] for part-of-speech tagging. The extracted features and their description is given in Table 3. 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} 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).

Feature type	**Generated feature**
Uni	W_i, W_{i+1}, W_{i+2}
Uni+Bi	$W_i, W_{i+1}, W_{i+1}, W_iW_{i+1}, W_{i+1}W_{i+2}$
Uni+Bi+Tri	$W_i, W_{i+1}, W_{i+1}, W_iW_{i+1}, W_{i+1}W_{i+2}$, $W_iW_{i+1}W_{i+2}$
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	$W_i, POS_i, W_{i+1}, POS_{i+1}, W_{i+2}, POS_{i+2}, W_iW_{i+1}, W_iPOS_{i+1}, POS_iW_{i+1}, W_{i+1}W_{i+2}$, $W_{i+1}POS_{i+2}, POS_{i+1}W_{i+2}, W_iW_{i+1}W_{i+2}, W_iW_{i+1}POS_{i+2}, W_iPOS_{i+1}W_{i+2}$, $POS_iW_{i+1}W_{i+2}, W_i, POS_{i+1}POS_{i+2}, POS_iW_{i+1}POS_{i+2}, POS_i, POS_{i+1}W_{i+2}$

Table 3: 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.

We extracted the above features (Table 3) 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).

[5] http://www.personal.psu.edu/pxb5080/dataSubj.html

[6] http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm

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 (Hall et al., 2009). 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:

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

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 4 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 5 reports the results of our classification model. We achieve an overall accuracy of 77.01%, a precision of 0.763 and 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 6 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

Trip–Advisor

	t			I			t+I			t+I+R		
	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1
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

	t			I			t+I			t+I+R		
	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1	Pr.	Re.	F-1
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

Table 4: Classification performance of different baseline features (Table 3) 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.

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 6, our classification model outperforms the best performing baseline (U+B for the t+I+R setting, refer to Table 4), 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 4) 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.)

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

Table 5: Classification results.

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

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 7 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 dialogue act features, for the Trip–Advisor forum whereas for the Ubuntu forum, dialogue 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, the combined performance of all the features is better than the performances of all the individual types of features.

	Trip–Advisor			Ubuntu		
Class	**Precision**	**Recall**	**F-1**	**Precision**	**Recall**	**F-1**
Structural	0.741	0.75	0.742	0.692	0.697	0.67
Dialogue 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
All	0.762	0.768	0.763	0.692	0.7	0.684

Table 7: Classification results for NYC and Ubuntu datasets obtained using different types of features.

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 8 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 (NumSubReply, 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 dialogue act feature (NumSol) in the top 10 list, whereas for Trip–Advisor, none of the dialogue act features are in the list.

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

Table 8: Top 10 features ranked by chi-square values for the two datasets.

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 semi-supervised 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

Aikawa, N., Sakai, T., and Yamana, H. (2011). Community qa question classification: Is the asker looking for subjective answers or not? *IPSJ Online Transactions*, 4:160–168.

Banea, C., Mihalcea, R., and Wiebe, J. (2010). Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bhatia, S., Biyani, P., and Mitra, P. (2012). Classifying user messages for managing web forum data. In *15th International Workshop on the Web and Databases(WebDB)*.

Bhatia, S. and Mitra, P. (2010). Adopting inference networks for online thread retrieval. In *AAAI 2010, Atlanta, Georgia, USA, July 11-15*, pages 1300–1305.

Bruce, R. and Wiebe, J. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

Duan, H. and Zhai, C. (2011). Exploiting thread structures to improve smoothing of language models for forum post retrieval. *Advances in Information Retrieval*, pages 350–361.

Gurevych, I., Bernhard, D., Ignatova, K., and Toprak, C. (2009). Educational question answering based on social media content. In *Proc. of the 14th International Conf. on Artificial Intelligence in Education*, pages 133–140.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hassan, A., Qazvinian, V., and Radev, D. R. (2010). What's with the attitude? identifying sentences with attitude in online discussions. In *EMNLP 2010*, pages 1245–1255. ACL.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *SIGKDD 2004*, pages 168–177. ACM.

Jeong, M., Lin, C.-Y., and Lee, G. G. (2009). Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1250–1259, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joty, S., Carenini, G., and Lin, C.-Y. (2011). Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 1807–1813. AAAI Press.

Li, B., Liu, Y., and Agichtein, E. (2008a). Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 937–946, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, B., Liu, Y., Ram, A., Garcia, E., and Agichtein, E. (2008b). Exploring question subjectivity prediction in community qa. In *SIGIR 2008*, pages 735–736. ACM.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 978–1420085921.

Ly, D., Sugiyama, K., Lin, Z., and Kan, M. (2011). Product review summarization from a deeper perspective. In *JCDL 2011*, pages 311–314.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *ACL*.

Mukund, S. and Srihari, R. K. (2010). A vector space model for subjectivity classification in urdu aided by co-training. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 860–868, Stroudsburg, PA, USA. Association for Computational Linguistics.

Seo, J., Croft, W. B., and Smith, D. A. (2009). Online community search using thread structure. In *CIKM 2009*, pages 1907–1910, New York, NY, USA. ACM.

Somasundaran, S., Wilson, T., Wiebe, J., and Stoyanov, V. (2007). Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *ICWSM 2007*.

Stoyanov, V., Cardie, C., and Wiebe, J. (2005). Multi-perspective question answering using the opqa corpus. In *EMNLP 2005*, pages 923–930. ACL.

Su, F. and Markert, K. (2008). From words to senses: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 825–832, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

Wiebe, J., Bruce, R., and O'Hara, T. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *ACL 1999*, pages 246–253. ACL.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, pages 486–497.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210. 10.1007/s10579-005-7880-9.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhai, Z., Liu, B., Zhang, L., Xu, H., and Jia, P. (2011). Identifying evaluative sentences in online discussions. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.