Finding Thoughtful Comments from Social Media

Swapna Gottipati, *Jing Jiang*
School of Information Systems, Singapore Management University, Singapore
swapnag.2010@smu.edu.sg, jingjiang@smu.edu.sg

ABSTRACT
Online user comments contain valuable user opinions. Comments vary greatly in quality and detecting high quality comments is a subtask of opinion mining and summarization research. Finding attentive comments that provide some reasoning is highly valuable in understanding the user's opinion particularly in sociopolitical opinion mining and aids policy makers, social organizations or government sectors in decision making. In this paper we study the problem of detecting thoughtful comments. We empirically study various textual features, discourse relations and relevance features to predict thoughtful comments. We use logistic regression model and test on the datasets related to sociopolitical content. We found that the most useful features include the discourse relations and relevance features along with basic textual features to predict the comment quality in terms of thoughtfulness. In our experiments on two different datasets, we could achieve a prediction score of 79.37% and 73.47% in terms of F-measure on the two data sets, respectively.

KEYWORDS: Opinion mining, Information Extraction, Text Classification.

Proceedings of COLING 2012: Technical Papers, pages 995–1010,
COLING 2012, Mumbai, December 2012.

1 Introduction

In recent years sentiment analysis and opinion mining has been extensively studied in natural language processing (Pang and Lee, 2008), largely because of the availability of a huge amount of opinionated text in online product reviews, blogs, social networking sites, forums, etc. Most work on opinion mining is about mining reviews of products and services, and the focus of these studies has been on a few important subtasks such as sentiment classification (Pang et al., 2002; Cui et al., 2006; Jiang et al., 2011) and opinion extraction (Popescu and Etzioni, 2005; Choi et al., 2006; Wu et al., 2009).

When we go beyond product review mining and consider the general problem of opinion mining from social media, many other subtasks and challenges arise. One of them is how to assess the quality of online comments and select high quality ones for further analysis and summarization. Consider the problem of mining the comments found in online social media towards a political speech such as Obama's State of the Union address. By restricting the search space to politically active blogs and forums and by using queries such as "*State of the Union*," likely we are able to retrieve highly relevant comments to the speech. However, not every comment contains valuable insight into the public's opinions regarding the sociopolitical issues addressed in the speech. Comments such as "*innovate and innovation appeared 10 times*" and "*To him..investment = more deficit spending*" are subjective but lack thoughtful explanations to support their claims. In comparison, comments like "*You want to really drive innovation, job growth and entrepreneurs? Make education, health care and retirement less of a burden on the average family, adopt more socialist policies like Norway (paid for by higher taxes, especially on the rich), and watch our standard of living rise at last!*" provide much more insightful reasoning that government policy makers may find highly valuable in understanding the general public's sentiment. Hence, we define thoughtfulness as insightful reasoning with relevance to the issues discussed in the article. So, detection of the comments with reasoning or justification is the focus of our task. Thoughtfulness is accessed only for relevant comments. This problem of finding *thoughtful* comments from social media is what we study in this paper. Formally, a *thoughtful comment* is relevant to the target document and has a justification or an argument to the issue(s) in the target document. It is particularly important for sociopolitical opinion mining because of the complexity of sociopolitical issues.

Intuitively, finding thoughtful comments is related to measuring text quality. There has been a large body of previous work on text quality prediction, but the methods are usually applied to student essays (Attali and Burstein, 2006) and news articles (Tang et al., 2003), (TREC novelty track 2003 and 2004). In social media mining, there have also been a number of studies on finding high quality reviews (e.g. (Kim et al., 2006; Agichtein et al., 2008)), but the focus in not on finding thoughtful comments, which requires us to look for reasoning in text. Presumably, a thoughtful comment should be logically well organized and coherent. We therefore hypothesize that discourse relations such as comparison, expansion and contingency will play an important role in finding thoughtful comments. A well organized comment is not always thoughtful. Comments such as, "*He is a great speaker as he writes the speech by himself and also delivers it very confidently*" are justified but are not relevant to the issues discussed in the article. Hence we hypothesize that relevance factors play an important role in detecting thoughtful comments.

We adopt a supervised learning approach and consider a diverse set of factors ranging from lexical usage to discourse relations, all derived from the textual content of comments. Many

of the factors we consider are based on the study by (Pitler and Nenkova, 2008). In addition, we also consider a relevance feature because of the nature of our problem. We construct two data sets to evaluate the various factors, one based on Singapore Prime Minister's National Day Rally speech[1] and the other on US President's State of the Union address[2].

Empirical evaluation reveals that discourse relations and relevance scores together with the standard textual features aid in better prediction of thoughtful comments. We could achieve a prediction score of 79.37% and 73.47% in terms of F-measure on the two data sets, respectively. We further tested our model across data collections. Our test result shows that the model with combined textual, discourse and relevance features still performs better than textual features alone.

The rest of the paper is organized as follows. We present the related work in Section 2. In Section 3, we formally define our problem and give an overview of our solution. We present the various features we consider in Section 4. The data set details are presented in Section 6. Evaluation and results are presented in Section 7. Finally we conclude and discuss some future work.

2 Related work

Our work is related to a large body of literature on measuring text quality in NLP, but our problem has some essential differences. The main difference is that in traditional sense, high-quality text should be grammatical, coherent and readable. For online comments, we focus more on the insightfulness or thoughtfulness of comments.

Many recent studies examined the challenges on the quality of comments. (Kim et al., 2006) studied how to predict the helpfulness of product reviews. They found that a helpful review should describe the features of the products and the pros/cons of the features. A more elaborative review that provides the complete details of the product is more likely to be considered high quality. Our problem is more general and the comments are not necessarily about products. Moreover, comments on sociopolitical articles need not elaborate on all the issues in the article. Therefore products and their features are not relevant to our problem. Another study by (Ghose and Ipeirotis, 2011) on review helpfulness looked into factors related to the reviewer, such as reviewer characteristics and reviewer history. In our work, we focus on features observed from the text only. Other social factors such as a commenter's profile or past behavior are complementary to our method.

Our work is also related to opinion retrieval (Zhang and Ye, 2008; Huang and Croft, 2009; Macdonald et al., 2009), which aims at automatically finding attitudes or opinions about specific targets, such as named entities, consumer products or public events. In most existing work on opinion retrieval, only relevance and subjectivity are considered, whereas we propose that quality in terms of thoughtfulness is also an important factor.

Work on measuring quality of social media content considers not only the quality of the content itself but also its authority in the social network through the author's authority, its popularity, etc. (Hsu et al., 2009). Several researchers explored the social network together with the content of the reviews to predict the review quality. (Bian et al., 2009) proposed a mutual reinforcement learning framework to simultaneously predict content quality and user reputation, whereas (Lu et al., 2010) proposed a linear regression model with various social contexts

[1] http://www.pmo.gov.sg/
[2] http://www.whitehouse.gov/

for review quality prediction. They combined textual and social context information to evaluate the quality of individual reviewers and to assess the quality of the reviews. We do not consider these factors as we want to focus on textual cues first. These additional features can be factored in as an independent step. Similar line of work can be seen by (Chen et al., 2011; Liu et al., 2007; Bian et al., 2009).

Our work is similar to (Amgoud et al., 2011) where they introduced argument analysis together with opinion. In their task, properties of a person or product (honesty, rigor, friendliness, etc.) are treated as arguments. The task is oriented towards aggregating features related to the product and supporting arguments to detect polarity. The task we address in this paper is quite different from their work in two main aspects. Firstly, for sociopolitical issues, the policy makers look for insightful reasoning text to understand the public sentiment in which case, properties are insufficient. Secondly, we study the attentiveness of the comments but not the polarity. Polarity orientation is a separate task which can be studied individually.

3 Problem Definition and Overview of Solution

We assume the following general definition of the task of finding thoughtful comments: Given a comment c made with respect to a *target* document d, we would like to determine whether c is a thoughtful comment. We will explain in Section 6 how we instruct the human annotators to label thoughtful comments. Generally speaking, a thoughtful comment is relevant to the target document and has a justification or an argument to the issue(s) in the target document.

While the task defined above is certainly not trivial, and theoretically speaking one would need a deep understanding of both the target document and the comment as well as relevant world knowledge to be able to judge whether a comment is thoughtful. Here we take an empirical approach and test whether features defined at lexical, syntactic, discourse levels and relevance factors have correlations with the thoughtfulness of comments and whether they can be used to achieve decent prediction accuracy. A large portion of the linguistic features we consider are inspired by existing work on measuring text quality. Indeed, at first glance our problem may appear to be the same as measuring text quality. News articles and student essays are formal and usually lengthier, whereas online comments are usually much shorter and less formal. In traditional sense, high-quality text should be grammatical, coherent and readable. Our problem seems to be text quality assessment which is defined as above, but we are not looking for grammar and readability. Instead we look for insightful reasoning with relevance to the article, as mostly user comments are not formal in social media.

We adopt a supervised learning approach to our problem. Specifically, we assume that we have a set of N training examples $\{(d_i, c_i, y_i)\}_{i=1}^{N}$, where d_i is a target document, c_i is a comment on d_i, and y_i is a binary label indicating whether c_i is a thoughtful comment with respect to d_i. With a set of feature functions, we can represent (d_i, c_i) by a feature vector $\mathbf{x}(d_i, c_i)$ (which we refer to as $\mathbf{x}_i$). We can then use standard classification algorithms to learn a classifier from $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. This classifier can be used to predict y for any unseen pair of d and c. In the following sections, we will explain in detail the features we consider and the classification algorithm we use.

4 Features

There have been many studies on measuring text quality and many features have been proposed to capture text quality. As mentioned previously our methodology is based on existing work on this topic. In particular, we follow the work by (Pitler and Nenkova, 2008). They

conducted a systematic study on text quality using various linguistic features and Wall Street Journal articles. Based on the major findings of their study, we take the following features as our starting point.

4.1 Structural Feature

Structural features are generated from the comment structure. (Pitler and Nenkova, 2008) tested various structural features including the average number of characters per word, the average number of words per sentence, the maximum number of words per sentence, and article length. According to their findings, article length was the only significant factor with good correlation with text quality. Hence, we define our first feature F_1 as the number of words in the comment.

4.2 Lexical Feature

Lexical features aim to capture the lexical usage of a piece of text compared to some reference corpus. (Pitler and Nenkova, 2008) used a lexical feature based on unigram language models, which provide a principled way to statistically model text. Specifically, it is assumed that there is a reference corpus that represents high quality text, e.g. a corpus of Wall Street Journal articles. A unigram language model, denoted as θ_r, can be estimated from this reference corpus. The lexical feature is defined as the log likelihood of the comment based on θ_r, calculated as:

$$\sum_w n(w,c) \log P(w|\theta_r), \tag{1}$$

where $P(w|\theta_r)$ is the probability of word type w according to θ_r, and $n(w,c)$ is the number of times word type w appears in comment c. We call this feature F_2.

4.3 Syntactic Feature

(Pitler and Nenkova, 2008) examined various syntactic features including the average parse tree height, the average number of noun phrases per sentence, the average number of verb phrases per sentence and the average number of subordinate clauses per sentence. They found that the average number of verb phrases per sentence was a useful feature with high correlation with text quality. So, the third feature F_3 we use for our study is the average number of verbs per comment.

We also experimented with other syntactic features like average number of noun phrases and noun to verb ratio calculated from the user's comments. We found that the average number of verbs per comment had the highest correlation with comment quality, and therefore we do not consider these other syntactic features in our experiments.

4.4 Discourse Features

Previous study by (Pitler and Nenkova, 2008) found that discourse relations were also correlated with text quality. Discourse relations aim to capture textual structures such as comparison, elaboration, cause-effect explanations and examples. They are considered key for the ability to properly interpret or produce discourse. For the problem of finding thoughtful comments, we hypothesize that discourse relations may play an even larger role because a logical argument will likely rely on coherently connecting textual units through discourse relations.

Discourse relations are divided into four major semantic classes (Prasad et al., 2008):

Expansion covers those relations where the second argument expands the discourse of the first argument or move its narrative forward.

Comparison relations highlight prominent differences between the two arguments of a relation.

Contingency is marked when one of the situations described in an argument causally influences the other argument.

Temporal relations are marked when the situations described in the arguments are related temporally, either synchronously or sequentially.

It has been found that a large portion of discourse relations can be detected through connectives, i.e. cue words and phrases (Pitler et al., 2008). We use a list of such connectives compiled by (Prasad et al., 2008) and study the statistics of our corpus to discover the discourse relations. Table 1 shows that the statistics of discourse relations in our dataset.

DR Class	Singapore	US
COMPARISON	44.50%	45.87%
EXPANSION	16.75%	15.47%
CONTINGENCY	38.75%	38.66%
TEMPORAL	5.70%	5.72%

Table 1: Discourse relations statistics in our corpus

Table 1 shows that the frequency of temporal relations is low in our corpus. It is not surprising because for many online comments the arguments are not temporal. Hence, we ignore the temporal class for the rest of our paper, and restrict our attention to only the other three major classes, namely, expansion, comparison and contingency. At the same time, many relations are explicit and can be discovered using the connectives/words as used in other applications (Saito et al., 2006).

The full list of the phrases for each class are shown in Table 2. This list is collated from (Prasad et al., 2008). We observed that some words are ambiguous: 'if', 'and', 'but', 'as' etc. In our study, such words are counted only once while combining the classes for the feature generation.

In the earlier work by (Pitler and Nenkova, 2008), the Penn Discourse Treebank was used for computing the discourse features. For us, we take a simpler approach and count the number of discourse relations in a comment. This becomes the F_4 in our experiments.

4.5 Relevance Feature

One of the important differences between our problem and standard text quality assessment is that the quality of a comment also relies on its relevance to the target of the comment. In our problem definition, the target is also a piece of text. For example, consider comments made to Obama's State of the Union speech. A comment such as "We are very lucky to live in the USA. I always have and always will support our president" is not directly related to any issue addressed by Obama in his speech, and therefore is not considered to be a thoughtful comment. Hence, for thoughtful comment prediction, we also consider a relevance feature

Class	Phrases
COMPARISON	although, as though, but, by comparison, even if, even though, however, nevertheless, on the other hand, still, then, though, while, yet, and, meanwhile, in turn, next, ultimately, meantime, also, as if, even as, even still, even then, regardless, when, by contrast, conversely, if, in contrast, instead, nor, or, rather, whereas, while, yet, even after, by contrast, nevertheless, besides, much as, as much as, whereas, neither, nonetheless, even when, on the one hand indeed, finally, in fact, separately, in the end, on the contrary, while
EXPANSION	accordingly, additionally, after, also, although, and, as, as it, as if besides, but, by comparison, finally, first, for example, for one thing, however, in addition, in fact, in other words, in particular, in response, in sum, in the end, in turn, incidentally, indeed, instead, likewise, meanwhile, nevertheless, on the one hand, on the whole, overall, plus, separately, much as, whereas, ultimately, as though, rather, at the same time, or, then, if, in turn, furthermore, in short, turns out, while, yet, that is, so, what's more as a matter of fact, further, in return, moreover, similarly, specifically,
CONTINGENCY	and, when, typically, as long as, especially if, even if, even when, if, so, when if only, lest,once, only if, only when, particularly if, at least partly because, especially as, especially because, especially since, in large part because, just because, largely because, merely because, not because, not only because, particularly as, particularly because, particularly since, partly because, because, simply because, since, then, after, one day after, reportedly after, consequently, mainly because, for, thus, apparently, in the end, in turn, primarily because, largely as a result, as, because, therefore, only because, particularly, when, so that, thereby, presumably, hence, as a result, if and when, unless, until, in part because, now that, perhaps because, only after, accordingly,

Table 2: Discourse relations

in addition to text quality features. There are many ways to measure relevance, and here we choose KL-divergence score, a principled measure for relevance commonly used in information retrieval tasks.

The KL-divergence score between a comment c and a target document d is defined as the KL-divergence between the unigram language models θ_c and θ_d estimated from c and d, respectively:

$$Div(\theta_c||\theta_d) = \sum_{w \in \mathcal{V}} p(w|\theta_c) \log \frac{p(w|\theta_c)}{p(w|\theta_d)}, \tag{2}$$

where $\mathcal{V}$ is the vocabulary.

KL-divergence using only nouns: We hypothesize that the topical relevance between a comment and its target relies more on the overlap of nouns in the two pieces of text. Hence, we consider another KL-divergence measure using only nouns in c and d. Specifically, we use the unigram language models that are defined over nouns only. Let θ_c^N and θ_d^N denote the two language models. We have

$$Div(\theta_c^N||\theta_d^N) = \sum_{w \in \mathcal{V}} p(w|\theta_c^N) \log \frac{p(w|\theta_c^N)}{p(w|\theta_d^N)}. \tag{3}$$

We define the fifth feature that uses only nouns, F_5=-$Div(\theta_c^N||\theta_d^N)$ and the sixth feature which is based on all words, F_6=-$Div(\theta_c||\theta_d)$.

KL-Divergence between comment and average comment: (Lu et al., 2010) proposed conformity features in which the comment c is compared with other comments by looking at the KL-divergence between the unigram model of the comment c, and unigram model of an "average" comment that contains the text of all comments for an article. We did a preliminary analysis to study the impact of conformity on the quality. We found that this KL-divergence score has low correlation with comment quality on our data sets, and therefore we do not consider it in the rest of this paper.

5 Logistic Regression

So far we have introduced six features, which are summarized in Table 3.

Feature Set	Description
F_1	Comment length
F_2	Comment likelihood
F_3	Average number of verbs
F_4	Number of discourse relations
F_5	Relevance score using nouns only
F_6	Relevance score using all words

Table 3: Full feature set for comment representation

Once features are defined, we can use a classification algorithm to learn a model from the training data and apply the model to unseen data for thoughtful comment prediction. In this paper we use logistic regression as our classification algorithm.

As we have pointed out earlier, a comment c together with its target document d can be represented by a feature vector $\mathbf{x}$. A logistic regression classifier models the probability of observing a discrete label y for a given $\mathbf{x}$ as follows:

$$p(y|\mathbf{x};\mathbf{w}) = \frac{1}{Z(\mathbf{x},\mathbf{w})}\exp(\mathbf{w}_y^T\mathbf{x}), \tag{4}$$

where

$$Z(\mathbf{x},\mathbf{w}) = \sum_{y\in\mathscr{Y}}\exp(\mathbf{w}_y^T\mathbf{x}).$$

Here $\mathbf{w}$ is a weight matrix and $\mathbf{w}_y$ is the weight vector corresponding to class y, and $\mathscr{Y}$ is the set of class labels.

Given training data $\{\mathbf{x}_i, y_i\}_{i=1}^N$, we learn a weight matrix by minimizing the following objective function:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\left[\lambda\|\mathbf{w}\|^2 - \frac{1}{N}\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i;\mathbf{w})\right], \tag{5}$$

where $\|\mathbf{w}\|^2 = \sum_{y\in\mathscr{Y}}\|\mathbf{w}_y\|^2$ and λ is a regularization parameter that is empirically set.

6 Data Set

6.1 Data Collection

Our objective is to study how the thoughtfulness of a comment is reflected in the various linguistic factors including discourse relations and relevance factors to the article. As we have mentioned earlier, measuring the thoughtfulness of a comment is especially important for sociopolitical opinion mining. We therefore collected two data sets in this domain for our evaluation. We first acquired the following two political speeches: (1) Singapore Prime Minister's National Day Rally Speech in 2010. (2) US President's State of the Union address in 2011. We further broke down each speech into several segments based on topical boundaries. The topics of each speech are listed in Table 4.

Speech	**Topics**
Singapore	Economy & Productivity, Immigration, Congestion, Housing, Education, National Service (NS), Singapore Spirit, Founding Fathers, Youth Olympics
US	Economy, Innovation & Research, Education, Rebuild, Spending & Taxes, Debts, Military

Table 4: Topics in the two speeches.

To collect an unbiased sample of comments for each speech, we use two search queries ("national day rally speech 2010" and "president state union address 2011") and Google API to obtain a list of top 50 URLs. We further manually selected URLs from online forums and blogs. We cleaned the data and removed short comments with no more than two words.

For F_2, the lexical feature, we need a suitable reference corpus. For the Singapore data set, we collected 1200 news articles from AisaOne.com to form our reference corpus. For the US data set, we used a set of 1358 New York Times articles to form the reference corpus.

6.2 Annotation

We engaged two human annotators to judge the comments we had collected. The annotators were asked to judge (1) whether a comment was relevant to each segment of its corresponding speech, and if so, (2) whether the comment was a thoughtful one. In other words, we treat each segment of a speech as a target document. For each pair of a comment c and a target document (i.e., a speech segment) d, we obtained two binary labels: a label z that indicates whether c is relevant to d, and a label y that indicates whether c is a thoughtful comment with respect to d.

To judge whether a comment was thoughtful, the annotators were asked to use the following criteria:

1. Is the comment a mere repetition or a rephrase of the speech text? For example, "PM says that we should stay open for the foreigners" is a repetition of the text from the article in passive voice. Such comments are relevant to the article, but not insightful.

2. Does the comment contain opinions of the commenter? For example, "Eliminating the deficit-Im sure this makes Mitch happy" is about the topic "Spending and Taxes" but without any insightful opinion. Such comments are relevant but not insightful.

3. Does the commenter provide argument to support her opinion? For example, "All this deficit crap reminds me of when Reagan ran for president ; how the deficit was terrible etc, etc, and after he got elected he ran up the biggest deficit ever. This was mostly due to spending on the military and tax cuts for the rich. This was even after he slashed domestic spending. If the US would wake up to the fact that we can't afford the wars, we might be able to move forward."

In total, the human judges annotated 1350 pairs of issue-comments for PM speech and 1150 pairs of issue-comments for Obama Speech. Since the annotation is still subjective, we calculated the inter-annotator agreement level using Cohen's Kappa coefficient. Cohen's kappa on quality is 0.8965 for all comments. On relevance the kappa is 0.7355 for all comments. We use the judgment from the judge who is stricter as our ground truth. The statistics of the labeled data are shown in Table 5.

Comment type	**Singapore**	**US**
Thoughtless	63.35%	68.25%
Thoughtful	36.65%	31.75%

Table 5: Comment statistics for both articles

7 Experiments

To check whether the features we have defined correlate with the thoughtfulness of comments based on human judgement, we first compute the Pearson correlation coefficients for all the features summarized in Table 3. The results are shown in Table 6. We observe that all features are positively correlated with the thoughtfulness of comments.

Feature	**Singapore**	**US**
F_1	0.3744	0.3594
F_2	0.3782	0.3755
F_3	0.3639	0.3911
F_4	0.3913	0.3554
F_5	0.1606	0.2437
F_6	0.1191	0.2146

Table 6: Pearson correlation coefficients between the features and the thoughtfulness of comments.

In the remaining of this section, we show our experimental results that answer the following questions:

RQ1: Does the KL-divergence relevance score based on nouns work better than the KL-divergence score based on all words?

RQ2: Which discourse relations have bigger impact on the performance?

RQ3: Which combination of various features gives the best prediction of thoughtfulness?

For all the experiments below, we use the standard precision, recall and F-score as our performance measures.

7.1 Relevance Model

To answer our RQ1, we first tested the performance on finding relevant comments on the Singapore dataset for both KL methods discussed in Section 4.5 . For this evaluation, we used only the labels z from the human judgment, i.e. the relevance judgment. We tested both relevance models: KL-divergence using all words and KL-divergence using only nouns. The results are shown in Table 7. We used the F-measure to evaluate the results. If score is greater than $\tau >$ -2.2 (set empirically), the comment is relevant to the topic. We can see that using nouns to compute the KL-divergence score works better. So, for the succeeding experiments we use F_5 which is the feature based on KL-divergence score between a comment and a target speech segment using nouns only.

Feature	Model	F-1
F_5	KL-Divergence using nouns only	**0.634**
F_6	KL-Divergence using all words	0.611

Table 7: Comparison between the two KL-divergence scores on the Singapore dataset.

7.2 Discourse Relations

To answer our RQ2, we studied the influence of various discourse relations on the F-measure of the comment thoughtfulness using the logistic regression model. For this evaluation, we used the labels y from the human judgment, i.e. the thoughtful comment. Table 8 shows the comparison of all three classes of discourse relations (Comparison, Expansion and Contingency) on comment quality. We can see that for both data sets, when *comparison* relations are used, the accuracy is the highest for both data sets. For the subsequent experiments, we use only the *comparison* relations to form our discourse feature, i.e. F_4 is set to be the number of *comparison* relations in a comment.

DR-Level	Singapore	US
All	0.6186	0.6464
Comparison	**0.6313**	**0.6538**
Expansion	0.5824	0.6111
Contingency	0.6213	0.6309

Table 8: Comparison of different classes of discourse relations using F-measure.

7.3 Thoughtful Comment Study

To answer RQ3, we conducted a detailed analysis on all the feature combinations we summarized in Table 3. We tested the thoughtfulness of the comments for a given article using the logistic regression model. The results of our experiments are shown in Table 9 for Singapore and in Table 10 for US. For all our experiments we performed 5-fold cross validation and with all the combinations of the features. For better analysis, we show only the most important combinations in the results.

For the Singapore data set, using linguistic features alone (F_1+F_2+F_3) leads to a F-score of 73.33%. Our hypothesis is that discourse relations play important role in detecting thoughtful comments. The results confirm that using discourse relations together with linguistic features

Feature Set	Recall	Precision	F-1
F_1+F_2+F_3	0.7097	0.7586	0.7333
F_1+F_2+F_4	0.8065	0.7143	0.7576
F_1+F_3+F_4	0.8387	0.6667	0.7429
F_2+F_3+F_4	0.8065	0.6944	0.7463
F_1+F_2+F_3+F_4	0.7742	0.7273	0.7500
F_1+F_2+F_3+F_5	0.7419	0.7667	0.7541
F_1+F_2+F_4+F_5	0.8065	0.7813	**0.7937**
F_1+F_3+F_4+F_4	0.7419	0.7931	0.7667
F_2+F_3+F_4+F_5	0.7742	0.7500	0.7619
F_1+F_2+F_3+F_4+F_5	0.7742	0.8000	0.7869

Table 9: Prediction results of thoughtful comments for Singapore using various feature combinations.

yields (F_1+F_2+F_3+F_4) a 75% F-score. But the model performs slightly better without syntactic features (F_1+F_2+F_4)) with 75.76%, which is a 0.76% increase over combined features and 2.43% higher than the linguistic features. Our second hypothesis is that relevance factors play an important role in detecting thoughtful comments. The results confirm that using relevance scores together with linguistic features and discourse relations (F_1+F_2+F_3+F_4+F_5) leads to 78.69% F-score, which is a 3.69% increase compared to linguistic together with discourse relations. Here again, we notice that the model has better performance without syntactic features (F_1+F_2+F_4+F_5) with F-score of 79.37, which is a 4.37% increase compared to linguistic together with discourse relations and 6.04% higher than linguistic features alone.

Feature Set	Recall	Precision	F-1
F_1+F_2+F_3	0.6522	0.6818	0.6667
F_1+F_2+F_4	0.7826	0.6429	0.7059
F_1+F_3+F_4	0.7391	0.6296	0.6800
F_2+F_3+F_4	0.7391	0.5862	0.6538
F_1+F_2+F_3+F_4	0.7826	0.6207	0.6923
F_1+F_2+F_3+F_5	0.7391	0.6296	0.6800
F_1+F_2+F_4+F_5	0.7826	0.6923	**0.7347**
F_1+F_3+F_4+F_4	0.7826	0.6429	0.7059
F_2+F_3+F_4+F_5	0.7391	0.6538	0.6939
F_1+F_2+F_3+F_4+F_5	0.7826	0.6667	0.7200

Table 10: Prediction results of thoughtful comments for US using various feature combinations.

For experiments on US data set, using linguistic features (F_1+F_2+F_3) alone leads to the F-measure of 66.67%. Discourse relations together with linguistic features (F_1+F_2+F_3+F_4) yields 69.23% F-measure which is a 2.56% increase over features without discourse relations. Here, we also notice that the model performs slightly better without syntactic features (F_1+F_2+F_4) with F-score of 70.59% which is 1.36% increase over combined features. Using relevance scores together with linguistic features and discourse relations (F_1+F_2+F_3+F_4+F_5) leads to 72% F-measure which is 2.77% increase compared to linguistic together with discourse relations. We also notice that the model has better performance with out syntactic

features (F_1+F_2+F_4+F_5) with F-score of 73.47, which is 4.24% increase compared to linguistic together with discourse relations.

During our analysis, we observed that the US data is less verbose compare to Singapore data. Another thing we also noticed is that the US data users focus more on the speech delivery rather than the actual speech issues. At the same time, they tend to discuss mostly one issue in each comment where as, Singapore users tend to combine many issues in their comments. So, even if one issue is justified in the comment, the comment is treated as thoughtful comment. This explains the performance differences in the two data sets. It will be interesting to study more fine grained opinion analysis at the comment level and we leave it for our future work.

7.4 Cross Collection Experiments

We further perform cross data collection experiments to test the performance of our model. We tested Singapore using USs' 5-feature model and viceversa. To compare the cross data collection results with original results, we depict Table 11 which shows the F-measure prediction on thoughtful comments. Singapore performed with a quality prediction F-measure lowered by 4.49%, whereas the US performance decreased by 5.11% compared to actual model.

		train	
		Singapore	**US**
test	**Singapore**	0.7937	0.7488
	US	0.6836	0.7347

Table 11: Cross data collection comparison. F-Measure for thoughtful comments.

We show some sample comments in Table 12 for both datasets. Due to space constraints we show 2 sample thoughtful and thoughtless comments for two issues in each articles.

Topic	**Quality**	**User Comment**
Innovation	Thoughtless	I love these plans on energy, but alas, the energy secretary appears to be asleep.
	Thoughtful	You want to really drive innovation, job growth and entrepreneurs? Make education, health care and retirement less of a burden on the average family, adopt more socialist policies like Norway
Spending & Taxes	Thoughtless	..Oh, so Obama "compromised" on the tax cuts for the wealthy
	Thoughtful	Low taxes aren't helping the vast middle and working class and aren't creating more jobs, it's a policy that only benefits the rich.
Housing	Thoughtless	By the way did anybody count the no of flags on a HDB flat. believe me 95% of the time u will take less than 10 sec to do it
	Thoughtful	I am glad that to hear more HDB houses to be built. But do I got a taste of this pie? What about those who are genuine to upgrade their existing 3 room flat but not 1st timer?..
National Service	Thoughtless	i'm still waiting before the budget and erection. otherwise i'll vote oppo. 9k is ?
	Thoughtful	Just 9000 for NSman. Those foreign scholar in NUS NTU got tution non-subsidize fee alone is 20000 one year. That even exclude lodging and return ticket fully paid by PAP.

Table 12: Sample comments: First two topics are from US and last two are from Singapore

7.5 Parameter Sensitivity

The regularization parameter λ in Equation 5 is set empirically. We study the optimal value and tuned it by regular cross validation. Figure 1 shows our experiments for both Singapore and US datasets. We get optimum results when we set λ to 0.1 or 1. We choose 0.1 for both datasets as it generates higher prediction performance in general.

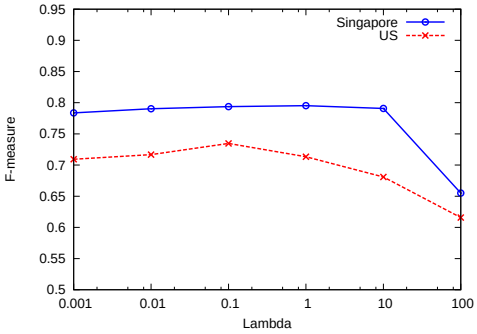

Figure 1: Regularization parameter sensitivity study

Conclusion and perspectives

Detecting thoughtful opinions is an important subtask of opinion mining and summarization. We perform an empirical study using syntactic, vocabulary, discourse and relevance features for prediction and combination of all is substantally better than the baseline surface features. Moreover, through our cross data collection experiments, we show that prediction using our approach achieves competitive performance.

Currently, we use KL divergence to compute the similarity but users tend to use abbreviations for some words and this impacts the performance of KL-divergence scores. We want to try other similarity techniques based on topic modeling and enhance the relevance performance. Extending the problem to identify the sentiment orientation is another useful subtask of opinion mining which we want to try next. In the future, we would like to extend our work to application base, and investigate the usage of thoughtful opinions in opinion summarization.

Acknowledgments

This research/project is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, New York, NY, USA. ACM.

Amgoud, L., Bannay, F., Costedoat, C., Saint-Dizier, P., and Albert, C. (2011). Introducing argumention in opinion analysis: Language and reasoning challenges. In *Sentiment Analysis*

where AI meets Psychology, pages 28–34, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment (JTLA)*, 4(3).

Bian, J., Liu, Y., Zhou, D., Agichtein, E., and Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In Quemada, J., Leon, G., Maarek, Y. S., and Nejdl, W., editors, *WWW*, pages 51–60. ACM.

Chen, B.-C., 0002, J. G., Tseng, B. L., and 0015, J. Y. (2011). User reputation in a comment rating environment. In Apte, C., Ghosh, J., and Smyth, P., editors, *KDD*, pages 159–167. ACM.

Choi, Y., Breck, E., and Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439.

Cui, H., Mittal, V., and Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, pages 1265–1270.

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowl. Data Eng.*, 23(10):1498–1512.

Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *CSE (4)*, pages 90–97. IEEE Computer Society.

Huang, X. and Croft, W. B. (2009). A unified relevance model for opinion retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 947–956, New York, NY, USA. ACM.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 151–160.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 423–430, Sydney, Australia. Association for Computational Linguistics.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342. Poster paper.

Lu, Y., Tsaparas, P., Ntoulas, A., and Polanyi, L. (2010). Exploiting social context for review quality prediction. In Rappa, M., Jones, P., Freire, J., and Chakrabarti, S., editors, *WWW*, pages 691–700. ACM.

Macdonald, C., Ounis, I., and Soboroff, I. (2009). Overview of the trec 2009 blog track. In *TREC*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *EMNLP*, pages 186–195. ACL.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. K. (2008). Easily identifiable discourse relations. In Scott, D. and Uszkoreit, H., editors, *COLING (Posters)*, pages 87–90.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *LREC*. European Language Resources Association.

Saito, M., Yamamoto, K., and Sekine, S. (2006). Using phrasal patterns to identify discourse relations. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 133–136, New York City, USA. Association for Computational Linguistics.

Tang, R., Ng, K. B., Strzalkowski, T., and Kantor, P. B. (2003). Automatically predicting information quality in news documents. In *HLT-NAACL*.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541.

Zhang, M. and Ye, X. (2008). A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 411–418, New York, NY, USA. ACM.