Exploring Generic Features For Online Large-Scale Discussion Forum Comments

by

Adithya Aggarwal

A Dissertation Presented in Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2016 by the
Graduate Supervisory Committee:

Ihan Hsiao, Chair
Claudia Lopez
Erin Walker

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Online discussion forums have become an integral part of education and are large repositories of valuable information. They facilitate exploratory learning by allowing users to review and respond to the work of others and approach learning in diverse ways. This research investigates the different comment semantic features and the effect they have on the quality of a post in a large-scale discussion forum. We survey the relevant literature and employ the key content quality identification features. We then construct comment semantics features and build several regression models to explore the value of comment semantics dynamics. The results reconfirm the usefulness of several essential quality predictors, including time, reputation, length, and editorship. We also found that comment semantics are valuable to shape the answer quality. Specifically, the diversity of comments significantly contributes to the answer quality. In addition, when searching for good quality answers, it is important to look for global semantics dynamics (diversity), rather than observe local differences (disputable content). Finally, the presence of comments shepherd the community to revise the posts by attracting attentions to the posts and eventually facilitate the editing process.

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor Dr. Sharon(Ihan) Hsiao, from the School of Computing, Informatics & Decision Systems Engineering at ASU. The door to Dr. Hsiao's office was always open whenever I ran into a trouble spot or had a question about my research or writing. She consistently allowed this paper to be my own work, but steered me in the right direction whenever she thought I needed it.

I would also like to thank Dr. Claudia Lopez, from the Universidad Tcnica Federico Santa Mara, who was involved in the validation part of the research without whose passionate participation and input, the results would not have been substantiated.

I would also like to acknowledge Dr. Hasan Davulcu, ASU and Dr. Erin Walker, ASU as the second readers of this thesis, and I am gratefully indebted to their very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my friends, Varsha Iyengar, Arundhati Kanakagiri, Devika Krishnamurthy and Kruthika Tanikella for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Online discussion forums have become an integral part of education and are large repositories of valuable information. They facilitate exploratory learning by allowing users to review and respond to the work of others and approach learning in diverse ways [4]. For decades, discussion forums have been widely used as a communication medium in and outside classrooms to facilitate learning. The benefits derived from such platforms have extended to open, free & fast-growing online communities (homework-help sites, discussion forums for MOOCs courses etc.) These sites have grown steadily over the past few years and have formed sizable repositories of problem solving-solutions. They are filled with thousands of problem-solving tips such as how-to questions [28],people-valued examples as well as the examples explanations [25]. To better support information seeking and learning, a majority of these sites utilize social mechanisms to filter and point out the best solutions/answers instead of providing the set of steps or scaffolds required for learning [27, 26]. In large-scale discussion forums like StackOverflow where solutions are crowd-sourced, the role of social interactions in the form of comments plays a vital role in shaping a post. The research presented in this paper originates from trying to connect the dots between the social interaction features and the post content quality. We investigate the different comment semantic features and employ key quality identification features from relevant literature survey to find the effect they have on the quality of a post in a large-scale discussion forum.

In a broader sense, the work aims to understand how students use (or could use) these sites to learn to program. Particularly, we focus on how assorted levels of content on StackOverflow can be used to better support scaffolding. StackOverflow is one of the biggest online programming Q&A communities and currently hosts a massive amount of heterogeneous definitions, solutions, and examples in various programming languages.

Today, data mining techniques have noticeably enhanced the ability to organize and analyze the mass amount of content. For instance, sentiment analysis has been widely explored and applied in mining large-scale social media data. Opinion mining can successfully detect trends, assist decision-making, discover spam reviews [16, 21, 12] etc. Q&A sites have social features that focus on finding the best answers, however, they emphasize less on highlighting the diverse points of views among the content sea. For instance, unclear questions might tend to raise doubts amongst users leading to the usage of controversial words such as "I'm not sure", "I doubt that", etc; duplicate questions may lead to contentious arguments between experts and novices. On the other hand, a healthy argument between experts commenting on a post could also incorporate controversial phrases; a constructive criticism or a debate leading to editing a post a number of times in order to improve the quality of the post. However, widely-used social mechanisms, such as crowd-sourced evaluation typically ranks all solutions, are useful to point out the best and the worst contents, but they do not necessarily help to discern the gray area. For instance, often times there is more than one "correct" answer, especially for code review questions [25]. Even though the social metrics (votes, acceptance, favorite) point to the correctness of a possible solution, readers have differing backgrounds and varying degrees of previous knowledge which influence which answer is the best for that particular individual. Therefore, these quality indicators may not be universally applied to all users. For example, an accepted answer may be too complicated for a novice to understand.

A low-ranked answer can approximately indicate a low quality of its content as a solution to the problem, but this can also overlook the learning-value of showing obvious mistakes. In the context of learning to program, students can learn by constantly being exposed to similar or different solutions. This enables them to compare and contrast alternatives in order to identify what is important in a problem, distinguish different features of code, and paint a road-map to the problem [20, 17]. Multiple pieces of code can allow a novice to construct patterns and schemas [14], and even erroneous examples may assist learning at various levels [16]. Therefore, in this work, we aim to harness the controversy dynamics along with other comment semantic features in discussion forums to help us explore and understand the quality of a post.

1.1 Motivation

There are primarily three kinds of users in StackOverflow, i) Users who ask questions ii) Users who answer the questions i.e the experts and iii) Users who are looking for answers; normally not involve in the discussions. It can be observed that these discussions are mainly held by the first two categories of users and in many cases, help in re-structuring the questions. For instance, questions that are ambiguous in nature are often clarified by a back and forth exchange of comments. A "poor" quality question is scrutinized by a group of experts until the question is either edited or removed. The exchange of comments thus provides insight into the quality of the post itself. Figure 1.1 illustrates a scenario where a user is criticized for asking a bad quality question.

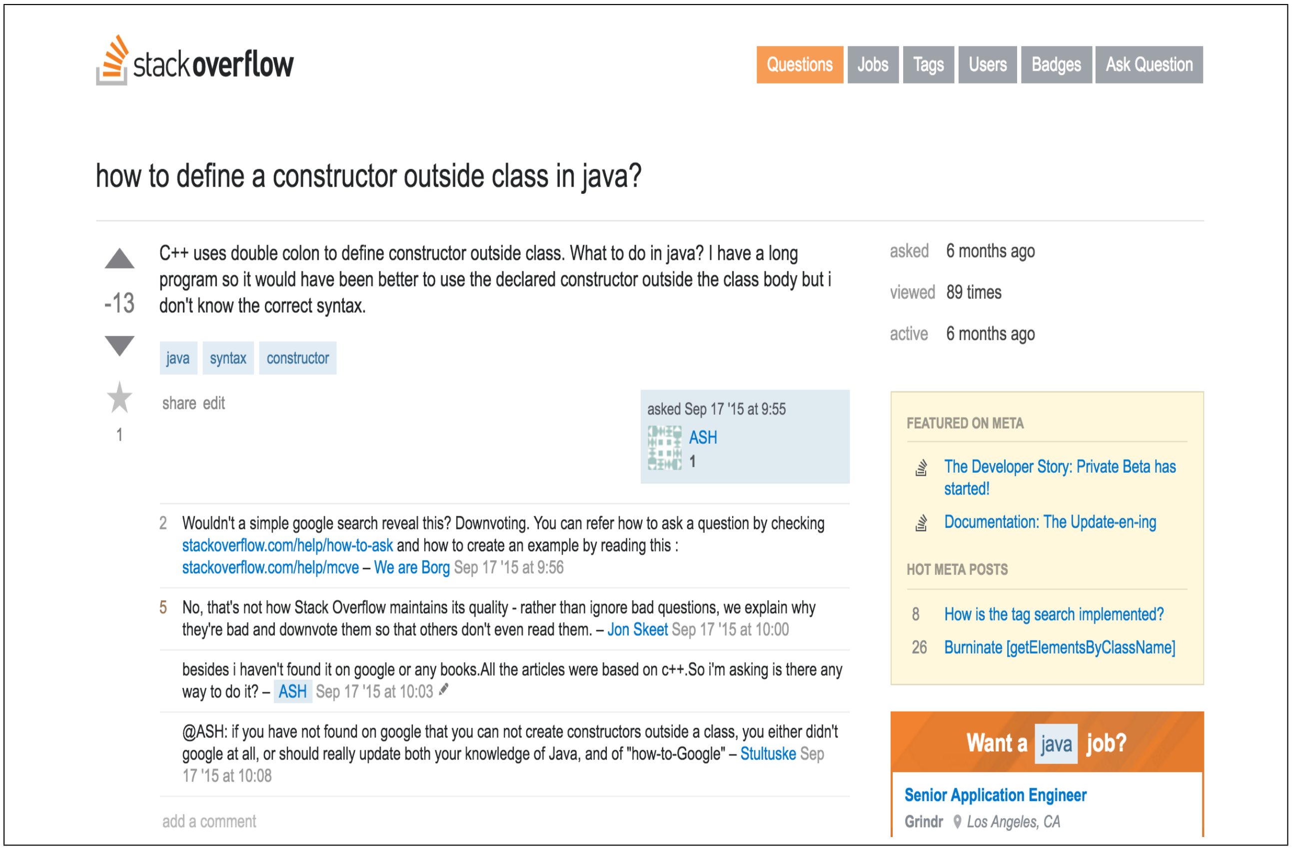

Figure 1.1: Stackoverflow Snippet Illustrating a Poor Quality Post. The Figure Highlights the Fact That Poor Quality Post's Are Down-voted and Receive Negative Comments by Users.

Moreover, users rely on StackOverflow's social features like 'votes', 'accepted answer' and 'reputation' when selecting an answer. The number of votes is directly proportional to its chance of being picked. An 'accepted answer' essentially implies that it worked for the user asking the question and thus prompts other users to pick the answer. These quality features are crowd-sourced and they stabilize over a period of time. Comments appear instantly due to the vast multitude of people using StackOverflow; roughly 5,395,936. Our evaluation reveals that a comment appears approximately fifteen minutes after a post has been rendered by a user. In order to gain reputation, a user needs to answer unanswered questions. This motivates the experts to answer these questions as soon as it is posted thus the average time a comment appears was observed to be less than fifteen minutes. The time taken for prediction thus provided further incentive to analyze the quality of the post.

This paper investigates if comments semantic features like comment sentiment and controversies along with StackOverflow's social and syntactic features can help establish certain patterns for interpreting post quality. Broadcasting this information on the forum could increase the confidence of the user in electing an answer until StackOverflow's features have stabilized.

StackOverflow follows a crowd-sourced model where every time a user's answer gets up-voted, his points aggregation, labeled as "reputation", increases. Reputation is of significant value in the community, indirectly indicating the level of expertise of a user. This system might negatively impact the users encouraging them to give quick but low-quality answers for the sake of gaining reputation, falling prey to gaming [9, 19]. Moreover, the rating systems in online discussion forums are subjected to malicious behavior [30]; users might individually or collectively promote or demote a post resulting in unfair rating scores. We are aware of these discrepancies that crowd-sourced models such as StackOverflow contain but this remains out of the scope of the current research.

1.2 Research Questions

As of today, content quality assessment of online forums is gauged either by a group of experts or is crowd-sourced. There lacks a system that automatically determines quality solely based on semantic and syntactic features. This research work is a step towards filling this void by exploring the effect of various components, like the presence of controversial comments, on post quality. To achieve this feat, we aim to answer the following questions:

- Can comment semantic features predict quality as good as the crowd-sourced metrics?
- Can we rely on the crowd-sourced evaluation metric for instance, 'votes', to assess our model?
- Can we find universal norms and patterns in current data in the form of comment features that users could follow, eventually enhancing post quality?
- How can we extract more generalizable features that can be extended to other online forums?

1.3 Contribution

- Identifying and engineering comment semantic features that contribute in detecting post quality.
- Utilizing data mining techniques like sentiment analysis to predict post quality.
- Ciphering a number of benchmarks, inferences and features that directly impact post quality, for instance, a) Higher the number of commenters per post, higher is the post quality b) Entropy is a better indicator of finding good content than controversy c) Reconfirming the literature survey that Edit count strongly facilitates content quality.

We extract representative quality features based on surveyed relevant literature in StackOverflow content modeling and construct new models in capturing disputative forum posts semantics. The rest of the research is structured with reviewing two streams of related work: StackOverflow content quality modeling and sentiment analysis in Q&A forums. We then present the methodology by describing data collection and feature engineering. Finally, we present the evaluation results and draw conclusions on the findings.

Chapter 2

LITERATURE REVIEW

2.1 Quality of the Content on Stackoverflow

StackOverflow has been defined as the epitome of Q&A sites and has been extensively studied. Previously a lot of research has gone into investigating its design [13], its value for software development [29] and content quality [6]. The thesis closely lines with the latter group. Hence, the literature review is focused on this strand of research

Prior research [6] has already classified the many studies on the quality of content into three categories: (1) finding the best answer, (2) ranking all answers, and (3) assessing the quality of the questions. Given that the focus of this research is on identifying different points of views in the answers can be a valuable learning opportunity, related work that examines the quality of answers (categories 1 and 2) have been the most informative for purposes this research. Extensive data mining has been used to identify the main factors that are significantly associated with an answers ranking, which is given by the votes that the answer has received on the site. These factors can be grouped according to the entity they describe, such as the user who provided the answer (e.g., reputation, expertise), the answers' text (e.g., length, structure, style) and the amount of review the answer has had (e.g. edits, comments). Hasan et al., in their paper [6], exploit user feedback to learn how answers can be ranked in Q & A websites. Their research attempts to automate the process of identifying loose edits that could arise due to the feature of free edit in discussion forums. They propose a learning to rank (L2R) approach for ranking answers in Q&A forums that uses Random Forests. The model was trained with user feedback given to answers in Q&A forums. Factors related to the user and the review process were found the most relevant to predict ranking. Finally, their research performs a comprehensive study showing that text features are useful in assessing the quality of answers, hence bolstering the foundation of the hypothesis of the thesis. Rather than taking a data mining approach to replicate these findings, the proposed research aims to extend this line of research by further understanding the impact of the review process, while controlling for other aspects that are known to affect the perceived quality of an answer. Beyond these factors, contextual factors such as the number of alternative answers and the time that the answers were provided have also been found related to the number of votes an answer gets [2, 23].

A major stumbling block for large-scale Q&A forums is of having questions without an accepted answer or having answers with low votes. As a result, users are hesitant in selecting answers for such questions. In some cases, the question is deleted from the forum's database resulting in loss of could-be valuable information. This scenario bolsters the purpose of this research work and asks for the existence of a system that could predict the presence of quality answers at times when the community-based features of StackOverflow might fail. Tian et al., in their paper [23] address the above-mentioned issue by predicting the best answer in the lot using semantically and syntactically tailored features. Their work's first contribution is towards identifying the features that could lead to an answer achieving the feat of the best answer. They assess three major qualities 1) quality of the answer content 2) whether the answer solves the question or not 3) how it competes with other answers. Their second contribution involves designing and evaluating a learning approach using these features.

Yao et al., in their paper [32] express the concern of asking good questions and how they could attract good quality answers. They hypothesize that quality of a question could correlate to that of its answers. An interesting question might obtain more attention from potential answerers and possibly have a better chance to receive high-quality answers. Moreover, predicting the quality of existing questions and answers, especially soon after they are posted, becomes an essential task for both information producers and consumers. From the perspective of information producer (e.g., who asks or answers questions), predicting the quality as early as possible could help the questions that are potentially of high quality to attract more high-quality answers by recommending these questions to experts. Users are always looking forward to answering good questions and identifying question quality would accelerate the process also as an aftereffect, indicate useless and spam posts. The work presented by Yao et al., further fosters the basis of this research.

One of the prime motivations of carrying out this research was to find a way to rank new answer posts before they received any community reputation. Tian et al., in their paper [24] expresses similar concerns. Their approach predicts the best answerer for a new question. Users of CQA sites post their questions and wait for other users to post answers to the question, which may take several days. Even if there are answers, without votes it is difficult for the user to show confidence in using it. The asker is sometimes not satisfied with the quality of the answer. In both cases, a system that could recommend questions to experts who have a higher likelihood of answering them is needed. Their method considers both user interest and user expertise relevant to the topics of the given question. User interests on various topics are learned by applying topic modeling to the previously answered questions while his expertise is learned by leveraging collaborative voting mechanism of CQA sites. They claim to outperform the state of the art approach of TF-IDF .

2.2 Sentiment Analysis in Q&A Forums

Sentiment analysis has been widely applied in social media datasets in order to detect trends, assist decision-making, discover spam reviews [16, 21, 12] etc. Most of the sentiment classification approaches are based on linguistic language structures (i.e Linguistic Inquiry and Word Count), semantic network relations (i.e. WordNet, sentiWordNet [3]), or build a domain-specific opinion dictionary by defining positive and negative adjectives relevant to the corpora [11, 7]. There is only a few explored opinion mining in open discussion forums with students' learning. Wen et al. [31] investigated students' opinions towards a course. The researchers introduced subjectivity and sentiment mining based on educational data mining findings, such as boredom was associated with poorer learning; frustration was less associated with poorer learning, which attitude, in general, played an important factor in effective learning. The authors found a correlation between sentiment ratio on daily forum posts and student drop out each day. Munezero et al. [15] coded eight derivative human emotions from Plutchiks emotion psycho-evolutionary theory [18] to track emotions in a students learning diaries, which are made available for instructors as a source of informative feedback. They further visualized the variations of students emotions over a period of time and helped faculty narrow down those students whose anxiety/frustration levels seem increasing.

Our research work can be considered as an extension to Hsiao's et al., work in [11]. Their paper aims to study automatic methods for identifying useful content to learn to program from a discussion forum. The hypothesis of their paper is that identifying learning-inductive content in large scale programming discussions will prevent learners from searching for a needle in the haystack and reading a mess of new detail. Their research shows that preexisting forum quality indicators might (votes, acceptance, favorite) point to the correctness of a possible solution but, readers have differing backgrounds and varying degrees of previous knowledge which could influence which answer is the best for that particular individual. Therefore, these quality indicators may not be universally applied to all users. They further explain this with an example; an accepted answer may be too dense for a novice to grasp. They delve into discovering artificially intelligent methods to evaluate content quality that could provoke learning in CQA. Their approach involves them building a constructive lexicon library to capture comparing & contrasting words, explanation, and justification & elaboration words. They then identify useful content that might evoke learning by constructing a model to capture the quality of the contents that signal constructive behavior. In our research work, we make use of the constructive measure used in this paper to investigate into the constructiveness of the comments to reveal post quality.

In this work, in order to investigate the effects of disputable perspectives among social interactions, we incorporate SentiWordNet3 [8] to compute the generic language sentiments from discussions. It is an enhanced version of SentiWordNet [3]. SentiWordNet3 is a lexical resource explicitly devised for supporting sentiment classification and opinions mining applications. More than 300 research groups and a variety of research projects use it. It is the result of automatic annotation of all the synsets of WORDNET according to the notions of positivity, negativity, and neutrality. Each of the synsets s is correlated to three numerical scores: Pos(s), Neg(s) and Obj(s), which indicate respectively the positivity, negativity and neutrality of the terms contained in the synset. Each of the three scores range in the interval [0.0,1.0], and the sum is 1 for each synset.

Our research uses Somasundaran et al's., work [22] label controversial semantics. Their work includes analyzing opinion categories like sentiments and arguments in meetings. They use the AMI corpus consisting of 6504 sentences to carry out all their analysis. Their research hovers around identifying arguments in conversations as little research has been done in this field compared to identifying sentiments. First, they manually annotated categories and then developed genre-specific lexicons using interesting function word combinations for detecting the opinions. They developed an arguing lexicon as a new knowledge source for automatically recognizing the argument category. Their work resulted in them proving that dialog structure interacts with the expression of opinions and they confirmed this through machine learning experiments.

Another one of the relevant works was by Agrawal et al., [1] where they address confusions in MOOC Discussion Forums. They developed a system (YouEDU) which presents a unified pipeline that automatically classifies forum posts across multiple dimensions. They particularly detect the presence of confusion and then present the users in that post with one-minute-resolution video snippets that are likely to help address the confusion. Besides describing the extent of confusion, each entry in the MOOCPosts set indicates whether a particular post was a question, an answer or an opinion, and gauges the post's sentiment and urgency for an instructor to respond. Evaluation of their system showed that YouEDU performed well and addressed the issue of manifestation of confusions in MOOC discussion forums.

Chapter 3

METHODOLOGY

3.1 Research Platform

Unlike discussion forums like Yahoo! Answers and Quora that cater to a broad range of interests, StackOverflow is a question and answer forum designed mainly for professional and enthusiast programmers. It is an aggregated library known to contain the widest range of programming questions and answers. The most striking (aspect) is the largely engaged user community that collaboratively manages and edits the posts on the site. As the questions in StackOverflow require an in-depth knowledge of various domains in the programming field, a large part of this community comprises of experts in the domain. The quality of the content is actively conserved through this community; any question which is of questionable quality is generally detected by an expert and put on a hold. These questions then stack up in a review queue from where it goes to experts to decide its existence. The popularity of StackOverflow is largely based on the numerous features it provides to its users. Through membership and active participation, the website allows its users to up-vote or down-vote posts. Users of StackOverflow can earn reputation points and "badges"; for example, a person is awarded 10 reputation points for receiving an up-vote on an answer given to a question, and can receive badges for their contributions [17], which represents a kind of gamification of the traditional Q&A forum. The feature of closing questions 3.2 differentiates other discussion forums from StackOverflow. Figure 3.1 shows a snippet from the discussion forum.

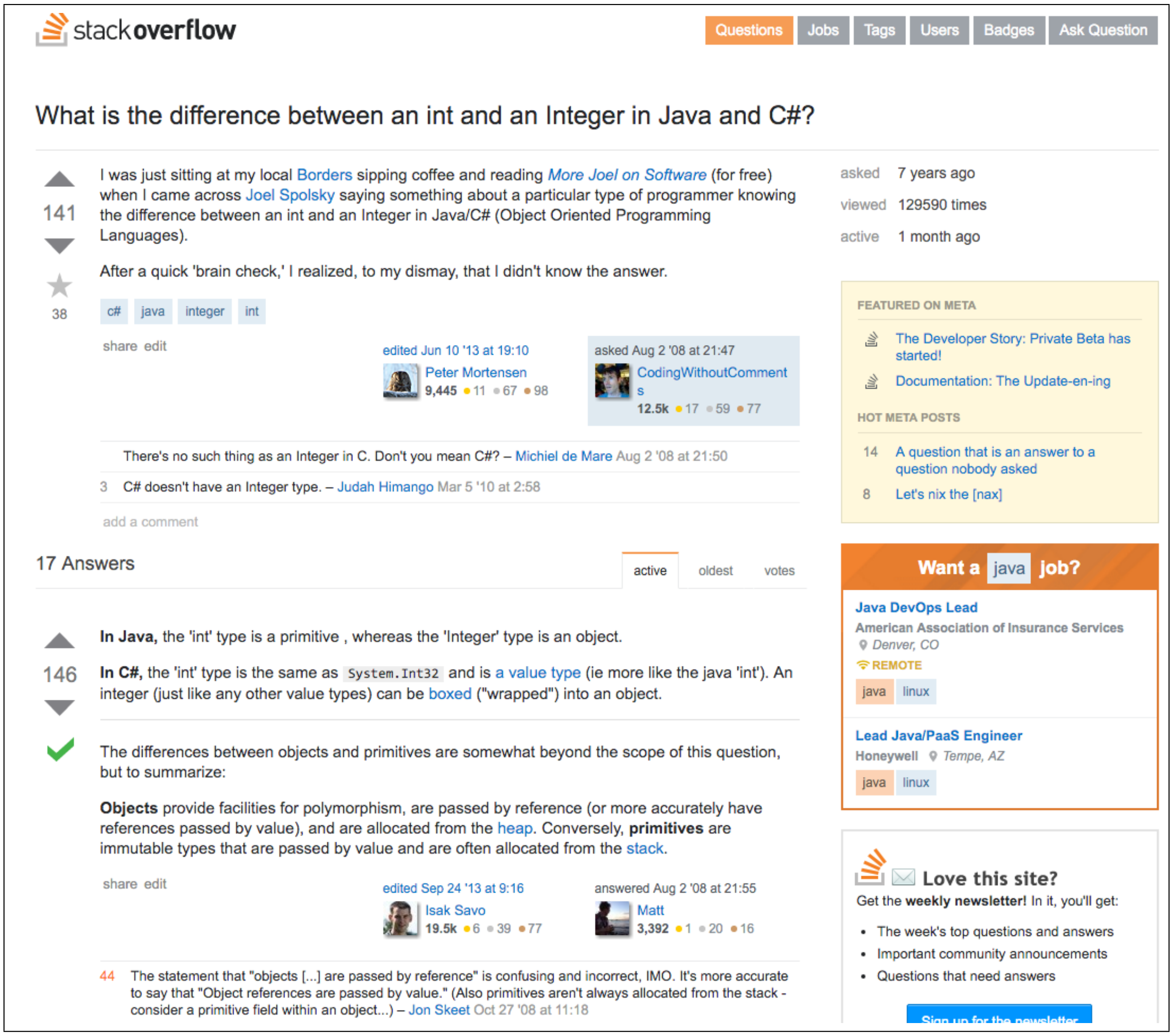

Figure 3.1: Snippet of a Page from Stackoverflow.com - The Image Displays a Question Post, Comments on the Question Post, an Answer Post and Comments under the Answer Post.

3.2 Data

The focus of this study is in the comments of the posts. The comments section in SO is utilized when users have something worth saying about the post that could raise a discussion, asking the questioner for clarity or having a procedural point to

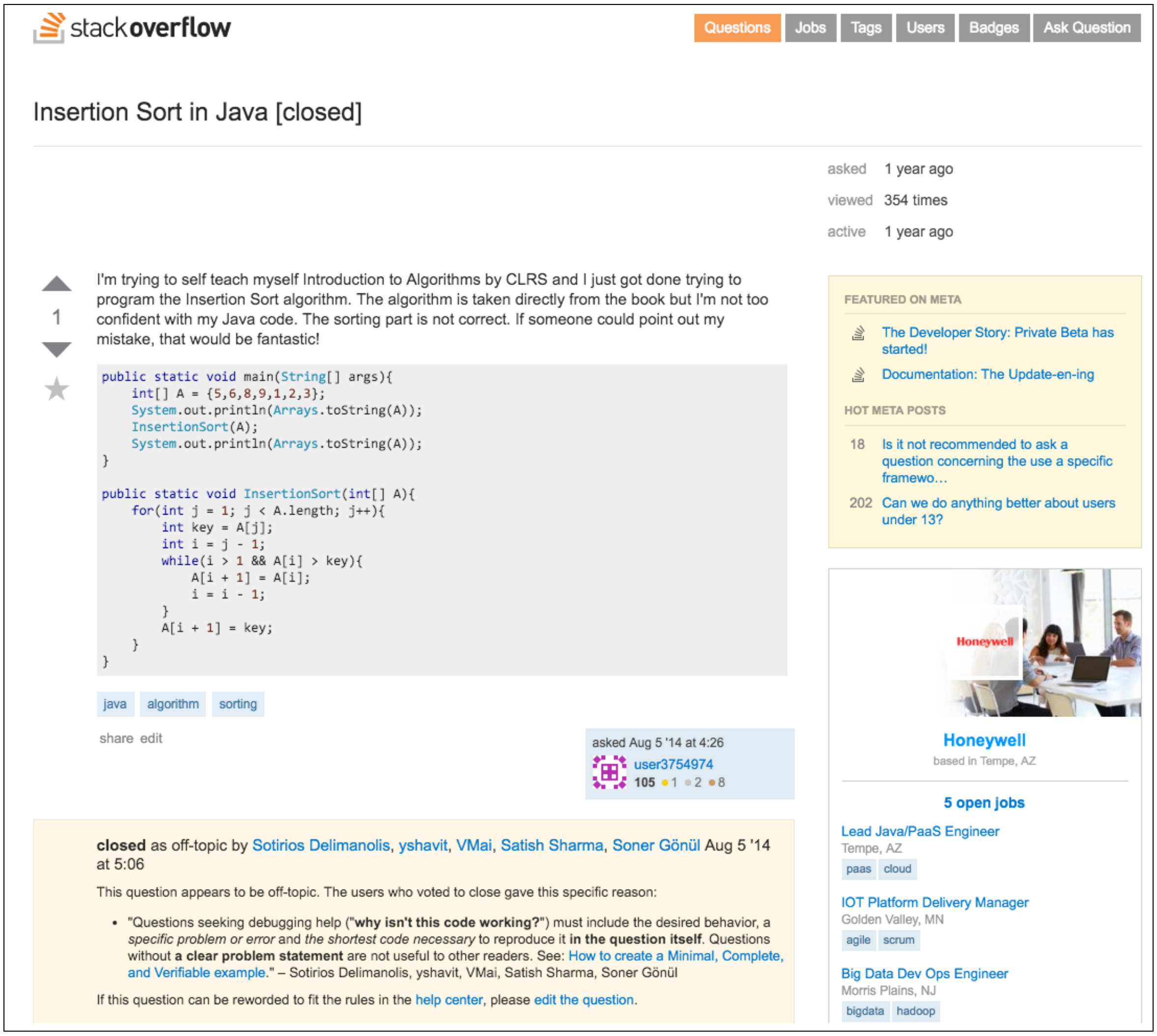

Figure 3.2: Snippet of a Page from Stackoverflow.com - The Image Displays a Question Post That Has Been Closed by Users Having Enough Reputation.

make about the existing post. Below is a sample list of data fields used to scrape StackOverflow data from Stack Exchange API.

```
<post_id>: numeric value
<parent_id>: numeric value
<posttypeid>: integer
```

```
<post_body>: varchar
<post_title>: varchar
<acceptedanswerid>: numeric value
<comment_text>: varchar
<post_creation_date>: date
```

We show that the presence of controversial comments under a post strengthens the prediction of post quality (#Votes). We use the MPQA (Multiple Perspective Question Answering) opinion corpus developed by University of Pittsburgh and adopt the arguing lexicon to investigate the existence of controversial phrases in a comment [22]. The arguing lexicon consists of 17 arguing lexicons of which our prime focus is on the presence of 3 strongly controversy lexicons (doubt, inconsistency, and contrast). Table 3.2 shows a sample of comments being tagged with lexicons. Lexicon entries are in the form of regular expression patterns. We perform our analysis on 8038 unique posts from StackOverflow, 5025 of these are question posts. Table 3.1 depicts the frequencies of the occurrence of Lexicons in the comments of posts in our data set. Java being the most popular of the programming languages, for our analysis we have used posts that have been tagged with 'java' in SO, dated January 01, 2014 to January 10, 2014. We queried a list of features using StackExchange Data Explorer API from the StackOverflow database. Since our research is centered towards regulating the importance of controversial comments under a post, we divide our data into post level and comment level.

In order to retrieve the value of comment sentiment, we used sentiWordNet which is an opinion mining resource. It uses the principle of automatic annotation of all synsets of WORDNET according to the notions of "positivity", "negativity", and "neutrality".

	Lexicon	**Frequency**		**Lexicon**	**Frequency**
1	assessment	395	**9**	generalization	66
2	authority	49	**10**	inconsistency	945
3	causation	2947	**11**	inyourshoes	41
4	conditionals	354	**12**	necessity	3314
5	contrast	1472	**13**	possibility	1450
6	difficulty	0	**14**	priority	408
7	doubt	166	**15**	rhetorical question	171
8	emphasis	802	**16**	wants	0

Table 3.1: Arguing Lexicon Frequency

Comment	**Lexicon**
I feel sorry for you. If it's due in a few hours then you are not going to get it working, and SO most definitely won't help you cheat.	possibility; emphasis; causation
@JimGarrison I'm not trying to cheat!! I just want to figure out what I'm doing wrong.	contrast
Thanks @Jeroen, I changed it back but it still gives the error.	inconsistency
@user1973167: I am not convinced you are doing it,correctly from your description of the problem.	doubt

Table 3.2: Lexicon Tagging of Comments

3.3 Feature Engineering

The content semantic features in the analysis are extracted from either the literature in the field or through observational inferences. The syntactic features, as well as the social features inherent to StackOverflow, were also considered. The above categorization of features was applied to the post as well as comment level.

The post-level syntactic features extracted were 'Post Length', 'Timestamp' and 'Type of Post'. It is observed that question posts have a distinguishing characteristic from answer posts in terms of edit counts, the median time of comments etc., hence, it becomes necessary to distinguish an answer post from a question post, justifying our use of the feature. The above is evaluated in the later part of the thesis. 'Timestamp' is of the essence as it helps us understand the swiftness with which questions get answered on StackOverflow. From the literature survey, it was found that 'Post Length' and 'Code Percentage' could be crucial predictors of content quality, our analysis ascertains the following. 'Votes' and 'Constructiveness Score' formed the basis of our evaluation metric.

Votes is a crowd-sourced metric to measure the quality of a post, from relevant literature survey it is proven to be an effective way of measuring post quality. Therefore, we employ this as one of the dependent variables while conducting regression.

"Constructing meaningful posts that require cognitive thinking is a constructive activity" [11]. Based on the ICAP (Interactive, Constructive, Active, Passive) framework [5] the measure of constructive activity include the following possible underlying cognitive processes: inferring, creating, integrating new information with prior knowledge, elaborating, comparing, contrasting, analogizing, generalizing, including, reflecting on conditions and explaining why something works. According to these cognitive processes, Hsiao et al., in their paper [11] build a constructive lexicon library to capture comparing & contrasting words, explanation, justification & elaboration words. Adding few of their own models they extended an opinion mining technique [10] to automatically identify post constructiveness based on the constructive lexicon library. We harness this model in this research to produce the constructiveness scores of a post which forms the second dependent variable we perform regression on.

Our consideration of 'Edit Counts' relied on Observational inference and relevant literature survey, implying towards it being an important variable in detecting quality content. Correlating the existence of controversial comments to post quality forms the hypothesis of this research and we later go on to prove that controversial comments can be imperative to post quality. Table 3.3 lists down all the post-level features alongside their definitions.

Analogous to post-level syntactic features we consider 'Comment Length' and 'Timestamp' as comment-level syntactic features. 'Comment Score' and 'User Reputation' formed the comment-level social features, their contribution towards predicting the quality of post remained unsubstantial. These two features were considered as a result of observational inferences. 'No. of Interactions' in a comment initially gave out an impression that it might be a useful predictor, but it proved else wise. The literature survey was consistent in suggesting 'Comment Sentiment' and 'Entropy' being critical assets in assessing post quality. We relate the term 'Entropy' as a global measure as it's value is calculated on the basis of other comments under the considered post. On the other hand 'Controversy Score' which is a subset of the post-level semantic feature 'No. of Controversial comments' is termed as a local measure due to calculation of controversy being independent of other comments. Table 3.4 lists down all the comment-level features alongside their definitions.

Features	**Description**
Syntactic Features	
Post Length	Number of words including Code words
Timestamp	Creation date and time of the post
Type of Post	Post could be a question, answer or an accepted answer post.
User Characteristics and Social features	
Votes	Crowd-sourced content quality evaluation metric.
Edit Count	Number of Edits a post has,encountered
Content Semantic Features	
Constructiveness Score	The number of constructive word counts that associate with constructive learning activities per post; the underlying cognitive processes are described in detail in [11].
Code Percentage	The amount of code text in a post.
No. of Unique users	Number of unique users commenting on a post.
No. of Controversial comments	The amount of controversial comments per post.

Table 3.3: Post Level Features

Features	**Description**
	Syntactic Features
Comment Length	Number of words including code words.
Timestamp	Creation date and time.
	User Characteristics and Social Features
Comment Score	Crowd-sourced comment quality evaluation metric.
User Reputation	Crowd-Sourced trust measurement.
	Content Semantic Features
Median User Reputation	Median value of reputation of unique users under a post.
No. of Interactions	Number of @ mentions that a user has been referenced in a post.
Comment Sentiment	SentiWordNet value of comment text [3]
Controversy Score	Number of controversy lexicon matches in a comment.
Entropy	Shannon Entropy to represent comment diversity.

Table 3.4: Comment Level Features

3.4 Descriptive Statistics

Our dataset includes 5,025 questions and 3,013 answers, which sum up a total of 8,038 posts. Table 3.5 and figure 3.3 show the descriptive statistics by each kind of posts. Compared to questions, a larger proportion of answers had positive counts of votes. About two-thirds of all answers had a positive count of votes (more than one vote) and a majority of them (60%) had a total count of one and four positive votes.

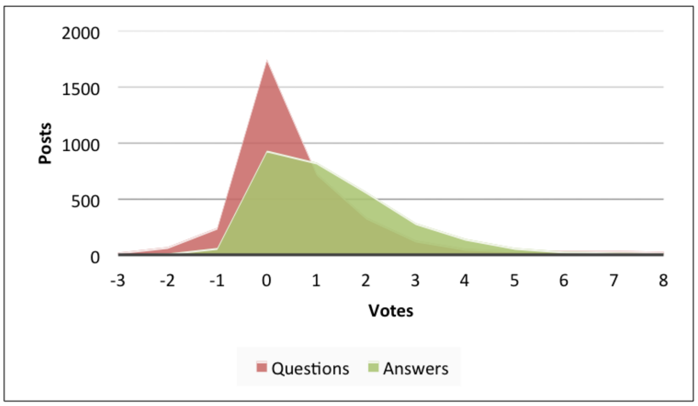

Figure 3.3: Votes Distribution

Variable	**Answers**	**Questions**
Post with negative votes	78 (2.59%)	424 (8.44%)
Post with zero votes	932 (30.93%)	1,750 (34.83%)
Post with between zero and four votes	1,801 (59.77%)	1,250 (24.88%)
Post with more than four votes	202 (6.70%)	1,601 (31.86%)

Table 3.5: Votes by Kind of Post

Questions and answers had similar constructiveness scores on average, but Questions tend to have more words than answers (83.1 vs. 78.9, p = 0.03) and are edited more often (1.8 vs. 0.7, $p < 0.001$) on average (see Table 3.6 and 3.7). The median time of comments are shorter for questions (15 mins), which means that people comment questions quicker than answers (29 mins). We consider the median time of comments instead of mean because the distribution of the timing of comments is very skewed.

A possible explanation can be the most of the comments to questions are clarification type of comments, as a result, they tend to be easier to attend to and result in question askers frequent edits to enhance the content presentation. However, even both answers and questions received comments from 1.9 users on average, users who commented answers provided slightly more comments than those who commented questions (1.6 vs. 1.3 comments per user, $p < 0.001$). In addition, Answers comments also have larger average entropy than questions answers. These reasons lead us to focus on only Answer posts' comments, and how they affect its quality, if any. One interesting note is that Answers in our dataset often included code, and was provided quickly after a question was posted and had few alternative answers. Overall, four out of five answers included code (80%). The average answer appeared 4.5 minutes after a question is posted on the site. However, the number of alternative answers appeared in a skewed distribution and ranged from 0 to 9, with a mean of 0.8 answers. Contrasting to their comments statistics, it indicates the challenge to provide solutions to programming-related problems.

To further investigate the value of comments to answers, we randomly sample 22,756 comments with posts topic Java in the year 2014. These comments are associated with 8,038 posts. 13,189 comments were related to 5,025 questions and 9,567 comments were associated with 3,013 answers. On average, a question received fewer comments than an answer (2.6 vs 3.2, $p < 0.001$). The descriptive

	Answers			
Variable	**mean**	**sd**	**min**	**max**
Post constructiveness	3.6	2.4	1	23
Post length	78.9	98	2	1
Edit count	0.7	1.2	0	19
With code	0.8	0.4	0	1
Time to answer the question	273.2	167.6	1	578
Alternative answers	0.8	1.1	0	9
Number of unique commenters	1.9	0.9	0	11
time of comments (seconds)	876881.7	4726272	0	5.25e+07
Median reputation of commenters	2088.2	1222.56	2	4207
Ratio comments to commenters	1.6	1.1	1	14.5
Ratio negative comments to comments	0.1	0.2	0	1
Mean entropy of comments	4.2	0.2	2.6	5.1

Table 3.6: Descriptive Statistics of Post Variables by Kind of Post:Answers

statistics of our comment variables are shown in Table 5. While 30% of the questions comments had, at least, one positive vote, only 10% of the answers votes have positive votes. Nevertheless, the constructiveness score of answers' comments is larger than questions comments, on average (1.9 vs 1.3, p <0.001). Compared to answers' comments, questions' comments are more positively voted; however, they have a lower score of constructiveness. The mismatch between voting patterns and constructiveness scores again shows that questions comments' can be easier to attend to, but answers' comments may require more cognitive processes (i.e. arguing, contrasting, comparing, etc.) We dig deeper into Answers' comments

	Questions			
Variable	**mean**	**sd**	**min**	**max**
Post constructiveness	3.5	1.9	1	9
Post length	83.1	86.4	1	1
Edit count	1.8	2	0	15
With code	0.8	0.4	0	1
Number of unique commenters	1.9	1.7	0	18
time of comments (seconds)	400689.9	3218941	0	5.24e+07
Median reputation of commenters	2114.1	1230.2	1	4209
Ratio comments to commenters	1.3	0.7	0.5	10.5
Ratio negative comments to comments	0.1	0.2	0	1
Mean entropy of comments	3.2	0.7	2.3	5

Table 3.7: Descriptive Statistics of Post Variables by Kind of Post:Questions

	Answers				Questions			
Variable	**mean**	**sd**	**min**	**max**	**mean**	**sd**	**min**	**max**
Comment with positive votes	0.1	0.3	0.0	1.0	0.3	0.4	0.0	1.0
Comment constructiveness	1.9	1.2	1.0	10.0	1.3	1.2	0.0	9.0
Comment length (words)	24.8	20.7	0	276.0	16.3	17.9	1.0	216.0
Time to comment (seconds)	780583.5	4795113	0.0	5.38e+07	356365.5	3209177	0.0	5.39e+07
Controversy	0.2	0.9	-5.9	6.8	0.1	0.7	-5.6	5.7
Entropy	4.2	0.3	0.8	5.5	3.7	0.9	0.9	5.9

Table 3.8: Descriptive Statistics of Comment Variables by Kind of Post

semantics. We found that Answers comments are more controversial and have more entropy than questions comments. On average, an answers comment is twice as controversial as a questions comment (0.2 vs. 0.1, $p < 0.001$). Answers comments are having more distinctive content (higher entropy) than questions comments (4.2 vs. 3.2, $p < 0.001$). Therefore, even eventually the answers are not accepted, the associated comments left traces of commenters efforts. Such traces can be valuable resources for readers, for instance, commenters trial and error experiences.

Chapter 4

EVALUATION RESULTS

To address our research questions, we conducted regression analysis to model the quality of answers in terms of votes (see Table 4.1) and constructiveness (Table 4.2).

Regression for both votes and constructiveness was carried out for two cases, column one represents regression with post semantic features and column two includes post semantic features with comment semantic features. The reason for considering the above mentioned cases while building the regression model was to highlight the effect comments have on post quality. We used a linear regression to model the number of votes, which could take negative and positive values. Regarding constructiveness, we conducted logistic regression analysis to estimate a binary dependent variable that represented a median split of the constructiveness score.

4.1 The Value of Comments

4.1.1 More Heads Are Better than One

Regarding the comments, we found that the number of unique users who provide comments is associated with the votes an answer received. The more users comment, the higher the number of votes the answer gets. An increase of one unit in the number of commenters increases the number of votes by 0.295. Additionally, the timing of comments is also plays a role. The longer the median time of comments, the more votes the answer is expected to get. The effect size of this measure of time is an increase of 0.098 by every second after the answer is provided. This indicates that the amount of people who participate in the discussion of an answer and the speed at which this discussion takes place are strongly related to the attraction and positive perception that the answer will obtain. Such results are in fact aligned with the literature [2, 23]. There are other findings that are also aligned with prior research. For instance, we found evidence that the presence of code in the answer and the number of edits of the answer have positive relationships with the number of votes. This further suggests the importance of the community review process (number of edits) is a significant factor. However, how do these comments shepherd the edits and achieve higher votes? Is it due to comments diversity or comments controversy?

4.1.2 Comments Semantics Matter to Post Quality: Diversity Is Better than Controversy

To address the comment semantics effects based on each individuals opinions, we evaluate comments entropy and comments controversy. We found that the mean entropy of an answers comments is significantly and positively associated with the answers votes. A unit of increase in the mean entropy is predicted to increase the number of votes by 0.070 units. Yet, neither the ratio of comments to authors nor the ratio of negative comments to total comments has significant effects. The community would consider more diverse information as good quality and then provide higher ratings, but the more disputable comments would actually indicate the more questionable the contents are. The outcome confirms our hypothesis that controversy can be an index in judging post quality, however, in order to find good content, diversity will be a better indicator than controversy. In another word, when searching for valued content, it is important to look for entropy, rather than observe controversies in comments. Such finding can be a huge help in auto-assessing fast growing online discussions.

4.1.3 Comments Amplify Explanatory Power

We found that the number of unique users who provide comments, their reputation, and the timing of the comments positively relate to the constructiveness of the answer (Table 4.2). Answers that attract more commenters, with higher median reputation and that garner half of the comments in longer periods of time have higher odds to have above median constructiveness. Such finding again confirms literature that good quality content provokes cognitive process is the one involves with the authorship. Among the control variables, the number of editions and the answer length are also related to constructiveness. Longer answers and those with more edits are associated with above median constructiveness. Together the results indicate the characteristics of the comments help to explain the variability of both the number of votes and constructiveness, thus providing support to our main hypothesis that comments could be used as a proxy to find value in answers.

4.1.4 Comments Facilitate Editing Process

As we discussed earlier (section 3.4), users actually put in efforts in giving answers comments, do those efforts yield in good answers that provoke learning? We have already found that the comments semantics are associated with the answers votes, but not with constructiveness (Table 4.1, column (2) & Table 4.2 (2)). Such results were initially counter-intuitive to our understanding. However, we

Regression Analysis: Votes			
		(1)	**(2)**
	Variables	**Number of votes**	**Number of votes**
Post Variables	Answer length (log)	0.035	-0.165***
	With code	0.236***	0.206***
	Time to answer the question (log)	-0.006	-0.006
	Alternative answers	-0.024	-0.017
	Edit count	0.162***	0.127***
Comment Variables	Number of unique commenters		0.295***
	Median time of comments (log)		0.098***
	Median reputation of commenters (log)		-0.000
	Ratio comments to commenters		-0.027
	Ratio negative comments to comments		0.003
	Mean entropy of comments		0.070**
	Observations	**2983**	**2951**
	Pseudo R-squared	**0.094**	**0.155**
	Exponentiated coefficients * p<0.05, ** p<0.01,***p<0.001		

Table 4.1: Regression Analysis for Answers' Votes, Column (1) Constitutes Only Post Semantic Features, Columns (2) Considers Both Post and Comment Semantic Features.

Regression Analysis: Constructiveness			
		(1)	**(2)**
	Variables	**Above Median Constructiveness**	**Above Median Constructiveness**
Post Variables	Answer length (log)	1.294***	1.437***
	With code	1.054	1.084
	Time to answer the question (log)	1.044	1.027
	Alternative answers	0.961	0.983
	Edit count	1.399***	1.434***
Comment Variables	Number of unique commenters		0.819**
	Median time of comments (log)		1.046**
	Median reputation of commenters (log)		0.934*
	Ratio comments to commenters		0.937
	Ratio negative comments to comments		1.034
	Mean entropy of comments		1.166
	Observations	**2983**	**2951**
	Pseudo R-squared	**0.046**	**0.054**
	Exponentiated coefficients * p<0.05, ** p<0.01,***p<0.001		

Table 4.2: Regression Analysis for Constructiveness, Column (1) Constitutes Only Post Semantic Features, Columns (2) Considers Both Post and Comment Semantic Features.

reason that votes and comments semantics are strongly connected because both of them represent the readers perceptions of an answer. On the other hand, constructiveness of an answer could only be affected by the comments semantics if the answers are edited as a result of the comments. Therefore, we now turn to assess whether there is a connection between comments and edit count. Table 4.3 shows that indeed there is an association between the mean entropy of comments and the edit count. The less entropy in the comments, the more the number of edit counts. A unit of increase in the mean entropy of an answers comments is expected to decrease the count of edits by a factor of 0.577. In plain English, when there is a lack of point of views in the comments, answers tend to remain the way it is without editing. Without community shepherds discussions based on the feedback from comments, it is harder to shape answers thought-provoking content.

4.1.5 Comments Make Answers Become Attractive

Other aspects of an answers comments are also factors on the number of edits. The number of unique commenters, the median time of comments and the ratio of comments to commenters are expected to increase the count of edits. For example, a unit of increase in the number of unique commenters is predicted to increase the number of edits by a factor of 1.238. When the ratio of comments to authors increases by one unit, the count of edits is expected to increase by a factor of 1.145. Additionally, the longer it takes for the answer to obtaining half of its comments, the higher the number of edits. We interpret all of these independent variables as attention attractors, which represent how attractive the answer is. The more people are involved, the more comments the community can possibly provide and the longer the answer can still receive comments. These are all hints that the answer

Poisson Regression: Edit Count			
		(1)	**(2)**
	Variables	**Answer's Edit Count**	**Answer's Edit Count**
Post Variables	Answer length (log)	1.341***	1.120**
	With code	2.497***	2.349***
	Time to answer the question (log)	1.029	1.033
	Alternative answers	1.097***	1.048**
Comment Variables	Number of unique commenters		1.238***
	Median time of comments (log)		1.056***
	Median reputation of commenters (log)		1.040
	Ratio comments to commenters		1.145***
	Ratio negative comments to comments		0.931
	Mean entropy of comments		0.577***
	Observations	**2983**	**2951**
	Pseudo R-squared	**0.063**	**0.086**
	Exponentiated coefficients * $p<0.05$, ** $p<0.01$,*** $p<0.001$		

Table 4.3: Poisson Regression Analysis for Answers Number of Edits, Column (2) Constitutes Comment Semantic Features, Column (1) Does Not.

has attracted attention, and therefore, there is more chance that the answer is further improved through more edits.

4.2 The Effects of Comment Quality

Finally, we conducted two regression analysis to assess the comments quality in terms of votes and constructiveness. Given that only a small proportion of comments had a total count of votes larger than zero, we defined a binary dependent variable to represent whether a comment had positive votes or no. We conducted a logistic regression to model this binary indicator. On the other hand, comment constructiveness was a count variable, therefore, we used a Poisson regression to estimate the effect of the independent variables on the constructiveness score. The results indicate that the larger the entropy score of a comment, the more likely that the comment will have a positive count of votes. A unit of increase in entropy is expected to increase the probability of the comment to have positive votes by a factor of 1.431 (Table 4.4 column (1)). However, the controversy score of the comment is not significantly related to the votes. Additionally, longer and earlier comments are more likely to have a positive count of votes. Meanwhile, the pattern of influence of the comments semantics is the opposite. While entropy does not have a statistically significant relationship with constructiveness, controversy does. More controversial comments are associated with more constructive comments. An increase of a unit in the controversy score is predicted to increase the count of constructiveness score by a factor of 1.036 (Table 4.4 column (2)). Together, these results mean that when searching for feedback to improve comments, people are likely to get provoked by controversial content. However, the community tends to cast the votes to appreciate different points of views, but not necessarily opposite points of view.

Regression: Votes & Constructiveness		
	(1)	**(2)**
Variables	**Comment with positive vote? (yes/no)**	**Comment constructiveness**
Comment length (log)	1.106*	1.562***
Time to comment the question (log)	0.906***	1.002
Controversy	1.069	1.036***
Entropy	1.431**	0.941
Observations	**9569**	**9569**
Pseudo R-squared	**0.011**	**0.075**
Exponentiated coefficients * p<0.05, ** p<0.01, *** p<0.001”		

Table 4.4: Regression Analysis of Comments Votes and Constructiveness

Chapter 5

CONCLUSIONS

5.1 Summary

- The evaluation section of this paper provides substantial evidence in considering comment semantic features to distinguish the quality of posts. Comment controversy can be an index in judging quality, however, entropy helps find the good quality content.
- By referring to some of the past research work on StackOverflow, it can be stated that StackOveflow's content quality detection features are reliable.
- Diverse comments attract more positive votes than controversial comments.

5.2 Conclusion

In this work, we investigate comment semantics from StackOverflow site and how does the semantic dynamics affect Answer quality. We survey the relevant literature and employ the key content quality identification features. We then construct comment semantics features and build several regression models to explore the value of comment semantics dynamics. We examine the value of comments for large-scale of discussion forums. We not only reconfirm the usefulness of several essential quality predictors, including time, reputation, length, and editorship, we also found the results show that comment semantics are valuable to shape the answer quality. They increase the model explanatory power and reassure importance of the commenting-editing ecosystem for large-scale programming discussion forums. Specifically, the diversity of comments contributes to the answer quality. More importantly, when searching for valued content, it is important to look for global semantics dynamics (entropy), rather than observe local differences (controversy). Although, controversial comments are not significant indicators to post quality, the disputable comments leave traces of commenters efforts. Such traces can be valuable resources for future readers, such as commenters trial and error experiences. Moreover, the presence of comments shepherd the community to revise the posts by attracting attentions to the posts and eventually facilitate the editing process. Currently, we only consider limited set of social features, such as reputation. We disregard several others, such as badges, tags etc. We emphasize that the goal of this study is to understand the value of comments to the posts, rather than predicting post quality. In the future, more comprehensive features can be considered. Most importantly, we plan to investigate how much students can learn by consuming different comments-supported content and what other semantics structure will affect content quality.

5.3 Future Work

As part of the future work, we would like to address the issue of StackOverflow's discrepancies in the rating system by identifying such posts before analyzing its quality.

The data considered for the analysis of this research is StackOverflow's java posts. Extending our method of evaluating quality to other programming language posts would make this research further reliable.

Assessing the scope of extending this research to various other online discussion platforms.

REFERENCES

[1] Agrawal, A., J. Venkatraman, S. Leonard and A. Paepcke, “Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips”, (2015).

[2] Anderson, A., D. Huttenlocher, J. Kleinberg and J. Leskovec, “Discovering value from community activity on focused question answering sites: a case study of stack overflow”, in “Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining”, pp. 850–858 (ACM, 2012).

[3] Baccianella, S., A. Esuli and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.”, in “LREC”, vol. 10, pp. 2200–2204 (2010).

[4] Breslow, L., D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho and D. T. Seaton, “Studying learning in the worldwide classroom: Research into edx’s first mooc”, Research & Practice in Assessment **8** (2013).

[5] Chi, M. T. and R. Wylie, “The icap framework: Linking cognitive engagement to active learning outcomes”, Educational Psychologist **49**, 4, 219–243 (2014).

[6] Dalip, D. H., M. A. Gonçalves, M. Cristo and P. Calado, “Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow”, in “Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval”, pp. 543–552 (ACM, 2013).

[7] Dray, G., M. Plantié, A. Harb, P. Poncelet, M. Roche and F. Trousset, “Opinion mining from blogs”, International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) **1**, 205–213 (2009).

[8] Esuli, A. and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining”, in “Proceedings of LREC”, vol. 6, pp. 417–422 (Citeseer, 2006).

[9] Farzan, R., J. M. DiMicco and B. Brownholtz, “Spreading the honey: a system for maintaining an online community”, in “Proceedings of the ACM 2009 international conference on Supporting group work”, pp. 31–40 (ACM, 2009).

[10] Ganapathibhotla, M. and B. Liu, “Mining opinions in comparative sentences”, in “Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1”, pp. 241–248 (Association for Computational Linguistics, 2008).

[11] Hsiao, I.-H. and F. Naveed, “Identifying learning-inductive content in programming discussion forums”, in “Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE”, pp. 1–8 (IEEE, 2015).

[12] Liu, B., *Sentiment analysis: Mining opinions, sentiments, and emotions* (Cambridge University Press, 2015).

[13] Mamykina, L., B. Manoim, M. Mittal, G. Hripcsak and B. Hartmann, "Design lessons from the fastest q&a site in the west", in "Proceedings of the SIGCHI conference on Human factors in computing systems", pp. 2857–2866 (ACM, 2011).

[14] Margulieux, L. E., M. Guzdial and R. Catrambone, "Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications", in "Proceedings of the ninth annual international conference on International computing education research", pp. 71–78 (ACM, 2012).

[15] Munezero, M., C. S. Montero, M. Mozgovoy and E. Sutinen, "Exploiting sentiment analysis to track emotions in students' learning diaries", in "Proceedings of the 13th Koli Calling International Conference on Computing Education Research", pp. 145–152 (ACM, 2013).

[16] O'Connor, B., R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.", ICWSM **11**, 122-129, 1–2 (2010).

[17] Patitsas, E., M. Craig and S. Easterbrook, "Comparing and contrasting different algorithms leads to increased student learning", in "Proceedings of the ninth annual international ACM conference on International computing education research", pp. 145–152 (ACM, 2013).

[18] Plutchik, R., "A general psychoevolutionary theory of emotion", Theories of emotion **1**, 3–31 (1980).

[19] Rashid, A. M., K. Ling, R. D. Tassone, P. Resnick, R. Kraut and J. Riedl, "Motivating participation by displaying the value of contribution", in "Proceedings of the SIGCHI conference on Human Factors in computing systems", pp. 955–958 (ACM, 2006).

[20] Schwartz, D. L. and J. D. Bransford, "A time for telling", Cognition and instruction **16**, 4, 475–5223 (1998).

[21] Siersdorfer, S., S. Chelaru, W. Nejdl and J. San Pedro, "How useful are your comments?: analyzing and predicting youtube comments and comment ratings", in "Proceedings of the 19th international conference on World wide web", pp. 891–900 (ACM, 2010).

[22] Somasundaran, S., J. Ruppenhofer and J. Wiebe, "Detecting arguing and sentiment in meetings", in "Proceedings of the SIGdial Workshop on Discourse and Dialogue", vol. 6 (2007).

[23] Tian, Q., P. Zhang and B. Li, "Towards predicting the best answers in community-based question-answering services.", in "ICWSM", (2013).

[24] Tian, Y., P. S. Kochhar, E.-P. Lim, F. Zhu and D. Lo, "Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting", in "Social Informatics", pp. 55–68 (Springer, 2013).

[25] Treude, C., O. Barzilay and M.-A. Storey, "How do programmers ask and answer questions on the web?: Nier track", in "Software Engineering (ICSE), 2011 33rd International Conference on", pp. 804–807 (IEEE, 2011).

[26] van de Sande, C., "Free, open, online, mathematics help forums: the good, the bad, and the ugly", in "Proceedings of the 9th International Conference of the Learning Sciences-Volume 1", pp. 643–650 (International Society of the Learning Sciences, 2010).

[27] van de Sande, C. and G. Leinhard, "Online tutoring in the calculus: Beyond the limit of the limit", education **1**, 2, 117–160 (2007).

[28] Vasilescu, B., "Software developers are humans, too!", in "Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing", pp. 97–100 (ACM, 2014).

[29] Vasilescu, B., V. Filkov and A. Serebrenik, "Stackoverflow and github: Associations between software development and crowdsourced knowledge", in "Social Computing (SocialCom), 2013 International Conference on", pp. 188–195 (IEEE, 2013).

[30] WALES, N. S., "Detecting, representing and querying collusion in online rating systems", arXiv preprint arXiv:1211.0963 (2012).

[31] Wen, M., D. Yang and C. P. Rosé, "Sentiment analysis in mooc discussion forums: What does it tell us", Proceedings of educational data mining **1** (2014).

[32] Yao, Y., H. Tong, T. Xie, L. Akoglu, F. Xu and J. Lu, "Want a good answer? ask a good question first!", arXiv preprint arXiv:1311.6876 (2013).