Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5383–5391
Marseille, 11–16 May 2020

Assessing Users' Reputation from Syntactic and Semantic Information in Community Question Answering

Yonas Woldemariam
Dept. Computing Science, Umeå University, Sweden
yonasd@cs.umu.se

Abstract

Textual content is the most significant as well as substantially the big part of CQA forums. Users gain reputation for contributing such content. Although linguistic quality is the very essence of textual information, that does not seem to be considered in estimating users' reputation. As existing users' reputation systems seem to solely rely on vote counting, adding that bit of linguistic information surely improves their quality. In this study, we investigate the relationship between users' reputation and linguistic features extracted from their associated answers content. And we build statistical models on a Stack Overflow dataset that learn reputation from complex syntactic and semantic structures of such content. The resulting models reveal how users' writing styles in answering questions play important roles in building reputation points. In our experiments, extracting answers from systematically selected users followed by linguistic features annotation and models building. The models are evaluated on in-domain (e.g., Server Fault, Super User) and out-domain (e.g., English, Maths) datasets. We found out that the selected linguistic features have quite significant influences over reputation scores. In the best case scenario, the selected linguistic feature set could explain 80% variation in reputation scores with the prediction error of 3%. The performance results obtained from the baseline models have been significantly improved by adding syntactic and punctuation marks features.

Keywords: Reputation Prediction, Community Question-Answering, Competence Analysis, Syntax and Semantics

1. Introduction

Although many CQA forums aim to provide reasonably high-quality content, their underlying mechanisms to control the quality and reliability of such content lack lots of important features (Shah and Pomerantz, 2010). Particularly, linguistic features are quite pertinent to the very nature of such content, as textual content is the most significant as well as substantially the big part of CQA forums. Moreover, even if it is quite apparent how linguistic quality contributes for gaining reputation, existing reputation systems do not seem to take such quality into account in reputation scores estimation.

General CQAs' data usage trends show that the number of users actively visiting such sites and the rate at which users regularly posting content are dramatically increasing (Baltadzhieva and Chrupala, 2015b). For instance, right now, the StackExchange (SE) Stack Overflow's traffic shows it is receiving nearly 10 million page views and 6.4k questions per day, there are also many forums in SE with almost 100% question answering rate. Despite such huge amount of data are consumed by constantly growing number of users, the data suffers from serious quality (e.g., inaccuracies and grammatical errors) and provenance (e.g., data ownership and trust- worthiness) issues that potentially mislead peoples looking for right answers for their questions.

In many CQA sites including SE, users' reputation is computed by simply counting votes given for posts (detailed in **Section 4**) without analyzing their actual content. Since largely the actual content generated by users is natural text, ensuring the linguistic quality of the text might add more evidence and improve data trustworthiness assessment mechanisms. That in turn enhance the quality of the entire services (recommendation and ranking) provided by the platform.

Probably, the reason why the linguistic quality of user-generated content seems to be ignored in estimating reputation scores, is the relationship between users' content and their associated reputation is not clearly understood. While there exists little research (Chen and He, 2013; Dascalu et al., 2008; Chen et al., 2014) on predicting user-specific features from on-line forums posts or other sources of text, the vast majority of studies on CQA sites focus on content related attributes, particularly content quality. Some focus on assessing question quality (Baltadzhieva and Chrupala, 2015a), whereas others focus on answer quality (Shah and Pomerantz, 2010). Although features considered in these studies are important to characterize content quality, they do not give clear clues how they affect users' reputation. Moreover, most of these features are non-textual surface features (meta-information) that can be directly extracted from CQA sites, rather than linguistic information hidden in the actual textual content.

In this study we assess users' reputation on the basis of linguistic features (syntactic, semantic, and punctuation marks) extracted from CQA textual content. The research is an extension of our previous work (Woldemariam et al., 2017) where we attempted to predict users' competence from linguistic data collected from the crowd sourcing platform Zooniverse, in particular Galaxy Zoo and Snapshot Serengeti. We aim to further investigate the methods presented in (Woldemariam et al., 2017) using SE datasets. Unlike Zooniverse, SE has rich features that are directly associated with users' performance and content quality. Among these features, reputation is widely regarded as a measure of competence in answering and asking questions (MacLeod, 2014; Zhang et al., 2007).

In our experiments, multiple linear regression models have been trained on various configurations of the selected linguistic features sets, validated and evaluated on in-domain (e.g., Server Fault, Super User) and out-domain (e.g., English, Maths) datasets. We found out that linguistic features have quite significant influences over reputation scores, according the performance evaluation of the learned models.

In all test cases, we obtained **the coefficient of determination (R-squared)** ranges 0.63-0.80 and normalized **root mean squared error (RMSE)** ranges 0.03-0.98.
We summarize related work in Section 2, the discussion of the notion and estimation of reputation in Section 3, the feature selection and extraction, and experimental setup in Section 4, the evaluation results and analysis in Section 5 and 6 respectively, finally, the conclusion and future work in Section 7.

2. Related work

There exist several studies which conduct analytics over CQA forums archives with a purpose of improving content quality using both textual and non-textual features. The vast majority of them focus on predicting content related features rather (e.g., answers/question quality) than user-related features (e.g., reputation) by using various natural language processing (NLP) techniques and tools. Here, we focus on literature that attempt to exploit particularly textual content to improve various services provided by CQA forums.
In addition to that, since it seems to be difficult to find literature on predicting reputation from users' textual content, we consider studies carried out on other types of forums and sources of text such as crowd sourcing sites and medical reports, where authors attempt to predict user-specific attributes, particularly users' performance (competence) from linguistic features. The reason we are stressing on the competence aspect of users is, as mentioned in **Introduction**, among other features in CQA forums, reputation is the fairly good representative and indicator of users' competence (expertise or performance) in CQA (MacLeod, 2014; Zhang et al., 2007; Movshovitz-Attias et al., 2013). And we're particularly interested in better understanding the relationship between users' competence and their associated text.
Compared to other on-line discussion forums such as short-message based chat rooms (or social media) CQA forums seem to have a large body of text characterized by longer and complete sentences, and rich linguistic information. That makes convenient for tasks of modeling and predicting content (questions or answers) quality using various NLP techniques. Baltadzhieva and Chrupala in (Baltadzhieva and Chrupala, 2015b), present a survey of papers on evaluating the quality of questions. That summarizes early research works (Suryanto et al., 2009), as well as recent ones (Asaduzzaman et al., 2013) over the years (2006-2015). The survey presents various methods of how content quality is perceived, interpreted and measured within CQA forums. Authors in (Li et al., 2012) provide good explanations of the notion and main aspects of content quality in CQA. According to the survey, among other question-related features, the number of answers received and a question score (calculated from up/down votes) are primarily used metrics to measure questions quality. These metrics in (Correa and Sureka, 2014) also used to identify and reason about unanswered questions, and play a great role in the decision of deleting questions on CQA sites. The authors also indicate that these metrics are influenced by textual features (e.g., number of sentences, word counts). Similarly, the most significant question-related features (e.g., questions' length and tags) that are widely used for predicting questions quality have been identified in the survey.
Researchers have also explored methods for predicting answers quality by extracting meta -information and linguistic features from questions-answers (QA) pairs. And statistically analyzed with machine learning tools. Jeon et al. (Jiwoon et al., 2006) for example, predict answers quality based on textual features (e.g., answer length) and non-textual features (e.g., answer view counts). They trained statistical models on manually annotated (as good or bad) answers. Thereby, human annotators judge the quality of answers by looking at the corresponding QA pairs on the basis of for example, relevance and clarity. Their experimental results indicate that the models are able to detect more of good answers than bad ones due to data imbalance (good answers have higher prior probability than bad ones) in the training set. Moreover, their feature analysis also shows that the length of answers is the most correlated feature with answer quality scores than others. In comparison, authors in (Hu et al., 2016) give more attention to textual features (e.g., bag-of-words, average sentence length) than Jeon et al. and have richer information in their probabilistic models predicting answer quality. They experimented with varied combination of features and their results indicate that the classifiers (based on logistic regression) with linguistic features outperform than the models trained on non-textual features extracted from profiles of users. According to their report, word frequency count in QA pairs is most significant than other extracted linguistic features.
Other studies (Chen and He, 2013; Woldemariam et al., 2017; Dascalu et al., 2008; Chen et al., 2014) have also explored methods for predicting user-specific attributes, particularly users' competence in various types of tasks (e.g., essay writing, classifying images, writing medical reports and so on). These studies assume that the level of the expertise of the users performing such tasks can be predicted from their text written in connection with their tasks. For example, authors in (Chen et al., 2014) attempt to infer the performance of medical students from their clinical portfolio by using machine learning methods. The authors trained various classifier models on linguistic (e.g., bag-of-words) and meta-data (e.g., the number of clinical notes) features extracted from an annotated corpus containing clinical notes. Their results indicate that the performance of the models vary across the selected competence domains.
Authors in (Woldemariam et al., 2017), investigate the problem of assessing the proficiency of volunteers in classifying images in a crowd-sourcing discussion forum from their chat messages. Thereby, the quality (competence level) of each volunteer has been estimated based the overall accuracy of the classifications of images made by the user, using a weighted majority voting scheme. The authors considered six different sets of linguistic features: bag-of-words, punctuation marks features and syntactic. And employed three different machine learning algorithms: k-Nearest Neighbors, Naive Bayesian, and Decision Trees (with gradient boosting) to measure to what degree high and low (average) competent users can be inferred from

their text. Their validation results indicate that the trained models are statistically significant across all feature set configurations. Their resulting models have been also evaluated on two related domain test sets and yielded consistent performance results that the models trained on bag-of-words and punctuation mark features provide reasonable results. Moreover, due to the nature of the forum where texts posted by volunteers tend to be quite short, and the quality measure (the ground truth estimation) of volunteers relies on majority voting and other heuristics and pragmatic decisions, the trained models get constrained.

3. Users' Reputation and Estimating Reputation Scores

Generally, CQA platforms consist of different integrated and interdependent components. Among other components, the reputation estimation component involves determining the quality of users as well as their content. Despite the multitude of various CQA forums in the world today, all share common logics to assign reputation scores for their users. Similarly, there seem to exist common understanding of reputation in many literature (MacLeod, 2014; Zhang et al., 2007; Shah and Pomerantz, 2010) as a measure of users' performance in a particular area of expertise. Like any award systems, they are assumed to bestow reputation points to users for their eloquently written answers/questions. However, they do not seem to have content analysis mechanisms that really deal with the actual content to make the distinction between high and low quality content (detailed in the below paragraphs).

In SE, users' reputation score is basically estimated from votes received for performing any of the three main activities in the SE communities: asking, answering and editing questions/answers. Detailed information on other activities in connection with reputation such as bounty awards and reputation limits, can be found in the SO[1] site. While getting up-votes from other users for any of such activities leads a user to gain reputation points, these activities carry different weights. Questions, answers and edits carry 5, 10 and 2 reputation points, respectively.

For example, a single vote on an answer awards 10 points for a user (answerer) posting that answer, and the total reputation point for that particular answer is then the product of the number of up-votes multiplied by 10. That holds true for all types of posts (questions, answers and edits), except their weights. Answerers get additional 5 points if their answer is accepted among other alternative answers provided for the same question. Contrarily, voting down results in users to lose their reputation scores. The overall reputation score for each user is computed by summing up all the scores earned from his/her posts into a single reputation score.

The overall reputation score also used to allow/block users to perform other activities (e.g, voting up and down, and commenting questions/answers). While the basic activities mentioned above can be performed by any registered user without any prior reputation requirement, voting up, down and commenting require at least 15, 50 and 125 reputation points, respectively.

Obviously, the reputation estimation mechanism is completely dependent on a voting scheme. Not only that, the effect goes to the ranking system where the answer with the highest vote get ranked first and is also marked as the best answer, no matter what the real quality of such answers. Moreover, that gets contagious and affects the recommendation and other components of the CQA system. That eventually, degrade the entire content quality of the site as other users tend to consider and accept answers provided by users with high reputation scores. The fact that CQA sites use the voting is, to make their site democratic and be governed by users, although it does not guarantee content quality.

Establishing trust mechanisms and relying solely on voting has several drawbacks, particularly on the quality of the content within CQA sites. Even if some of the content voted by many users might have reasonable quality in terms of accuracy and language, there is broad subjectivity among voters. Some users take time to check both the correctness and the overall quality of the content, whereas others might just focus only on the technical correctness of answers or relevance of questions. Therefore, reputation scores can be taken as expressions of subjective opinions of the voters. As a result, while the estimated user reputation scores provide rough clues about the trustworthiness of users as well as their content, that can be further improved or adjusted by using additional mechanisms that explicitly deal with the content quality.

As observed from the reputation system, answers are assigned the highest weights (10 and 15 points). That somehow leads to an intuitive conclusion that reputation points gained from answering question better show certain level of users? expertise than asking questions or editing posts. Therefore, for our users? reputation modeling task we focus on answers content.

4. Experimental Setup

4.1. Data

We collected data-dumps directly from the SE content repository[2]. While SE provides datasets archived (as XML files) from its 173 communities, only six datasets from various communities (categories) are used for our experiments. These categories include Stack Overflow, Server Fault, Super User, Ask Ubuntu, Mathematics and English. Stack Overflow (SO) is the oldest, and largest site among other SE' communities in terms of the number of users and number of answered questions. Because of that, it gives a high chance of getting statistically sufficient (large) amount of data required for training and evaluating models. On the other hand, we choose the remaining to evaluate our models on various (related and unrelated) domains. While Server Fault, Super User, Ask Ubuntu are examples of related domains with SO, Mathematics and English for unrelated ones. Evaluating models on such domains helps to see how well the models perform across various domains and leads to better conclusions about models' performance.

[1] https://stackoverflow.com/help/whats-reputation

[2] https://archive.org/details/stackexchange

We extracted about 361, 898 answers (nearly 51% of them got approved as accepted answers) from 17, 381 unique users. The entire text corpus contain 3 million sentences and 32 million words (around 8 % of them are unique words).

Following some pre-processing tasks that include parsing the collected XML data-dumps and loading into MongoDB[3], XML files representing database tables have been stored as MongoDB collections where each user corresponds to a MongoDB document. Since MongoDB is capable of supporting dynamic schema and relatively efficient (compared to others particularly NoSQL ones such as SQL Server, though commonly used to import SE data), the data-dumps have been transferred to MongoDB databases without re-defining or changing their structures. From each domain, we target those users having at least 5 answers to get reasonably sufficient amount of text from each user, and 1 of their answers must be accepted by other users.

Applying such threshold values also help intensify our focus of competence and filter those users who relatively (compared to other users who just only ask questions) exhibit some kind of expertise and commitment in answering questions. In addition to setting apart such users, answer related features that could best measure the quality of answers and characterize users' competence are extracted from users' profiles and included in our feature sets.

Then, datasets are randomly divided into three subsets using shuffled sampling: training, validation (development), and evaluation sets, 70%, 10% and 20% of the whole corpus, respectively.

4.2. Extracting Linguistic and Non-Linguistic Features

After the extraction of answers content from the selected users, the following linguistic and non-linguistic features have been extracted: syntactic (Syn),bag-of-words (BoW) and punctuation marks (Pun).

4.2.1. A Syntactic Feature Set (Syn)

Apart from answer related features (e.g., up/down vote scores) that can be directly extracted from users' profiles, considering the structures of such answers via syntactic parsing help further reveals answers quality. Moreover, the resulting syntactic features essentially help to make the distinction between answerers based on the grammatical behavior of their answers. For instance, users belonging to a certain range of reputation scores might follow a particular syntactic pattern (e.g., having many occurrence of adjectives(JJ) or declarative statements (S)) in their answers, whereas such categories might rarely occur in the answers of other groups of users. In general, that potentially provides some clues about the descriptiveness of users' answers and, eventually leads to a conclusion about their quality.

The Stanford shift-reduce parser (SRP) (Zhu et al., 2013) along with the Stanford CoreNLP (Manning et al., 2014) toolkit has been used for linguistic information annotation, constituency and dependency parsing. While constituency parsing builds phrase-structured parse trees from answers and extract syntactic categories (e.g., VB (verb), NN (noun)), dependency parsing constructs dependency-based parse trees where the dependency relationships between syntactic units (words) can be extracted(Marneffe et al., 2006; Nivre et al., 2016).

Computationally, it has been quite challenging to extract such syntactic features from 3M sentences. Therefore, in order to make the computation faster, we made practical decisions and choices during our experiments. For instance, switching from the Stanford probabilistic context-free grammar parser (Klein and Manning, 2003) to SRP and multi-threading the parsing process have significantly improved the performance. The aggregated answers from each user have been parsed together (as a single document) and characterized with the frequency counts of the extracted syntactic features. We have also analyzed how these features are related with reputation scores and, statistically the most significant one is illustrated in Figure 2(b).

4.2.2. A Bag-of-Words Feature Set (BoW)

The Bag-of-Words features provide information of frequencies and distributions of individual words in text (a corpus of answers). These features help identify most important terms that are quite frequent in individual user's answers as well as the entire dataset. Such terms, then used to measure whether there exist any similarity between users' answers content. Finding reasonable degree of similarity (pairwise word-vectors similarity) between answers' content received from users with close reputation scores, provides important evidence of causality between users' content and their associated reputation. Learning such relationship leads models to effectively assess (predict) users' reputation based on only the content of their answers. The assessment results could be taken as baseline estimated scores, and supplementing with other meta-features (e.g., number of votes) further improves the estimation.

Extracted BoW features results in a document-term matrix where each user's answers are represented with a row-vector. The size of a BoW feature set is as big as the number of unique words in the entire corpus of text. BoW features have numeric values of TF-IDF (term frequency inverse document frequency). TF-IDF is defined as the product of the frequency count of a word (W) in a document or simply TF (term frequency) and TDF (inverse document frequency). Where, IDF is the log of the ratio of corpus's size (documents' frequency count in a corpus) to the number of documents in which the word (W) occurs. While TF values indicate the relative importance of words in a particular document, IDF shows (weights) words' importance in a collection of documents (a corpus).

4.2.3. A Punctuation Marks Feature Set (Pun)

Punctuation marks have also been included in our experiments as part of linguistic information. However, the answers extracted from SO characterized by code snippets containing several programming symbols or a character set. Therefore, in order to capture both types of information, we attempted to combine punctuation mark features with some special characters that quite often occur in code snippets, and build a new feature set, though the set is not exhaustive.

[3]https://www.mongodb.com/

By taking both the individual feature set and their combinations, we came up with 6 feature set configurations: Bag-of-Words (BoW),punctuation marks (Pun), punctuation marks with Bag-of-Words (Pun+BoW), syntactic, syntactic with Bag-of-Words (Syn+BoW), and the combination of BoW, punctuation mark and syntactic (BoW+Pun+Syn).

4.3. Non-Linguistic Features

In order to enrich the selected linguistic features, we also consider non-linguistic meta-features that can be extracted from users' profiles. However, in our experiment, we only give more emphasis for those features that are intimately connected with users' answers.

4.3.1. Answer Related Features

The following are the non-linguistic features that best define the competence and commitment of users in relation to their activities with answering questions. Moreover, they are mostly used to measure content quality in CQA.

Number of answers (NoA) gives the total number of answers received from users for various questions. The threshold answer count value has been set to be at least 5 (as explained in Section 3).

Number of accepted answers is the number of accepted answers by other users (approved and marked as accepted). Only those users with at least 1 accepted answers have been considered.

NoA with >=100, 25, 10 votes counts answers up voted by at least 100, 25, or 10 users.

4.4. Training, Validation and Evaluation

Since we are aiming to solve a regression problem where the target variable (i.e., reputation) is continuous, multiple linear regression models (Freedman, 2005; Yan and Su, 2009) that learn such numeric value from the selected (non) linguistic features are trained. During training, model parameters i.e., slope coefficients and bias (intercepts) are computed using the training set, optimized using the validation set and evaluated with well established metrics. In our study, these parameters help understand which linguistic features could potentially explain the variation in reputation scores as well as their relative significance.

Models' performance has been measured with RMSE (root mean squared error) (Chai and Draxler, 2014) and $(R)^2$ (r-squared) also known as coefficient of determination (Yan and Su, 2009). They are widely used metrics to measure how well (goodness of fit) regression models perform (Freedman, 2005). While RMSE measures the differences or errors (aka residuals) between actual values and predicted values, $(R)^2$ estimates the degree to which the selected features explain the variation in the target variable, or simply the strength of the relationship between actual values and predicted values. The former ranges 0 (the best value)-positive infinity, the later 0-1 (the best value 1).

4.4.1. Training

Models' training has been done in two consecutive stages. The first stage trains baseline models and the second stage generates optimized models (explained in the next subsection). Various regression models have been trained on each feature set configuration.

4.4.2. Model Significance Tests and Validation

Following training baseline models, we have checked whether each model is statistically significant. The significance test involves determining whether there is a significant linear or non-linear relationship between reputation scores and the selected linguistic features. Among existing methods, a null-hypothesis (H_0) test is widely used (Alexopoulos, 2010). The test assumes that the slope of the models is 0. That means the selected linguistic features do not have any relationship with reputation. In order for rejecting the null-hypothesis and accepting a model, at least one of the features needs to have a positive or negative slope and a significant relationship with reputation scores with a threshold alpha value ($p - value$) of 0.05. In our experiments, each model satisfies the test, while some of those features (the least relevant) which fail to meet such value have been identified and removed during optimization.

Although the baseline models' evaluation results on the development set confirm their validity, in order to further improve their quality, each trained model has been validated and optimized using the development set. During optimization the most relevant linguistic features that potentially reduce errors and yield least squares (RMSE) have been identified from each feature set and used to train the optimized models. The validation results have been summarized in **Table 1**. Also, in the validation phase, models' coefficients have been (re) calculated to fit the reputation scores in the development set. Most of the models have been improved after the optimization both in terms of RMSE and R-squared.

4.4.3. Evaluation and Results

The learned models have been evaluated on test sets from 6 different domains of forums (categories) of the SE network including the SE dataset put aside for an evaluation purpose during the split of the entire corpus. There are 6 domains in SE: technology, art/life, culture/recreation, science, professional and business.

Primarily, our test cases aimed to target forums containing questions where a high percentage (almost 100%) of them got answered to maximize the chance of getting users who could meet the criteria set during building the corpus. Unfortunately most of such forums turn out to have less numbers of users compared to our main test set. Therefore, we selected forums that have the largest number of users from each domain. Evaluating the models on such heterogeneous domains has given a good insight and generalization on how well these models perform across in-domain (e.g., SO), related domains (e.g., Server Fault and Super User) and out-domains (English and Mathematics) datasets. For each selected forum, it has been possible to get up to 1800 distinct users that satisfy the criteria (users with at least 5 answers and 1 is accepted). To compensate the variation and standardize the measurement, RMSE scores have been normalized(briefly described in the next paragraph and further descriptions can be found in (Shcherbakov et al., 2013)).

In the evaluation experiments, we conducted the total of 36 (6 test sets * 6 feature sets) independent performance measurement runs as shown in **Table 1**. Each evaluation set representing one of the domains in SE has been evaluated across all linguistic feature sets. In principles, for conditions where test sets are different, for example in terms of domains and the range of a target variable, normalizing the RMSE scores helps make more logical comparison across such test sets (Shcherbakov et al., 2013). Depending on the cause that leads the difference most between the test sets, various normalizing techniques could be applied. For instance in our case, looking at the range of the reputation points across the test sets, shows wide gaps between them. Widely applied are normalizing by the range (Max-Reputation point - Min-Reputation point) or the mean values of the target variable of the evaluation sets. Both approaches have been considered in our study.

5. Analysis of Results and Discussion

As noted from the evaluation results presented in **Table 1**, all models score the R-squared value of ≥ 0.63 (on average 0.72), the range normalized RMSE value of ≥ 0.03 (on average 0.05) and the mean RMSE normalized value of 0.47 (on average 0.70). That means, in the former case, the 72% variation in reputation scores is due to the selected linguistic features. In the best case scenario (illustrated in **Figure 2(a)**) this score rises up to 80%, that is quite a good indication that the extracted linguistic features have a strong relationship with reputation scores. Moreover, is also possible to understand how the observed and the predicted reputation points pairs are strongly correlated by taking the square root of the R-squared values that gives the correlation coefficient (R) of nearly 0.85 and 0.90. RMSE-wise, the results seem to show that the models have low errors, that are between 0 and 1. On average, the predicted reputation points are off by such values from the actual observed reputation points. For better understanding and interpretation, this result (obtained from Stack Overflow test set) has also been compared with a simple benchmarking regression model that just predicts the mean reputation score of the training set, assuming that reputation points are normally distributed. In that case, all versions of the trained models far outperform such benchmarking model.

To the best of our knowledge no research work attempted to predict users' reputation from CQA data in from linguistic features. That makes it a bit difficult to find related benchmarking studies to compare with. Yet, we attempted to compare our results with some studies that attempt to model other important features related with reputation. For instance, Baltadzhieva and Chrupala in (Baltadzhieva and Chrupala, 2015a), build regression models to predict the quality of questions from meta-data features (e.g., questions' length and questions' tags) including reputation. Their best model has the R-squared value of 0.19 and the MSE (mean-squared error) value of 2.51. While R-squared seems to be quite a bit far from those models (including models built in this study) whose R-squared value is close to 1, the MSE suggests their model performs reasonably good. Also, authors in (Woldemariam et al., 2017) use similar linguistic features to predict users' competence from crowd sourced (not CQA) datasets, although the authors took it as a classification problem. Their performance evaluation results show that the trained classifiers learn competence from the selected linguistic features to some extent. However, since the experimental data used in (Woldemariam et al., 2017) is limited in many ways (e.g., ground-truth unavailability, text's length shortness, data scarcity), the models do not seem to be as effective as regression models trained in this study to assess users' expertise from linguistic features. More importantly, they identified potential areas where their results could be improved and some of them, for instance using dependency parsing to further enrich linguistic features, are applied in this study.

5.1. Impacts and Implications of Bag-of-Words Features over Reputation

Looking into all test cases where only BoW has been solely used, it gives the least performance values across all metrics, compared to other linguistic feature sets, as shown in **Figure 1 (a)** and **(b)**. Usually models trained on BoW gives only baseline results, unless supported with other linguistic information, though there might be exceptions in other text classification problems. That is due to the sever dependency of such models on the textual content used to train them. Also, since these content mainly contain domain-specific words, the models tend to perform poorly on other unrelated domains where such words occur rarely.

Our test results also partially confirm that truth. Evaluating the BoW-based model on SO dataset gives the best R-squared result next to Mathematics. Also, it gives the best RMSE (mean normalized), next to Server Fault and Super User. Surprisingly, yet, in terms of RMSE (range normalized), the BoW model performs poorly on the in-domain SO dataset (compared to other out-domain datasets). Probably, one of the possible reasons is, such metric favors for test sets with wide ranges of reputation points. Among other test sets, Server Fault and SO have the narrowest range of the observed reputation values. Therefore, considering the other metrics i.e RMSE (mean normalized) and R-squared avoids the bias in performance measurements and comparisons.

During models' validation, we identified the most significant terms/words (highest weighted terms assigned by regression models) that either positively (e.g., "try", "return", "void") or negatively (e.g.,"link", "edit", "answer") influence reputation scores. The majority of the terms seem to be keywords (programming constructs) of various programming and scripting languages. As these terms are extracted from the code snippets and psedcode posted by SO users, particularly from answerers, they seem to rarely appear in other domains such as English. Due to that, the models tend to better recognize words coming from related domains and get insensitive for the terms extracted from out-domain data sets like English (less effective in predicting reputation scores), compared to other related domains such as Server Fault. That implies, accompanying answers content with just a few programming terms, it is likely to achieve better predictions than using many generic terms.

Furthermore, among the highest weighted terms assigned

Metric	Domain	Linguistic Feature Set					
		BoW	Pun	Pun+BoW	Syn	Syn+BoW	BoW+Pun+Syn
R^2	Stack Overflow	0.70	**0.74**	**0.74**	0.73	0.70	**0.74**
	Server Fault	0.74	0.79	**0.80**	0.79	0.74	0.79
	Ask Ubuntu	0.68	0.69	0.69	**0.70**	0.66	0.64
	Super User	0.74	**0.79**	0.79	0.78	0.75	**0.79**
	Mathematics	0.65	0.71	0.71	**0.73**	**0.73**	0.63
	English	0.68	0.69	0.69	**0.70**	0.66	0.64
RMSE (range normalized)	Stack Overflow	0.07	**0.06**	**0.06**	**0.06**	0.07	**0.06**
	Server Fault	0.06	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
	Ask Ubuntu	0.03	0.03	0.03	0.03	0.03	0.03
	Super User	0.05	0.05	0.05	0.05	0.05	0.05
	Mathematics	0.06	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
	English	0.03	0.03	0.03	0.03	0.03	0.03
RMSE (mean normalized)	Stack Overflow	0.60	0.56	**0.55**	0.56	0.60	0.56
	Server Fault	0.53	**0.47**	**0.47**	**0.47**	0.52	0.48
	Ask Ubuntu	0.93	0.93	0.94	0.95	**0.90**	0.98
	Super User	0.58	0.54	0.53	0.53	0.58	**0.52**
	Mathematics	0.78	0.72	0.73	**0.71**	**0.71**	**0.71**
	English	0.93	0.93	0.94	0.95	**0.90**	0.98

Table 1: Models' Evaluation Results

by the regression models, some terms (e.g., "http"), indicate that, apart from providing code snippets, supporting answers content with web links to various sources play crucial roles in gaining reputation points.

We also found that answers' length (number of words) has significant influence on reputation scores than. That means, long answers are more likely to receive up votes than short answers, as they probably contain detailed descriptions, examples and so on. That seems to be quite intuitive, as such illustrations tend to make answers vivid and satisfy the desire of askers and other users looking for similar answers.

5.2. Impacts and Implications of Syntactic Features over Reputation

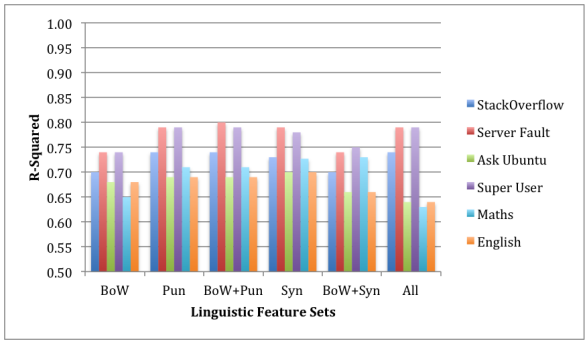
As evident from our experimental results, opposed to the content-based BoW model, models built on the syntactic feature set predict reputation effectively. In comparison, they also do not seem to be dependent on specific domains as syntactic structures are pretty much shared across different domains and languages. One of the best performance results in terms of R-squared has been obtained from Syn. RMSE (mean normalized) wise, Syn also gives the best result next to Pun and Pun+BoW feature sets, though RMSE (range normalized) does not seem to significantly change across feature sets. However, looking the results across the test sets (with respect to the domains), the models evaluation on SO gives the highest RMSE (range normalized).

Among the most significant syntactic features, tags extracted from the dependency parse greatly affected the Syn based models, compared to the constituency parse. While syntactic tags/relations (e.g., "parataxis", "mwe", "co") are positively related with reputation scores, tags/constituents (e.g., "SQ", "FRAG", "det:predet") have a negative influence on reputation. For instance, the occurrence of the dependency type "paraxis" in the dependency structure of sentences (answers), symbolizes that the parsed sentences are constructed with clauses (phrases) without being connected with linking words that coordinate them. We observe that such dependency type frequently appear in parsed CQA answers posted by highly reputable users. That could possibly provide some interesting facts about the linguistic nature of such answers as well as the reflection of other users (readers) voting the answers for that particular writing style. Firstly, such users seem to prefer to simple paratactic constructions in their answers' writing, probably for the sake of brevity. And then, the other readers seem to enjoy the simplicity and up vote such answers. Here, we assume that the reputation points gained by the answerers mainly from their answers.

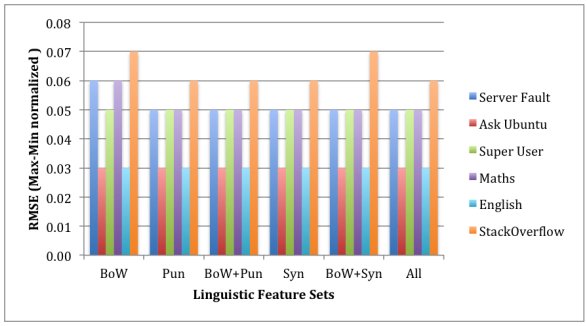
On the other hand, the constituent "SQ" appearing in constituency parse trees shows users are asking yes/no questions. However, observing such syntactic information (simply questions) in answers content seems to be quite unlikely, unless the question being answered is pretty vague. In other words, such interrogative answers do not seem to be preferred very much compared to simple declarative ones. Probably, that is the possible reason why the syntactic feature "SQ" negatively correlated with reputation scores. The regression analysis on such feature reveals that users tend to vote up more focused answers than answers containing questions, though such questions might arise from answerers for asking clarifications on the original question. Looking further into other most relevant tags also gives more insights on how the writing styles of users potentially control and affect their reputation within the CQA communities. Such relationships have been illustrated with scatter plots in **Figure 2(b)**.

5.3. Impacts and Implications of Punctuation Features over Reputation

Compared to the models trained on other linguistic feature sets, on average the model trained on Pun performs best re-

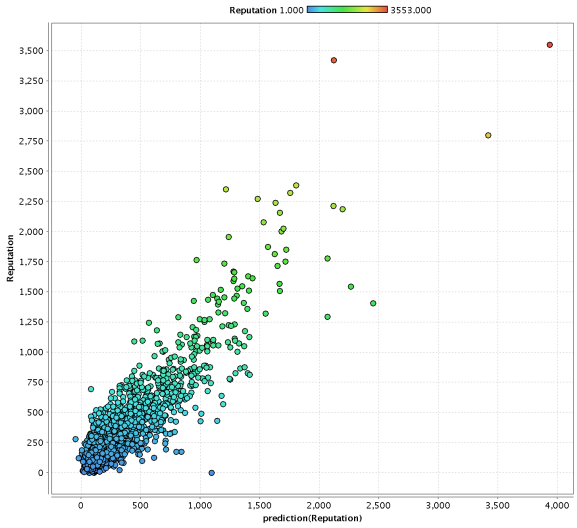


(a) R-squared

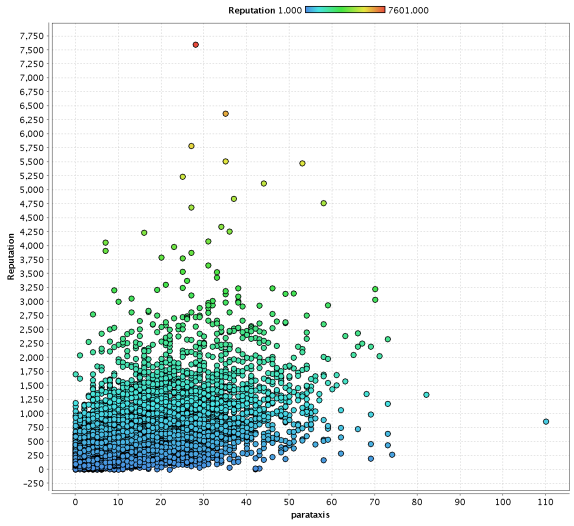


(b) RMSE (range normalized)

Figure 1: Performance Evaluation Results



(a) Observed Reputation Scores Versus Predicted Reputation (PR) Scores



(b) Reputation Scores Versus Parataxis (P)

Figure 2: Scatter Plots of Reputation Scores over PR and P

garding both versions of RMSE. R-squared-wise, there is no significant difference from its syntactic counterpart. Not only that, the baseline performance of the BoW model gets improved better when combined with Pun than Syn. The possible reason for that may be, its richness in terms of variety (e.g., common punctuation marks, special characters, character encodings). That potentially makes the Pun based models to capture the information present in both natural text and code snippets.

In the extracted punctuation mark feature set, special characters and character sets (character encodings) belong to different programming/scripting languages, cover a large part. Even if we observed that such features have a significant influence over reputation, we cannot tell exactly which type influence more than others, because of some overlapping between them. However, yet not difficult to see some characters such as tilda " " are extracted from answers containing code snippets rather than plain natural text.

Among the punctuation mark features, punctuation marks/special characters (e.g., comma, tilda) positively influence reputation scores the most, while other features such as (e.g., question mark, semicolon) negatively affect reputation scores the most. Interestingly enough, we note how the result obtained from Pun, particularly regarding the "question mark" feature, is quite consistent and coincide with the result obtained from the syntactic feature "SQ". That leads to make the same argument that we used to explain why the "SQ" tag negatively influence reputation scores.

6. Conclusion and Future work

We attempted to reveal and illustrate the relationship between users' reputation, and the syntactic and semantic representation of their associated CQA's content. We further analyzed the potential impacts of the selected linguistic features in the prediction of reputation scores from various perspectives. The methods presented in this study could be applied to further improve the quality of other important components (e.g., recommendation and ranking systems) related of the reputation system of CQA platforms. In the qualitative and quantitative analysis, more focus has been given for linguistic features, but as a future directions, it

would be also interesting to measure and provide detailed analysis of the impact of non-linguistic features.

7. Bibliographical References

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *HIPPOKRATIA*, Vol.14 Suppl 1:23–28.

Asaduzzaman, M., Mashiyat, A. S., Roy, C. K., and Schneider, K. A. (2013). Answering questions about unanswered questions of stack overflow. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 97–100.

Baltadzhieva, A. and Chrupala, G. (2015a). Predicting the quality of questions on stackoverflow. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 32–40.

Baltadzhieva, A. and Chrupala, G. (2015b). Question quality in community question answering forums: a survey. *SIGKDD Explorations*, 17:8–13.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Chen, H. and He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1741–1752.

Chen, Y., Wrenn, J. O., Xu, H., Spickard, A., Habermann, R., Powers, J. S., and Denny, J. C. (2014). Automated assessment of medical students' clinical exposures according to aamc geriatric competencies. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014:375–84.

Correa, D. and Sureka, A. (2014). Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 631–642.

Dascalu, M., Chioasca, E.-V., and Trausan-Matu, S. (2008). ASAP – an advanced system for assessing chat participants. In *AIMSA: International Conference on Artificial Intelligence: Methodology, Systems and Applications*, volume 5253 of *Lecture Notes in Computer Science*, pages 58–68. Springer.

Freedman, D. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Hu, Z., Zhang, Z., Yang, H., Chen, Q., and Zuo, D. (2016). A deep learning approach for predicting the quality of online health expert question-answering services. *Journal of biomedical informatics*, 71:241–253.

Jiwoon, J., Bruce, C. W., Ho, L. J., and Soyeon, P. (2006). A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 228–235, New York, NY, USA. ACM.

Klein, D. and Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Li, B., Jin, T., Lyu, M. R., King, I., and Mak, B. (2012). Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st International Conference on World Wide Web*, pages 775–782.

MacLeod, L. (2014). Reputation on stack exchange: Tag, you're it! In *Proceedings of the 28th International Conference on Advanced Information Networking and Applications Workshops*, pages 670–674.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., and Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 886–893.

Shah, C. and Pomerantz, J. (2010). Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 411–418.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., and Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, 24(24):171–176.

Suryanto, M. A., Lim, E. P., Sun, A., and Chiang, R. H. L. (2009). Quality-aware collaborative question answering: Methods and evaluation. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 142–151.

Woldemariam, Y., Björklund, H., and Bensch, S. (2017). Predicting user competence from linguistic data. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*.

Yan, X. and Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230.

Zhu, M., Zhang, Y., Chen, W., Zhang, M., and Zhu, J. (2013). Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443.

8. Language Resource References

Marie-Catherine Marneffe and Bill MacCartney and Christopher Manning. (2006). *Generating typed dependency parses from phrase structure parses*.

Joakim Nivre and Marie-Catherine de Marneffe and Filip Ginter and Yoav Goldberg and Jan Hajic and Christopher D. Manning and Ryan T. McDonald and Slav Petrov and Sampo Pyysalo and Natalia Silveira and Reut Tsarfaty and Daniel Zeman. (2016). *Universal Dependencies v1: A Multilingual Treebank Collection*.