Answers Quality – Predicting Best Answers in Yahoo! Answers

Nassi Abergel
Department of Management Information Systems
University of Haifa
Carmel Mountain, Haifa 31905, Israel
nassi9@gmail.com

Noffar Dahan
Department of Management Information Systems
University of Haifa
Carmel Mountain, Haifa 31905, Israel
noffardahan1@gmail.com

Abstract — Q&A sites provide a broad platform for asking and answering questions which vary by subjects, languages, users, education and demographic characteristics. Moreover, with no experts or quality assurance the content quality is naturally decreases. Q&A sites like Yahoo! Answers are offering a voting system which the askers and the participants can rank and choose the answers they prefer. Nevertheless, automatic answers quality evaluation can be very challenging because quality by definition is subjective and depends on the consumer satisfaction. In this paper an automatic approach is proposed, for distinguishing and predicting the best answers. Features are automatically extracted and analyzed, answers classification is made using machine learning techniques and finally we show that distinguishing can be made using these features. Best answers can be predicted in a good way using the automatically extracted features.

Keywords: *answer quality prediction, social Q&A, data mining, Yahoo! Answers.*

1. Introduction

Evaluating and predicting the quality of content is a subjective task. In online Community Question Answering (CQA) the answers and questions are mostly informal and individual by nature, since people seek for opinions or advices and not always for professional scientific answer. Moreover, in Yahoo! Answers there are no experts who review and fix these answers and many answers can be inappropriate, unhelpful or even harmful, and in many cases it could even be spam or advertisement. Thus evaluate the answer quality is very challenges and nontrivial task, a machine which automatically evaluates answer quality and accuracy is a very complex task, since it requires vast knowledge of the world, unlike human assessment which easily can evaluate content quality. In Yahoo! Answers there are millions of questions accompanied with dozens of answers, some of them are good answers and some are less. The best answer is selected by the questioners or by the participants, who have to decide what the best answer is. We want to examine what are the factors and criteria that influence the choice of the best answer and how a best answer can be predicted.

Yahoo! Answers has a system of points and levels, namely reputation system. Actions such as voting and answering increase user's points conversely actions which involve abuse or violation reduce points. The points help to progress in levels and help users to recognize how useful and beneficial other users answers are. Our motivation in the current study is the automatic approach that can be implemented in YA in a form of a tooltip that will indicate whether a suggested answer is chosen by Yahoo! as the best answer. In addition it can also display details of the features that automatically extracted from the answer. There is a high importance for automatic tools that evaluate the quality of user generated content, particularly in Web 2.0 sites, since most of them required their users to rate the content manually [9].

In order to understand better why and how users choose answers as best, we opened an account and we asked on Yahoo Answers[1]: "What makes you choose a particular answer as the best answer?"
The answers were (in short):

1) "I'm sure most people choose the answer that best helped to resolve their problem. If there are maybe 2 really good, basically the same, answers, May one provides a little something more that is helpful."
2) "An answer with more details, explanation, proof and convincing statements usually make me choose them. By the way, choose my answer."
3) "If a person gives me good insight on the topic I was talking about. Normally they tell me something I didn't know or give me a new perspective."

We can see just from these three answers that people define different criteria for best answers. There are three criteria that we can infer from these answers: Helpful, Informative and Novelty answers are a good candidate to be chosen as best.

1.2 Yahoo! Answers

Yahoo! Answers is a website based on questions and answers; people who want to use this website need to make their own user. This user gives them the opportunity to ask questions, see the answers people gave them and choose from them the best answer, make answers of their own for others questions, get points and more. Yahoo! Answers lunched in 2005 and it's available in 12 languages. In Yahoo! answer we can find millions of questions and a lot more of answers. Many studies have been conducted on Yahoo! Answers since the vast amount of the available data. There are many aspects that researches are exploring such as: predicting the category of a given question, deceptive answer prediction and answer quality prediction.

1.3 *The best answer in Yahoo! Answers*

When a user asks a question he may receive various answers from different users. One of the Yahoo! Answers mechanisms called "Best Answer" which is responsible for the best answers classification. When a question is asked, the option of best answer stays in open status for several days. The asker can choose the best answer or let the community vote for him. If the asker takes no action or a question doesn't receive any answers, the question expires and deleted; so best answer have to be chosen or else the question would be deleted. The asker can extend the period in order to let the community more time to answer. The asker also can designate if he satisfied with the answer he chose by writing comments. The best answer is not just a Boolean parameter which determine if the answer is best or not, it is also has rank from 1 to 5, in a form of stars. Once the best answer is chosen it can't be changed. Choosing the best answer is depending on how much the user or the community was satisfied from the answer. Although many answers can be of high quality only one chosen as best, but it does not guarantee that this best answer is necessarily has higher quality than others. In many cases high quality answers are provided by malicious users who mislead other users to choose their answers as best answers, or even select their own answers as best answers [4]. Thus, in this open and almost uncontrolled community, differentiate between levels of quality is much more complex than organized and structured community.

In Web 2.0 sites and especially in Yahoo! Answers where there are hundreds of categories, the content is produced by the users which usually share their personal experiences, opinions, advices and information, unlike other sites such as Wikipedia where there is no room for opinions or subjectivity. Hence, deduce unequivocal about criteria and features that distinguishing between the qualities of answers can be very difficult task.

[1] http://answers.yahoo.com/question/index?qid=20131204063121AAdD2fy

2. Literature Review

2.1. Quality in Community Question Answering

Community Question Answering (CQA) sites became very common and popular, there are many sites which are divided to categories and specific expertise, for instance HealthTap, Answerbag, Stack Exchange Network, Quora, Ask.com, WikiAnswers, etc[2]. These sites are also referred as knowledge exchange communities or social Q&A, where the quality control is applied with diverse and sophisticated mechanisms, which mostly based on the community itself. Harper et al [5] claims that in Q&A sites you get what you pay for, their study shows that in a fee-based Q&A sites, the answers quality usually higher than in the free sites. However, they also found that users in Q&A community such as Yahoo! Answers contribute to its success which outstripped individuals based Q&A sites, such as library reference services. Therefore, the field of predicting and evaluating the quality of answers attracted many researchers.

The recurrent issue in evaluation and prediction answers quality is rooted in the subjectivity criteria that indicate quality. An answer can include rich text and well formatted, including citations and references and yet it still can be incorrect at the same time.

Zhemin et al. claims that current NLP methods fail to explain quality in a scientifically approach [11]. We strongly agree with Zhemin, since quality can be defined in different ways and in many cases, it depends on the satisfaction of the consumer. Hence, many researchers used human assessment such as Amazon Mechanical Turk[3] to create a Human Intelligence Task (HIT) in order to evaluate the answers and define quality [2,8]. Other researches focused on the content analysis of the user's answers, they used criteria such as accuracy, politeness, originality, usefulness, objectivity, and completeness [11]. Some researchers focused on the spam and deceptive answers prediction [4] while others tried to evaluate quality via the network structure analysis [1,3].

Some researchers even tried to use a non-textual features to predict the answers quality [7], they used features such as print counts and click counts. Nonetheless, they also used the answer length as a feature, they claimed that this feature can be considered as textual feature, but they decided to add this feature anyway since it can be simply extracted without a deep analysis on the text and it is known to be helpful in assessing the quality of online texts.

Another study [10] on Yahoo! Answers tried to find the best answer selection criteria by analyzing the comments that askers write to their own questions when they choose the best answers. They developed a framework which based on the user-oriented relevance criteria to study the judgment of the askers. They used criteria such as content value, cognitive value, socio-emotional value and extrinsic value. In this paper we are taking different approach, we are extracting features that can be obtained automatically from the answers content and from the sources (references URLs) of the answers. We will try to automatically evaluate the quality of answers from the extracted features, with no human assessment. Finally we will try to predict the best answers using these features.

3.Research Design And Methods

In order to efficiently analyze the content of many answers, we want to hold a relative big DB of questions and answers. There are several approaches to access this data:

1) Yahoo! Research Webscope™ Program[4] makes few datasets accessible to university associated researchers.

2) Yahoo! Answers API[5] which allowing easily connect and get data.

3) Yahoo! Answers Datasets[6] which contains 2500 questions and four categories.

The third option didn't provide a filtered data according to our strict rules (see below filter rules section). Therefore, we developed a crawler in order to create

[2] http://en.wikipedia.org/wiki/List_of_question-and-answer_websites

[3] http://www.mturk.com

[4] http://webscope.sandbox.yahoo.com

[5] http://developer.yahoo.com/answers

[6] http://yahoodataset.sourceforge.net

our own dataset. The crawler used the Yahoo! Answers API (the second option) to collect questions and answers; we chose to use it from several reasons:

- Flexibility: the API provide data which cannot be fetched using static datasets.
- Testable: The API provides developers console which is easily can be monitoring.
- Relevance: in order to work with current data and updated questions and answers.
- Control: we control the data we want to work with- any category/users/questions/answers.

Filter Rules

Using the crawler we randomly downloaded questions and answers from the years 2013-2014, we didn't want to limit our study to only few categories, thus we produced a dataset which contains 313 categories and sub-categories, 17,509 questions and 72,293 answers. In order to create the dataset randomly, our crawler initialized with five categories numbers we have chosen arbitrarily from Yahoo! Answers site. For each category number we retrieved all the questions, for each question we retrieved all the answers and for each answerer we retrieved all the questions and so on until we have reached to the desire quantity of answers we wanted. This process was completely automatically. We filter non-English questions, questions with no best answer, and questions that have only one answer. The reason is that we think an answer need to have at least one more competing answer for a reasonable choice. This is ensures that the best answer wasn't chosen by default since it was the only answer. The data we collected is described in table 1.

Data	Amount
Categories	313
Questions	17,509
Answers	72,293
Best Answers	17,509
No Best Answers	54,784
Users	40,000
URLs	9,500

Table 1: Summary description of our dataset after filtering.

Zhemin et al. [11] suggested a model of 13 dimensions and 40 metrics for evaluating the quality of answers in social Q&A. they suggested dimensions such as politeness, completeness, readability, relevance, conciseness, level of detail, originality, and usefulness.
In order to develop this model they used user survey, experts and intuitions. Inspired by the proposed model and by the question we asked (see section 1) we developed the following six categories that we think describe quality in social Q&A sites regardless the category or question type, e.g. opinion, advice, information or expertise.
These categories are described in Table 2.

Category	Feature	Description
Informative	Tokens	Integer
	Reference Tokens	Integer
	Answer URLs	Integer
	Reference URLs	Integer
	Total URLs	Integer
	Page Rank	Integer
	Entities	Integer
Readability	Spelling Mistakes	Percentage
	Capitals Words	Integer
	Punctuations	Percentage
Convincing	Citations	Integer
Syntactic	Nouns	Percentage
	Verbs	Percentage
	Adjectives	Percentage
	Adverbs	Percentage
Relevance	Category Relevance	Percentage
	Answer Relevance	Percentage
	Page Relevance	Percentage
Reputation	User Questions	Integer
	User Answers	Integer
	User Best Answers	Integer
	Response Time	minutes

Table 2: A list of features we automatically extracted from the answers dataset.

Features

In this section we will explain about the automatically extracted features. As explained earlier, we used only features that can be extracted automatically from the questions and answers without any manual involvement. For each answer we extracted the features and we added a label with unique class of "yes" or "no" for indication whether it's the best answer or not.

Informative Features

Tokens: this feature is based on lexical analysis process. We developed a tokenizer which takes the answer string as an input and performs actions to remove stopwords, punctuations and any other unwanted data such as HTML and URLs. Many studies used the length of the content as a feature to predict quality [6,7,9]. In this study the length also being used as a feature, although we assume this feature alone cannot explains quality. Our tokenizer is based on TF-IDF which creates a unique BoW[7] for a document, in our case a unique BoW is created for each answer using unigram. We then count the total unique words in this BoW and assign it as the value of this feature.

Ref Tokens: when a user answers a question he can add a source (reference) to his answer in a specific source field, the source can be in form of URL, text or both. We took the source content and we applied the same process as the tokens feature.

The next three features are regarding to URLs that sometimes appear inside the answers, we claim that URLs can provide informative data to the asker and thus indicates for higher quality answers.

Answer URLs: this feature refers to the number of URLs that inside the content of the answer.

Ref URLs: this feature refers to the number of URLs that inside the source (reference) of the answer.

Total URLs: sum of the Answer URLs + Ref URLs.

Page Rank: this feature is based on Google Page Rank[8], PageRank is a metric for measuring the importance of page by counting its incoming links and their quality. For each answer we took all the URLs and we performed a web crawling to Google Page Rank API. The output of the API is the rank of the URL, represented by an integer number between 0-10. We crawled 9,500 URLs in total. If there was more than one URL for a given answer we considered the max page rank as the value of this feature.

Entities: this feature based on the information extraction (IE) field. We tried to identify entities using Named Entity Recognition (NER), during this task we were looking for entities such as organizations, locations, and persons in the content of the answers.

Readability Features

Spelling: we used Nhunspell[9] library in order to deal with spells mistakes. For each token in each answer we invoked the speller in order to count spells mistakes.

Capitals Words: we suggest that using capital words may infer as SHOUTING in the social media, and low readability of answers.

Punctuations: we count the number of punctuation marks in each answer and then we compute the percentage of punctuations, relative high number may difficult for readability of the answer.

Convincing Features

Citations: we counted the number of citation each answer has, simply by identifying the quotes pattern, e.g. I have to tell you that "some people are so poor, all they have is money".

Syntactic Features

The following four syntactic features are lexical categories, namely part of speech (POS) which extracted from the answers using WordNet. We count the percentage of: nouns, verbs, adjectives and adverbs in the extracted tokens (BoW).

[7] http://en.wikipedia.org/wiki/Bag-of-words_model

[8] http://en.wikipedia.org/wiki/PageRank

[9] http://nhunspell.sourceforge.net

Relevance Features

Page Relevance: besides the Google Page Rank that we used in order to measure the importance of the URLs. We claim that not only the rank of a page is important but also its relevance to the answer. Therefore, we performed a web crawling to the whole 9,500 URLs and we extracted from the pages three Metadata attributes[10] (Meta Tags) which are title, description and keywords. We combined the values of these attributes into single array of tokens.

Answer Relevance: the similarity measure between the answer and the question.

Category Relevance: the similarity measure between the category name and the answer.

In order to examine the relevance of the three features above, we used the Cosine Similarity measure.

For example: consider the following answer:

"What is the top videos music site?" and the answer is: "This music site is the best: [some link]".

The tokens of the question are: top, music, site, videos.

The tokens of the answer are: music, site, and best. Assuming the metadata of the link also contains the word video, we then measure the similarity and in this example the similarity is 86.6%, which is outstanding.

In the relevance features we used measure which based on lexical similarity rather on semantic similarity or sound like similarity such as Soundex. There are many others similarity methods, for instance: Levenstein Distance, Jaro–Winkler distance, Euclidean Distance, Smith Waterman and more. We chose to use Cosine Similarity since its being widely used in the text mining field, in particularly when the text is tokens based.

Reputation Features

For each answer in our data there is a user who provided this answer, and each user has a total questions he asked, and total answers he answered that have been chosen or not chosen as best.

User Questions: questions the user asked.

User Answers: answers the user answered.

User Best Answers: answers chosen as best answers.

Response Time: The time that took the answerer to reply, in minutes.

4. Experiments And Results

Features Analysis

In this section we will describe the results of each feature according to the best and regular answers.

Tokens

One of the most significant features that distinguish between best answers and regular answers is the tokens feature which belongs to the informative category. This is the answer length. The average tokens in regular answers is 15.7 while in the best answers is 29.5, which is almost twice.

URLs Analysis

In order to predict the best answer we used many features that involve URLs, e.g. URLs count inside the answer and inside the references, page rank and page relevance measure. We noticed that these features have minor influence on the prediction since only 7.5% of the answers (5448 answers) actually contain URLs. Nevertheless, the best answers yielded higher average values of URLs and page rank.

The results are as follow:

	Best	Regular
Answer URLs	0.17	0.06
Reference URLs	0.04	0.02
Total URLs	0.23	0.09
Page Rank	0.32	0.14

Readability Analysis

When we measured the spelling feature we noticed that the number of spelling mistakes was higher in the best answers, but we also encountered a strange situation (see in Difficulties section). The average of capitals words yielded higher values in the best answers- 0.41 compared to 0.21 in regular answers, this is simply can be explained by the fact that the best answers have much more tokens in average. The punctuations features produced relatively high average percent in both best and regular answers, both have 65% average punctuations, we assume the reason is that many users make an extensive usage of punctuations marks, e.g. "use this site….." Or "why are you asking??!!!!"

[10] http://en.wikipedia.org/wiki/Meta_element

Syntactic Analysis

In the part of speech (POS) features we didn't notice any significant difference between best answers and regular answers, the average percentage of the POS was almost the same. The results are as follow:

	Best	Regular
Nouns	51.7	49.8
Verbs	31.9	30.6
Adjectives	19.08	18.1
Adverbs	5.9	5.5

Relevance Analysis

The relevance features produced low average percent, although the relevance of the best answers was higher than the regular answers. The results are as follow:

	Best	Regular
Category Relevance	0.99	0.76
Answer Relevance	12.75	10.33
Page Relevance	1.19	0.54

The main problem with question answer relevance is that common methods such as Cosine Similarity and TF-IDF are based on lexical rather on semantic similarity. Fangtao et al [4] reported the same issue as well when they tried to measure question answer relevance, thus they developed a complex semantic models to overcome this problem.

Response Time Analysis

We have noticed an interesting feature that captures our attention, the average response time of the best answer was significantly lower than the answers that considered as not best answer. Our preliminary intuition for this finding is that users who are more experienced with high reputation are spending much more time on the site and therefore they are more available to answer. Moreover, these users should have more experience and therefore they have the ability to respond to questions more quickly. We assumed that this answers although their low response time was highly satisfactory for the askers.

After a deep analysis of the time feature, we observed that some questions were answered very long time after they were asked, approximately 7-8 years. This is a very long period of time after the best answer already been selected. We couldn't figure out why Yahoo! Answers allows adding answers even if there is already a best answer, particularly if there is only one best answer.

Obviously, these answers disrupted the "not best answers" average time statistics. As a result we decided to disregard answers that are written after the best answer has been chosen. We removed 180 answers from the dataset and we rerun the classifier again, now the results were much more reasonable. It seems that answers that chosen as best have high average response time, we suggest that high quality answers require more time to answer for several reasons:

1. The answer needs to be informative and detailed which require more time.
2. Finding relevant URLs involve the efforts of searching and browsing, we also can see that the "Total URLs" is actually higher in the best answers.
3. Organizing and formatting the answer also requires efforts and time.

If our assumption is true, we can say that best answers are requiring more time to write. An interesting question we could ask: is it possible that the "non-best answer" is usually sketchy or even copy/paste answers just to gain points in YA? Future research is required.

Convincing Analysis

The number of citations in the best answers was higher than the regular answers, 0.10 and 0.21 respectively**.**

It is possible that this feature affected by the length of the answer, and more deep analysis would reveal it.

Reputation Analysis

The average of user questions and user answers was lower in the best answers, but the average of user best answers was much higher in the best answers, twice from the regular answers.

The results are as follows:

	Best	Regular
User Questions	0.34	0.43
User Answers	8.55	9.11
User Best Answers	3.37	1.75

This finding led us to determine new heuristic that best answers will be chosen based on the user best answers. See the experiment at the next section.

Predicting the best answer

We used Weka for the machine learning tasks, using the following classification algorithms:

- Decision Trees: J.48 and Random Forest.
- SVM: Support Vector Machines (Weka SMO).
- Naive Bayes: Probabilistic Classifier.
- AdaBoost: Adaptive Boosting.
- Regression: Logistic & Simple Logistic.

Random classification: for the baseline test we used a random classification for choosing the best answer. The results were precision of 0.5 and recall of 0.23.We were able to overcome the random test at every classification algorithm. For the Naive Bayes algorithm we used an important feature of Weka which is discretization, grouping the feature values into new well-defined sets. Studies revealed that algorithms such as Naive Bayes works better with discretization. For the experiments we conducted 10-fold cross validation to all the classification algorithms for the heuristic and non-heuristic, we normalized the features vectors before the classification tasks. The results are shown in Table 3.

	Per	Rec	F-Meas	Accu
Random	0.50	0.23	-----	-----
J.48	**0.72**	**0.59**	**0.65**	**84.5%**
Random Forest	0.75	0.52	0.62	84.4%
SVM	0.60	0.14	0.22	76.9%
Naive Bayes	0.52	0.60	0.56	77.3%
Simple Logistic	0.61	0.22	0.33	77.8%
AdaBoost	0.83	0.40	0.54	83.6%
Logistic	0.61	0.23	0.33	77.9%

Table 3: experimental results of classification algorithms for best answer prediction.

The decision trees classification algorithms produced the highest results. J.48 with 0.72 precision and 0.59 recall and total accuracy of 84.5%. The Random Forest with 0.75 precision and 0.52 recall and total accuracy of 84.4%. These results are consider as good according to other researchers [2] who achieved 81.7% accuracy in predicting the best answer in Yahoo! Answers using different features and different dataset. Furthermore, they didn't mention the results of the precision or the recall. Other researchers [1] achieved 62% prediction accuracy based on the answers length alone. We also used Weka "Information Gain Ranking Filter" in order to rank the attributes (features) that have the most impact on the best answer prediction.

The results are as follows:

Rank	*Feature*
0.2260	User Best Answers
0.0524	Tokens
0.0310	Adjectives
0.0304	Adverbs
0.0301	Nouns
0.0274	Verbs
0.0243	Punctuations
0.0163	Answer Relevance
0.0149	Entities
0.0077	Capitals Words
0.0074	Total URLs
0.0067	Response Time
0.0060	Answers URLs
0.0057	Page Relevance
0.0053	Ref Tokens
0.0049	Page Rank
0.0046	Citations
0.0041	Category Relevance
0.0016	Reference URLs
0.0007	User Answers
0.0005	User Questions

Table 4: affecting features ranking

This ranking table helped us to understand better the classification task and what are the most affecting features. We have noticed that all the three reputation features are particularly intriguing, at the top of the table "User Best Answers" feature has the highest impact on the classification and at the bottom of the table the "User Answers" and the "User Questions" features have the lowest impact.

Heuristic classification

We noticed that the most affecting feature is the ***user's best answers***, hence we decided to test how it will affect the results if we will try to predict the best answer just by relying on the user who have the maximum number of ***best answers,*** rather on the actual best answer.

The results of the heuristic classification are shown in Table 5. The heuristic classification shows that the user's best answers is still the top affecting feature and even ranked with higher value, the tokens are also at the top three (Table 6). However, it seems that the ***user answers*** feature have high impact only in our heuristic, this can be explained by the fact that we chose the best answer by relying only on the maximum number of ***best answers*** the user have, and in Yahoo! Answers it goes hand in hand with the ***user answers***. The user best answers are derived from the user answers, or to be exact this is part of the user reputation.

	Regular classification				Heuristic classification			
	Per	Rec	F-Measure	Accuracy	Per	Rec	F-Meas*ure*	Accuracy
J.48	**0.72**	**0.59**	**0.65**	**84.5%**	0.65	0.54	0.59	84.5%
Random Forest	0.75	0.52	0.62	84.4%	0.65	0.46	0.54	81.1%
SVM	0.60	0.14	0.22	76.9%	0.73	0.39	0.51	81.8%
Naive Bayes	0.52	0.60	0.56	77.3%	0.56	0.44	0.33	74.6%
Simple Logistic	0.61	0.22	0.33	77.8%	0.72	0.42	0.53	82.1%
AdaBoost	0.83	0.40	0.54	83.6%	0.65	0.50	0.57	81.6%
Logistic	0.61	0.23	0.33	77.9%	0.72	0.42	0.53	82.1%

Table 5: Regular classification vs. Heuristic classification for best answer prediction.

Regular classification	
Rank	Feature
Top features	
0.2260	User Best Answers
0.0524	Tokens
0.0310	Adjectives
0.0304	Adverbs
0.0301	Nouns
Bottom features	
0.0046	Citations
0.0041	Category Relevance
0.0016	Reference URLs
0.0007	User Answers
0.0005	User Questions

Heuristic classification	
Rank	Feature
Top features	
0.2872	User Best Answers
0.0985	User Answers
0.0122	Tokens
0.0064	Adverbs
0.0057	Nouns
Bottom features	
0.0011	Category Relevance
0.0011	User Questions
0.0010	Ref Tokens
0.0008	Response Time
0.0001	Reference URLs

Table 6: Regular ranked features vs. Heuristic ranked features for the best answer prediction.

Different data sizes

Considering the results of our experiment, we decided to test the approach on different sizes of data. We divided the data into two parts, the first part include 24,000 answers and the second part 48,000 answers. The consideration for these sizes of data stems from the equal distribution of 72,000 answers (72,000 / 3). Eventually we ran the experiment on three sizes: 24,000/48,000/72,000. The results are shown in Table 7. We observed that the results of the two sizes do not differ much from each other and from the original data, regarding the precision, recall, f-measure and accuracy.

We also used the "Information Gain Ranking Filter" to check whether the impact of the features has been changed, it seems that most of the features remain the same except that in the 24,000 answers dataset the "User Best Answer" feature resulted higher (0.26) rank from the two other datasets (0.22). Which could be explained by the fact that we did a random data segmentation and probably this action produced more "coherent" dataset. From the results of this experiment it is also slightly noticeable that the values of the small dataset are higher than the two other datasets.

	24,000 Answers				48,000 Answers			
	Per	Rec	F-Measure	Accuracy	Per	Rec	F- Measure	Accuracy
J.48	0.74	0.60	0.67	85.5%	0.72	0.58	0.64	84.8%
Random Forest	0.76	0.55	0.64	85.1%	0.75	0.51	0.61	84.5%
SVM	0.57	0.11	0.18	76.5%	0.59	0.11	0.19	77.0%
Naive Bayes	0.54	0.59	0.57	78.3%	0.52	0.60	0.56	77.0%
Simple Logistic	0.60	0.21	0.31	77.6%	0.60	0.21	0.31	77.8%
AdaBoost	0.85	0.45	0.59	85.0%	0.83	0.41	0.55	84.0%
Logistic	0.60	0.21	0.32	77.7%	0.60	0.21	0.32	77.9%

Table 7: experimental results of classification algorithms for different data sizes.

5. Discussion

We conclude directly from the results that there are two significant features that distinguishing between best and not best answers. The first feature is the user best answers from the reputation category; the second is the tokens which belong to the informative category. This finding may suggest that if certain user with high reputation writes an informative answer, his answer probably will be chosen as the best answer. However, we also conclude that the users answers and questions were at least significant, which means it doesn’t matter how many questions the user asked or how many answers he answered, its matter how many users chose his answers as best. We also noticed that in the best answers almost every feature has relative higher value than the not best answers.

We could not explain whether each high value of the feature is necessarily good, but we definitely detected patterns in the best and not best answers. We assume that our research results were relatively good according to other researches. Chirag et al [2] have used human assessment which achieved classification accuracy of 79.50%, with dataset of 600 answers. When they used automatically extracted features with dataset of 5032 answers they achieved accuracy of 84.52% which is much better. Our relatively good results could be explained by several reasons: 1) we used large and diverse dataset. 2) We used a variety of features with emphasis on user's reputation and the answer content. 3) We used various classification algorithms for the machine learning tasks.

6. Limitations

We have found several limitations in our study. There are many attributes that are missing in our data, since Yahoo! Answers keep the users privacy, attributes such as age, education and reputation are missing and we are unable to measure their impact on the evaluation of answers quality. Moreover, these attributes are very dynamic which means that if we have questions from the last five years, we can't rely on the current attributes of the users. The age, education and reputation of the users has been changed at the last five years, and if we wanted to perform an exact measurement we should have access to these attributes at any given time, when the questions have been asked.

Difficulties

During the spell checking, we found several answers that written in other languages but English; although the data we requested from Yahoo! Answers was English only. We had to deal with language detection to remove these questions since they were irrelevant and created misleading data (high number of spelling mistakes). Sometimes users used the Spanish language but with English letters, or Spanish letters mixed with English answers, so we couldn’t remove these answers, for example:

Q: "Help me pick a meaningful screen name!?"
A1: BigBrownEyesHazelirisBlueEyedBeuty
A2: Åññâ

To deal with that problem we converted the non-English words into English and then we counted the total spell mistakes, if the number was relative high we examined the answers and if they were in other languages we manually removed these answers from our dataset. For example:

http://answers.yahoo.com/question/?qid=20081109033716AAX0TbM
http://answers.yahoo.com/question/?qid=20090820100943AAJV0PO
http://answers.yahoo.com/question/index?qid=20090629094757AALnBS8

During the spell checking procedure we also encountered with answers that contained lot of terms that the speller recognized as spell mistakes, especially in technical categories such as computers. Perhaps we should develop a unique speller for Q&A sites.

7. Conclusions and Future Work

In this paper we tried to automatically extract features and predict the best answers, we created categories that may indicate for quality in social Q&A sites, categories such as informative, readability, relevance and reputation. We show that two features have high impact on the best answers: the answers with high value of tokens and that are written by users with high reputation. Furthermore, we successfully predicted the best answers just by these two features and we also managed to predict the best answer with three different data sizes. Additionally we noticed that in the best answers almost every feature has a relative high value than the not best answer. In this research we tried to use various text mining methods, such as cosine similarity, TF-IDF, POS, BoW etc. and different machine learning algorithms such as Naive Bayes and SVM according to the educational material learned in class. There is a need to say that those methods helped us to analyze the data easily and to conclude the results.

In the future work, it will be interesting to test our approach with another social Q&A sites and compare it to Yahoo! Answers, and also to test our approach with different languages, although we will have to consider the use of specific speller and unique stopwords for each language. It will also be interesting to add semantic features in addition to the syntactic features we used. Using semantic similarity can be extremely challenging since it requires a unique ontology or taxonomy of the social Q&A world. In addition, we suggest analyzing the writing style of the answers, e.g. word usage, passive and active, conjunctions, prepositions, etc. Another two directions we suggest is to use sentiment analysis to identify positive and negative answers, and second to use subjectivity analysis to classify subjective and objective answers. These two sentiment analysis directions can provide assistance in predicting the best answers.

8. REFERENCES

[1] Adamic, L.A., Zhang, J., Bakshy, E., and Ackerman, M. (2008). "Knowledge sharing and Yahoo Answers: Everyone knows something", Proc. World Wide Web Conference, WWW2008.org.

[2] Chirag S. and Jefferey P. (2010) "Evaluating and Predicting Answer Quality in Community QA." *SIGIR'10*, July 19–23, 2010, Geneva, Switzerland.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. "Finding High-Quality Content in Social Media". WDSM'08, 2008.

[4] Fangtao Li, Yang Gao, George Zhou, Xiance Si and Decheng Dai, "Deceptive Answer Prediction with User Preference Graph". The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013).

[5] Harper, M. F., Raban, D. R., Rafaeli, S., & Konstan, J. K. (2008). "Predictors of answer quality in online Q&A sites". In Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (pp. 865−874). New York: ACM.

[6] J. E. Blumenstock, Size matters: "Word count as a measure of quality on Wikipedia", in Proceedings of the 17th international conference on World Wide Web, pp. 1095–1096, 2008.

[7] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answer with non-textual features", in Proceedings of the 29th annual international Special Interest Group on Information Retrieval, pp. 228–235, 2006.

[8] Liu, Y., Bian, J., & Agichtein, E. (2008). "Predicting Information Seeker Satisfaction in Community Question Answering." Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval.

[9] M. Weimer and I. Gurevych, "Predicting the perceived quality of web forum posts", in Proceedings of the Conference on Recent Advances in Natural Language Processing, pp. 643–648, 2007.

[10] Soojung K. and Sanghee O. (2008) "Users' Relevance Criteria for Evaluating Answers in a Social Q&A Site." journal of the american society for information science and technology, 60(4):716–727, 2009.

[11] Zhemin Z. and Delphine B. and Iryna G. (2009) "A multi-dimensional model for assessing the quality of answers in social Q&A sites." Technical Report TUD-CS-2009-0158, October, 2009.