SYS 6018 Final Project Report: Flagging Collapsed Answers on the Community Forum Quora

Group Members:
Shannon Mitchell, Mason Montgomery, Akshat Verma, Mike Voltmer
10 Dec 2015

Problem Description

- What is the problem?

Online question forums like Quora utilize user votes to determine the applicability of provided answers. In addition, human moderators flag inappropriate answers for reasons such as answers being not sincerely responding to the question, very short, difficult to read and understand, bad grammar, spacing, punctuation, spelling, unsupported personal opinions and assertions that provide minimal explanation among other things[1][2]. When a new answer is provided, it takes users' and/or moderators' time to determine whether it is applicable and appropriate. If a user has to spend a significant amount of time sifting through insufficient answers, it may affect the user experience and reduce the value they place on Quora's site and service. Automated categorization of answers could be helpful to reduce user time of reading incomplete, irrelevant or offensive answers. In addition, it could provide feedback to respondents if their proposed response is likely to fall into one of those categories.

- Why is it important?

The problem is important to Quora because the company wants to gain and retain users, and if the service is too cumbersome or time intensive, the users might not get their desired experience and decide not to return. It is important to users, as they want to minimize the time spent to find the answer to their question. Insufficient answers reduce the utility of the Quora service.

- Who cares about it?

This problem is important to users as well as the companies that run the online forums. From a company's standpoint, the number of employee moderators could be reduced as less-expensive automated pre-screening could take the place of human screeners. The displaced employees could then be deployed to other tasks. From the users' standpoint, the utility of the site degrades if it is too cumbersome to find relevant information amount insufficient answers. If the user experience is poor, they may not return to the site.

- Why does it remain unsolved?

Quora mainly relies on users and moderators for the laborious process of collapsing non-applicable answers. As of June 2014 they were testing a content review system for some questions, however users are not satisfied with the implementation[3].Models to assess content quality have been applied to technical forums such as Stack Overflow, so we investigated if similar results could be obtained for general forums such as Quora.

Objectives

- What are you proposing to do about the problem?

We proposed to create a predictive model capable of identifying unhelpful/low quality answers that should be collapsed in order to save the time of both users and moderators. This could potentially be used by Quora content review team to automate content moderation. We had a second objective to predict percentage of upvotes an answer is expected to receive with respect to total views, which could be used to provide immediate feedback to the answerer or to quantify the applicability of an answer in advance. However, after data collection we realized that collapsed answers did not contain number of views, so this analysis could not be conducted with information available.

- How will you measure the success of your work?

We selected precision and recall values as measures of success for classifying answers that should be collapsed. For this preliminary model, precision was more important than recall in order to provide answer owners with the benefit of the doubt. We would rather miss answers that should have been collapsed than flag an answer as insufficient that in reality could be useful to the Quora community. Also, recall was a better measure than overall accuracy owing to the inherent data skew against proportion of collapsed answers.

The criteria for success of this study was defined as precision higher than the rate of 66% published by Correa et al.[5] for predicting deleted answers in Stack Overflow. This prior study was selected as the baseline because it was also an early attempt at classification on an online question forum.

Related Work

- What have others (e.g., researchers, companies, etc.) done to address this problem?

Moderation in forums like Stack Overflow or Yahoo! Answers is typically done manually by human moderators. There have been efforts to automate this process to some extent. For instance, Yahoo! uses a rule-based engine to flag offensive content in the answers[7].However, an automated moderation to assess appropriateness and relevance of an answer to the question asked does not appear to be implemented yet, though some discussion is found in research. Stack Overflow identifies low quality posts manually. They have a system that flags posts to be added to the review queue, but that review is still done by a human[4].

- What are you doing that is similar to past work?

Like some of the prior research[4][5], we looked at text content of questions and answers, answerer data such as number of questions answered, and answer data such as uncollapsed/collapsed status, percentage of upvotes, timestamp, number of views, and tags.

- Are there commercial products that accomplish what you are trying to do? What are their characteristics? Where are their gaps?

There did not seem to be any commercial products addressing this problem, as each website will have its own unique challenges, and researchers have only begun tackling the problem in the past couple of years. Most forums seem to be handling this question internally for now.

- What about your work is novel? What gaps does it fill?

There does not yet seem to be any research applying to Quora the methods used on other forums. The closest thing so far seems to be, appropriately, a Quora question, but only general advice is given with no quantitative measures[6]. Most other research has focused on technical forums like Stack Overflow, or technical topics in general forums. Whereas the previously mentioned studies have looked at offensive answers on Yahoo! Answers and low quality answers on Stack Overflow, we look to combine those reasons for collapse into a single model for Quora.

Approach

- What data did you use? How did you preprocess it?

The top 50 Quora topics for 2015 are summarized in the answer to the Quora question “What are the most followed topics on Quora in 2015?” For the first page of each of the 50 topics, the top questions, roughly 15-20 for each topic were scraped using Python scripts incorporating the Python packages Urllib and BeautifulSoup. As quora web pages support dynamic content generation on scrolling and require user interaction to access the complete content, Selenium was incorporated to automate logging in to Quora, scrolling through all the answers for each collected question and clicking on 'Show collapsed Answers' link in order to scrape the desired text. Only those questions and all of their associated answers were included in the dataset which had any number of collapsed answers. This resulted in the collection of 4635 answers to 183 questions in 27 question categories for this analysis. In addition to the question and answers pages, the answerers' profile pages were scraped as well adding more initial features to the analysis.

For each answer, the following features were collected:

- Answer class: accepted or collapsed
- Question topic
- Question text
- Answer text
- Number of upvotes
- Number of views
- Number of questions asked by answerer
- Number of answers written by answerer
- Number of posts written by answerer
- Number of followers of answerer

The data was cleaned to reformat variable types and address missing and invalid values. The response variable was answer class. Therefore, if an answer did not have have a valid class, the observation was removed. Similarly, any observation that contained an uppercase alphabet characters for any of the "Number of" variables were removed. Empty answers were removed as their content could not be evaluated. Lastly, integers in the "Number of" variables were converted from strings to numeric by removing commas and multiplying numbers ending in "k" by 1000.

- What analyses did you perform?

Prior to building the prediction models, additional features were engineered based upon features that have been cited in literature as useful in predicting quality of answers in other community answer forums[4][5]. The following features were engineered:

- Number of characters in answer: includes spaces and punctuation, calculated using the stringr package in R, boundary set to "character"
- Number of words in answer: calculated using the stringr package in R, boundary set to "word"
- Number of sentences in answer: calculated using the stringr package in R, boundary set to "sentence"
- Number of spaces in answer

- Percentage of uppercase letters: $\frac{number\ of\ uppercase\ letters}{number\ of\ total\ letters}$
- Percentage of lowercase letters: $\frac{number\ of\ lowercase\ letters}{number\ of\ total\ letters}$
- Length ratio: $\frac{number\ of\ characters\ in\ question\ text}{number\ of\ characters\ in\ answer\ text}$
- Question words: type of question according to the presence of the following words in the question text - who, what, when, where, why, how, which
- Coleman Liau readability index: $0.0588L - 0.296S - 15.8$

    L is the average number of letters per 100 words
    S is the average number of sentences per 100 words
    Calculated using the koRpus package in R
- Gunning fog readability index: $0.4[(\frac{words}{sentences}) + 100(\frac{complex\ words}{words})]$, calculated using the koRpus package in R
- Shannon entropy: measurement of the randomness of the words in the answer. Calculated using the entropy package in R.
- Percentage of text abbreviations in answer: $\frac{number\ of\ text\ abbreviations}{number\ of\ words}$

    List of text abbreviations was scraped from www.netlingo.com/acronyms.php
    Abbreviations including multiple words and punctuation only were removed
- Percentage of stop words in answer: $\frac{number\ of\ stop\ words}{number\ of\ words}$

    The english list of stopwords from the tm package in R was used
- Semantic similarity index: measurement of semantic similarity between question and answer text using latent semantic analysis[8], calculated via an API from University of Maryland, Baltimore County Semantic Textual Similarity Service http://swoogle.umbc.edu/StsService/

From feature importances given by Random Forest and cross validation, seven factors were able to be removed:

- Percentage of lowercase and uppercase letters
- Percentage of slang words
- Questions by answerer
- Number of sentences
- Posts by answerer
- Type of question

The analysis performed was to classify whether an answer should be collapsed or not. Two types of factors were included in the analysis, one type we called intrinsic factors and the other extrinsic factors. The intrinsic factors included features engineered from answer content and comparison of answer content to the question content. The extrinsic factors included the profile of answerer, percentage of upvotes and views and question topic. Techniques from natural language processing like semantic analysis were used to construct our features.

- What models did you build?

We tested multiple models that tend to perform well on binary classification problems, including Radial Support Vector Machines and Logistic Regression. However, we found that a Random Forest provided the best results. Random Forests consist of many decision trees that are constructed on subsets of predictors and observations in the data. In doing so, the model does not overfit when the trees are aggregated together in the final model.

- What evaluation setup will you use?

We chose precision as our primary method of evaluation while maintaining an acceptable recall, because we want to hone in on our ability to classify collapsed answers. We care less about how the classifier does at predicting accepted answers. Precision is the ratio of true positive observations to all predicted true observations. Maximizing this score minimizes the number of Quora answers that are incorrectly predicted to be collapsed. We also evaluated recall, the proportion of collapsed answers that were predicted accurately. Using precision and recall as metrics are more specific than accuracy, as the tradeoff between false positives and false positives can be assessed.

Evaluation

- How well does your approach perform according to the metrics you describe in the Objectives section?

Our approach was successful and model performance is summarised in the table below.

Models after variable selection	Precision	Recall
Logistic Regression	74.9%	56.4%
Support Vector Machine	75.8%	60.5%
Random Forest with threshold 0.5	77.5%	79.0%
Random Forest with threshold 0.6	83.8%	70.0%

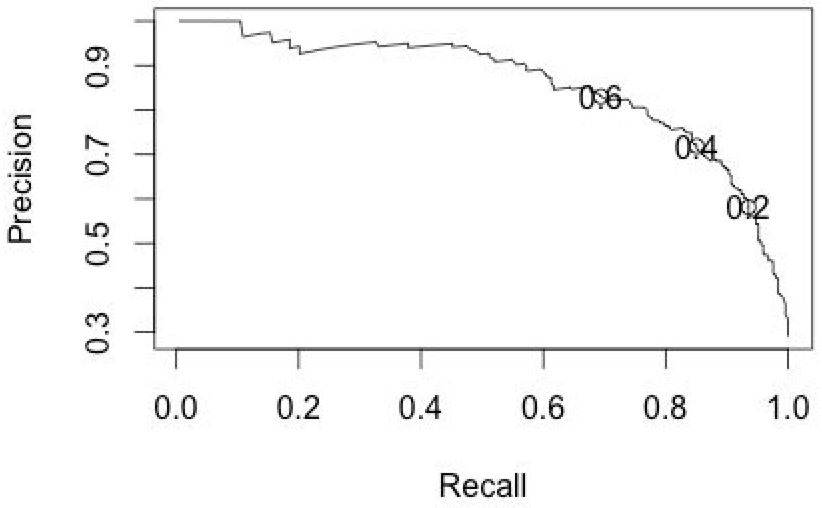
The first model we tried was a logistic regression model. The only statistically significant factors were entropy and percent stopwords at 0.05 level of significance. The model surpassed the precision target of 66% with a result of 74.9%, but recall was poor at 54.8%.

The next model we tried was a Radial Support Vector Machine (SVM). The Radial SVM slightly improved the precision over the logistic model to 75.8%, and recall also improved to 60.5%.

The final classification model we applied was a Random Forest model. This model again improved precision performance to 77.5%. There was a large increase in recall performance to 79.0%, which made the Random Forest the top overall model.

In order to account for the priority of precision above recall, the threshold was tuned. A precision/recall optimization with threshold values between 0 and 1 was conducted, and 0.6 was selected to maximize precision while maintaining an acceptable recall. The final precision value was 83.8% precision and recall was 70%. This threshold optimization could be performed by Quora depending on which type of error is more costly.

Collapsed Answer Precision vs. Recall
Random Forest Threshold Adjustment



- In what situations does your approach perform well? Provide examples.

Our final model does well at predicting collapsed answers within the sample dataset, given the precision and recall scores. It is likely the performance can be improved given a larger sample and more features. Our model also does a good job at classifying accepted answers. Given that the dataset was popular topics, we would expect the model to perform well on other topics with sufficient popularity however that assumption would need to be tested. Given the resources available, we were able to answer our question and pave the way for further research. With sound methodology, we maximized the prediction ability of our model by tuning parameters such as our response threshold. If inference is of interest to Quora instead of, or in addition to prediction, our logistic regression model showed promising precision performance and would be a good starting point for further work.

- Where does your approach break down? Provide examples.

The underlying data used to create this model was only a subset of the more than 6 million Quora question and associated answers. The data used was limited to the most popular Quora topics, therefore the number of answers, views and upvotes is likely higher than the average question on Quora. In addition, the response variable was collapsed versus not collapsed, therefore questions without collapsed answers were not included. Our analysis only included a small percentage of all the possible questions and answers and therefore cannot be assumed to apply to the general population of answers until it is tested against a representative sample. Also, since the best performance was achieved with a model which is not very interpretable, if Quora wants to understand the importance of the relevant factors in a collapsed answer better, they would need to utilize the logistic model and optimize it further.

- How does your approach stack up against other known approaches? Direct comparisons on shared test sets are best.

Our best approach performs with 83% precision and 70% recall, compared with our 66% goal. That goal was derived from the Correa et. al. paper[5], which achieved 66% precision and recall for a first attempt at classifying Stack Overflow questions. The later Ponzanelli paper[4] also analyzed Stack Overflow and broke down answers into quality categories, but its most comparable result was 68.9% precision. These are not direct comparisons since Stack Overflow is a highly technical forum, which attracts a different type of user than a general forum like Quora. A more apt comparison would be to Yahoo Answers, which was studied by Topa et. al. Their most comparable models performed at 78% precision in predicting low quality answers[9], which is similar to our results. However, Yahoo Answers does not have as much of a vetting process for quality Subject Matter Experts (SME), and less often deals with technical questions. Since there are technical questions and a SME system on Quora, there is a reasonable argument to be made that Quora is somewhat more serious than Yahoo Answers, and should be seen as standing somewhere between Stack Overflow and Yahoo Answers. Having similar precision on a more technical site can be considered an accomplishment.

Conclusions and Recommendations

- What have you learned by doing this work?

This initial work shows that prediction of answers that should be collapsed is feasible, and could be implemented in order to save both user and Quora moderator time. We learned that it is important to compare various model types, and tune the response threshold in order to optimize the tradeoff between precision and recall.

- What are your final recommendations with regard to addressing the problem you have identified?

Based upon the precision of 83.8% obtained using the random forest model, we believe this could be incorporated into the Quora review process to assist in identifying answers that should potentially be collapsed. Prior to implementing an auto-collapse feature using a model such as this, flagged answers should be queues for review by a moderator to ensure the flagging was appropriate. If the model achieves acceptable precision during a testing period, then auto-collapse could be implemented.

- What should be done to better address this problem in the future?

There are limitations to this study that could be addressed with future work. The sample questions and answers used to build the prediction model were from the top 50 topics on Quora, which likely have more views and answers than other Quora questions and answers. Therefore, the model should be tested on less popular topics to see if the performance is similar or if the model needs to be adapted. Additional factors such as part of speech counts and syntactic analysis could also be added to the model to determine if they improve the performance.

References

1. https://www.quora.com/Why-are-answers-and-reviews-on-Quora-collapsed
2. https://www.quora.com/Quora-Policies-and-Guidelines/What-kinds-of-answers-on-Quora-are-not-helpful
3. https://www.quora.com/How-does-Quora-Content-Review-work
4. Andrea Mocci, Luca Ponzanelli, Alberto Bacchelli, Michele Lanza and David Fullerton,Improving Low Quality Stackoverflow Post Detection, 2014
5. D. Correa and A. Sureka. Chaff from the Wheat : Characterization and Modeling of Deleted Questions on Stack Overflow. In Proceedings of WWW 2014 (23rd international conference on World Wide Web. ACM, 2014
6. https://www.quora.com/What-factors-increase-an-answers-chances-of-receiving-upvotes
7. http://buildingreputation.com/doku.php?id=chapter_10
8. Lushan Han, Abhay Kashyap and Tim Finin, Semantic Textual Similarity Systems, Proceedings of the Second Joint Conference on Lexical and Computational Semantics, 2013
9. Toba, Hapnes, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. "Discovering High Quality Answers in Community Question Answering Archives Using a Hierarchy of Classifiers." *Information Sciences* 261 (2014): 101-15. Web.