**Improving Crowdsourced Documentation:**

**Examining Answers on Stack Overflow**

Scott Crain, Niranjana Gajjelli, and Marc Hudak

Abstract

The goal of our project is to divorce questions from answers and focus on what makes an answer desirable. Our hypothesis is that highly voted answers reflect what people like to see in answers. To test this, we compare accepted answers against different tiers of top rated answers in order to determine what differences, if any, exist between top rated and accepted answers. We compare the answers using a set of metrics pulled from the stack overflow 2014 data dump.

1. **Introduction**

As an unofficial source of documentation, Q&A sites such as Stack Overflow are immensely popular among programmers, and that popularity is only increasing as time passes. As of October 19, 2014, Stack Overflow has over 8.2 million questions and 14 million answers. A mere year ago, Stack Overflow had over 5.5 million questions and about 10.5 million answers. Both of these are nearly 50% increases over a single year. Stack Overflow is constantly growing, and shows no signs of slowing down. However, as with all forms of community-driven content, the quality of the material on Stack Overflow is subject to questioning. With the increasing popularity of Stack Overflow as not only a Q&A forum, but as a form of unofficial documentation to those who come after, low-quality content is a very bad thing to have.

The primary forms of quality control on Stack Overflow are the "voting" and "accepted answer" systems. For any given post, be it question or answer, any Stack Overflow user can vote the post either up or down, and these votes are aggregated into a "score" on the answer. The simplest and most obvious use of this is that posts that score higher are more visible. Users are encouraged to participate and vote through the Reputation system. Therefore, generally speaking, a highly-voted post will be a high-quality, useful post as judged by the Stack Overflow community while a low (or negative) scoring post has been judged to be of little or no value. On the other hand, the accepted answer system is only available to question askers. The asker of a question can choose to mark any answer to his question as the "accepted answer", causing it to appear at or near the top of the answer list for his question, along with the highly-voted posts. Practically, this is very similar to the voting system, though there is the key difference that the only person who determines the accepted answer is the question asker, instead of the whole community.

1. **Problem Definition and Research Questions**

Our problem focuses specifically on the answers that have been deemed to be high quality. Specifically, we will be examining answers with a score of at least 15 and the set of accepted answers. These answers will be evaluated against a set of metrics to answer the question of "What makes for a good answer on Stack Overflow?”. The accepted answers and the highly-voted answers will, nominally, comprise two different data sets, but both are being measured in the same way. To this end, our primary goal is to answer the following question: “How do the factors that contribute to Accepted Answers differ from those that contribute to Highly Voted Answers?”

1. **Data Collection**

Our first step with this project was to convert the Stack Overflow data dump to a queryable Microsoft SQL Server database. To accomplish this, we used the “Stack Overflow Data Dump Importer” (SODDI), freely available from github user peschkaj. Using SQL Server Management Studio, we found that there were 13,684,117 posts marked as Answers, and these posts were the foundation of our data sets. We found the average (mean) score of all answers on Stack Overflow to be 2 and the standard deviation of scores to be 12.77. Even though the scores aren’t distributed normally (There were 239,445 answers with a score at least 1 standard deviation above the mean, which is roughly 1.7% of all answers), we felt that this cutoff provided enough answers to be meaningful. We also determined that there were 4,596,596 Accepted Answers in the dump by cross-referencing Post IDs and Accepted Answer IDs.

From there, we developed a small program that pulls each answer under consideration from the database and computes the metric set we’re using to examine each post for that answer and export the data to a comma-separated value file for easy examination. The metrics under consideration are “Noun count”, “Verb Count”, “Adjective Count”, “Total Word Count”, “Link Presence”, “Code Snippet Presence” and “Response Time”. Part of speech counting was accomplished by using the Stanford Part of Speech Tagger. Links were checked by examining the answer for an “<a href” bit in the answer, which would denote a hyperlink. Code snippets were similarly checked for by checking for “<code>”, because that HTML tag is used in Stack Overflow posts to denote a code block. Response time was computed by comparing the post time of the question to the post time of the answer.

1. **Research Methodology and Results**

We used a machine learning classification technique known as Alternating Decision Tree(ADT) to find an optimal model for parsing our results. We tested a few classification techniques (Decision Stump, J48, Simple Logistic, Bayesian Regression), but ADT offered a few distinct advantages. First, it offered a tree-based graphical representation of the model. Secondly, it was able to process a larger dataset and ran much quicker than other tree-based approaches (a difference of about 15 minutes for our smallest dataset). Importantly, this speed and size increase had no discernable affect on model accuracy (typically less than 1%). For analysis, we used cross-validation with 10 folds. Testing the results of cross-validation versus percentage split offered little difference in terms of stated model accuracy.

Our first task was to look at how well we could predict whether an answer would be accepted or top using all of our metrics. What we found was that as the score threshold for answers decreased, the difference between top answers and accepted answers decreased. We felt that accepted answers and top rated answers diverge as score increases is significant – why would a question asker not agree with the community? - and therefore focused our efforts on the highest tier of top answers versus accepted answers. In this tier, there were 57,156 top scoring answers. Because our goal was to determine which factors can be used to differentiate accepted answers from top answers, we selected 57,156 accepted answers to match our top scoring answers and provide meaningful classification percentage results. Because there is no obvious way to narrow the field of accepted answers, the sample of 57,156 were selected randomly from the accepted answer population.

Additionally, although we looked at 7 metrics initially, we found that 4 were strongly correlated and therefore narrowed these fields by using only one of these predictors for each data set. You can see the rather negligible results this had on results in the 1st tier section below.

* 1. Metrics

On the subject of metric dependency, running significance tests in R showed that the Noun, Verb, Adjective, Word and Response Time metrics are all statistically dependent across all answer sets under examination. However, this is a very reasonable result - not every word in a sentence can be a singular part of speech assuming proper grammar (Another reasonable assumption for highly voted and/or accepted answers), meaning that as the number of words under one part of speech go up, the others will naturally rise to fill out the sentences. Furthermore, longer answers simply take more time to write, therefore some level of dependence is to be expected. However, link and code snippet presence are independent of both each other and all other metrics under consideration, another logical result seeing that copy/pasting a hyperlink or code snippet can easily be done regardless of the rest of your answer.

* 1. 1st Tier Top (57156 answers) vs Accepted (57156 answers)

As discussed previously, we chose to narrow our set of predictors after correlation analysis. The impact this had on our results is negligible as seen in Table 1 below. The decrease in classification accuracy may be data set specific, or may reflect an insufficient weighting of the number of words by the ADT algorithm.

|  |  |  |
| --- | --- | --- |
| Impact of Condensing Dependent Variables | | |
|  | Correct (%) | Incorrect (%) |
| 7 Attribute Model | 70.3758 | 29.6242 |
| 4 Attribute Model | 70.2122 | 29.7878 |
| Difference | -0.1636 | -0.1636 |

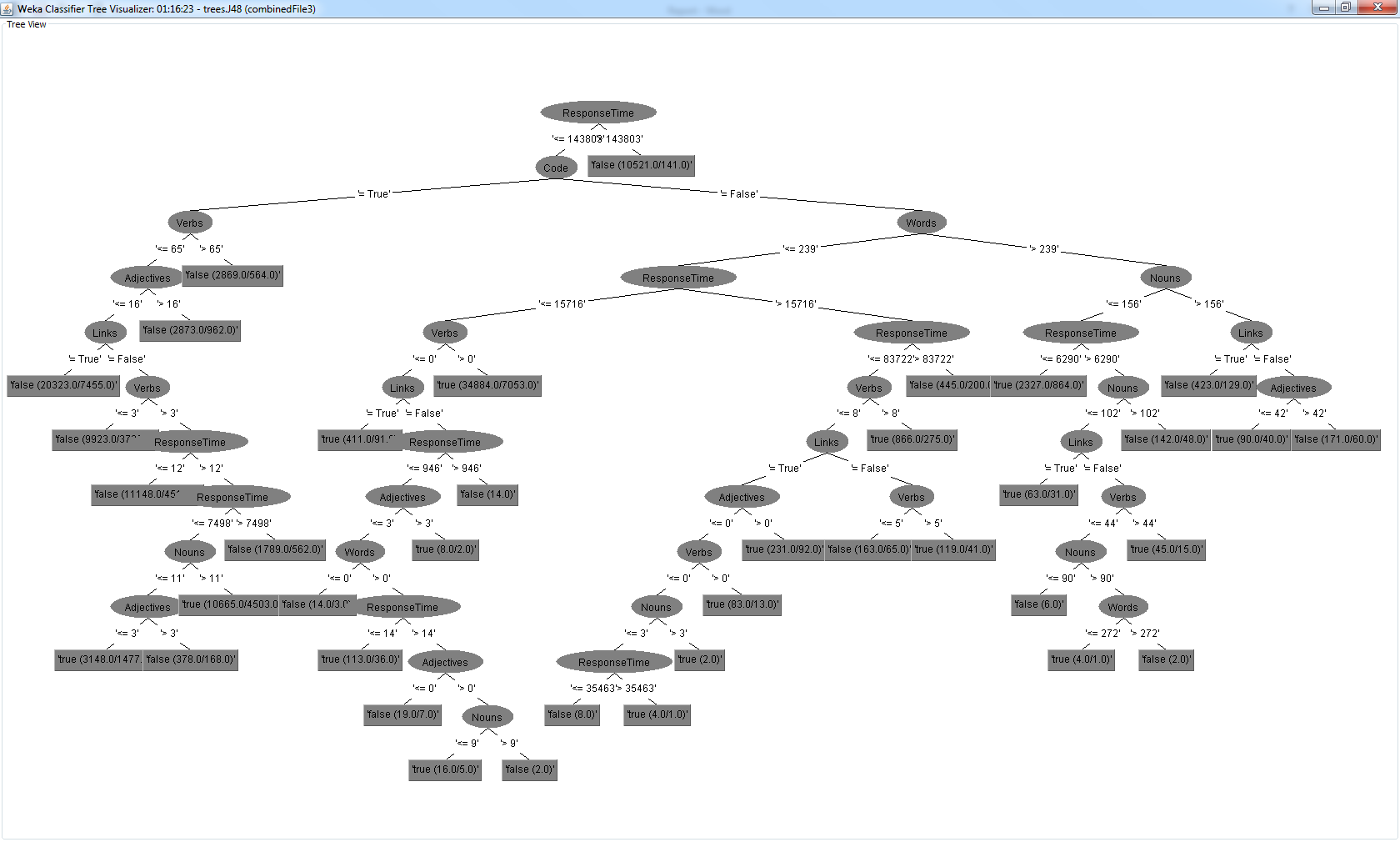
**Table 1:** Differences in model accuracy resulting from reducing the number of variables

For our classification testing, we chose to use chose to use 10-fold cross-validation in order to have a larger training and testing sample size as our sample size was relatively small for this test. This had a rather insignificant positive offset on our classification data. Assuming the negative difference seen in Table 1 is not due to chance, the positive difference associated with validation and the negative difference from variable pruning effectively cancel out.

|  |  |
| --- | --- |
| Cross-Validation versus Percentage Split | |
|  | Accuracy (%) |
| Cross-Validation (10-fold) | 70.3758 |
| Percentage Split(80/20) | 70.1382 |
| Difference | 0.2376 |

**Table 2:** Differences in model accuracy resulting from measuring accuracy against cross-validation or percentage split.

The model from our ADT, shown in graphical format in Figure 1, was able to correctly classify instances 70.64% of the time. The model incorrectly classified answers the remaining 29.36% of the time. A total number of 114312 instances were used to build and validate the model.



**Figure 1:** Alternating Decision Tree with 31 nodes where a final value closer to 1 corresponds to an answer being accepted and a final value closer to 0 corresponding with a top rated answer.

Using our 4 metrics, we looked at the impact each metric had on predicting whether an answer is an accepted answer. Table 3 and Table 4 contain the most relevant information we collected. Notable outtakes from the tables are that number of words had very little impact on accurate prediction (~6%) and presence of links had almost zero impact (~0.5%). Conversely, presence of code had the highest single variable impact (~17%) with response time also having a noticeable impact (~14%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Correctly Classified Instances (%) | | | | |
| w,l\\c, r | c,r | c,!r | !c,r | !c,!r |
| w,l | 70.2122 | 67.6237 | 64.4788 | 56.7176 |
| w,!l | 69.6401 | 67.6158 | 64.347 | 55.5901 |
| !w,l | 69.5171 | 66.7177 | 63.2972 | 50.503 |
| !w,!l | 69.0487 | 66.7177 | 63.2972 |

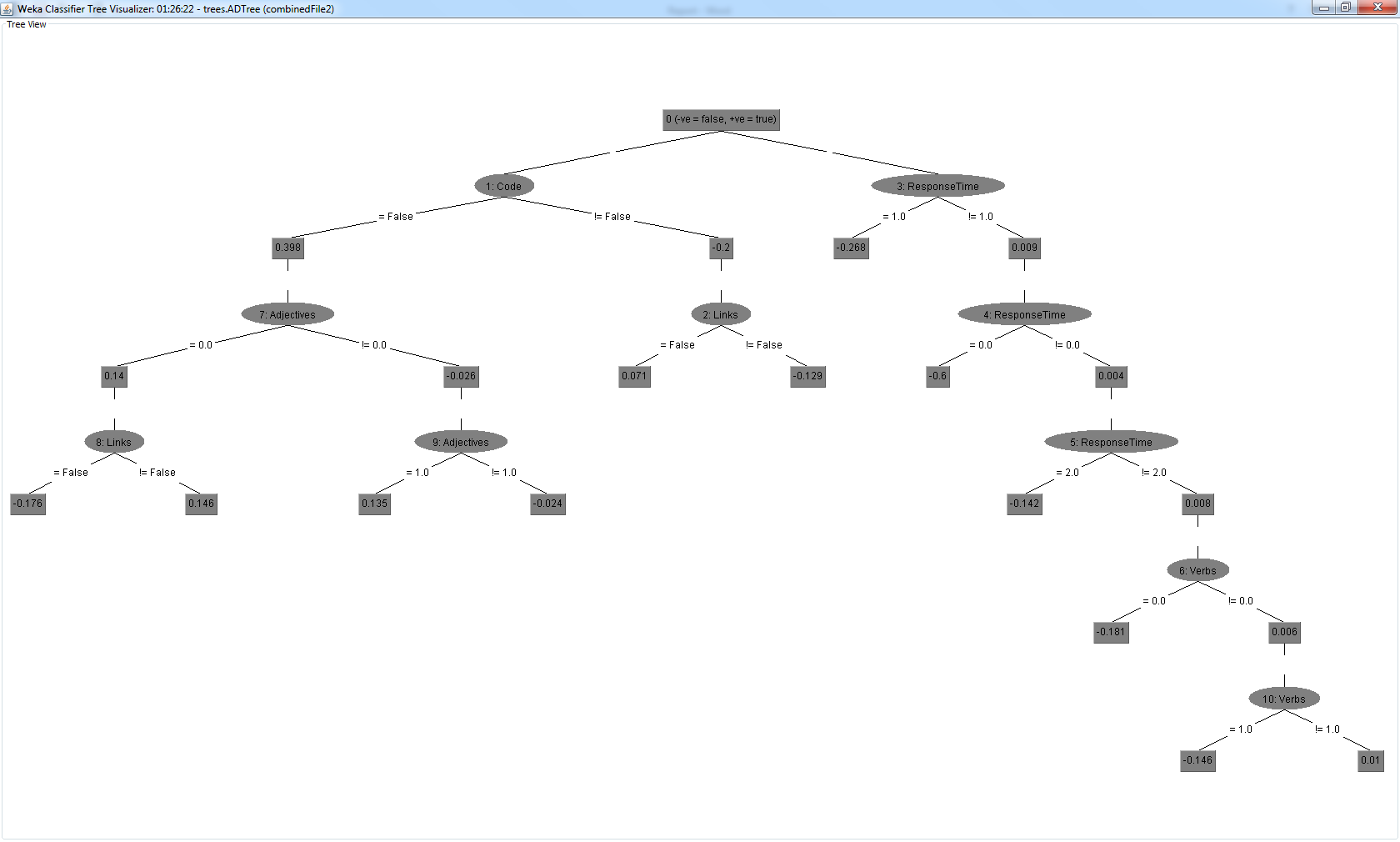
**Table 3:** Heatmap showing the impact of different variables on model accuracy (w= number of words; l = presence of links; c= presence of code; r= response time). Green is desirable, red is undesirable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Incorrectly Classified Instances (%) | | | | |
| w,l\\c, r | c,r | c,!r | !c,r | !c,!r |
| w,l | 29.7878 | 32.3763 | 35.5212 | 43.2824 |
| w,!l | 30.3599 | 32.3842 | 35.653 | 44.4099 |
| !w,l | 30.4829 | 33.2823 | 36.7028 | 49.497 |
| !w,!l | 30.9513 | 33.2823 | 36.7028 |

**Table 4:** Heatmap showing the impact of different variables on model accuracy (w= number of words; l = presence of links; c= presence of code; r= response time). Green is desirable, red is undesirable.

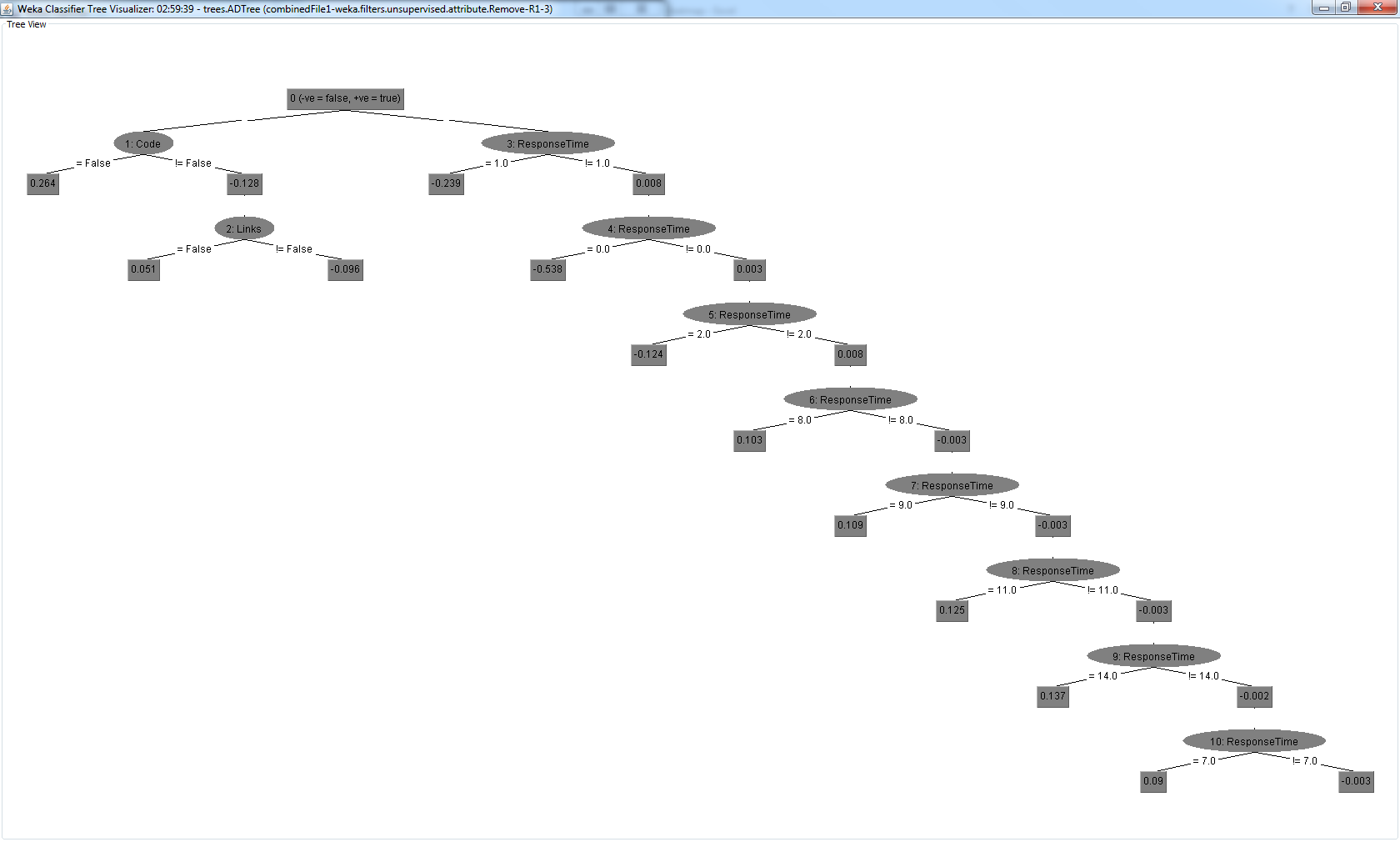
1. 2nd Tier Top (98194 answers) vs Accepted (98194 answers)

The model from our ADT for the second tier of top answers versus accepted answers, shown in graphical format in Figure 2, was able to correctly classify instances 62.99% of the time. The model incorrectly classified accepted answers the remaining 37.01% of the time. A total number of 196388 instances were used to build and validate the model. The accuracy of this model is much less significant than the highest tier of top answers (a drop of ~6%).

**Figure 2:** Alternating Decision Tree for the 2nd tier of highest scoring answers versus accepted answers. A final value closer to 1 corresponds to an answer being accepted and a final value closer to 0 corresponding with a top rated answer.

1. 3rd Tier Top (225000 answers) vs Accepted (225000 answers)

The model from our ADT for the second tier of top answers versus accepted answers, shown in graphical format in Figure 2, was able to correctly classify instances 58.59% of the time. The model incorrectly classified accepted answers the remaining 41.41% of the time. A total number of 425000 instances were used to build and validate the model; this number was rounded down due to hard memory constraints of the computer we were running tests on. The number of records was relatively small compared to the sample space (14446).The accuracy of this model is even further reduced compared to previous models (a drop of ~4.5% from the 2nd tier and ~12% from the 1st tier).



**Figure 3:** Alternating Decision Tree for the 3rd tier of highest scoring answers versus accepted answers. A final value closer to 1 corresponds to an answer being accepted and a final value closer to 0 corresponding with a top rated answer.

1. **Limitations and Threats to Validity**

There are a number of threats to the validity of our report which we realized late in our research.

First, we looked at the top rated answers overall. Based on the way that votes are distributed, it is likely that the pool of top rated answers is not distributed evenly across questions and many answers could have come from the same question. This means that there could have been a better written answer with lower visibility while our analysis instead took a less quality response because it had higher visibility.

We did not normalize our data set to remove duplicates (results that are both accepted and top rated). This means that the models we built may be more inaccurate because they were fed results that were identified as distinguishable, but in reality are not indistinguishable.

There’s an inherent limitation on score that must be considered and that we did not account for. Some questions will get more views and be more popular than others, and these questions will naturally draw more votes. Therefore, a well-constructed, correct answer to a highly popular question will score significantly higher than an answer of similar quality on a less popular question. This is a factor that is impossible for us or anybody else to account for, since question popularity is a highly variable and unpredictable factor and that information is also not contained in the data dump provided

1. **Conclusions**

From our analysis, there are a number of conclusions that can be made. First, because of the relationship between classifier accuracy and score threshold of top answers, we demonstrated that there is no significant correlation between a Stack Overflow answer being accepted and an answer having a high score. This can be partially explained by the failure to remove duplicate answers from both data sets, however, it stands to reason that questions with more traffic have multiple answers with high scores whereas only one high scoring answer per question can be selected. As for distinguishing highly rated and accepted answers, presence of code snippets and response time are decent predictors of accepted answers while presence of links is effectively useless as an indicator. It is interesting to note that there is even a difference in terms of content between these posts. From our model, we can see that question askers are relatively more interested in code and quick responses whereas the community are less inclined to include these features. Finally, as can be expected, words, adjectives, verbs, and nouns are highly correlated and are comparatively ineffective for analysis when used in unison. In finality, “what makes for a good answer?” is a difficult question to answer and really depends on who is asking. A question asker appears to want a quick reply with a code snippet – presumably an example or finished code. Both question askers and the community like to see links – presumably to learn more on the topic at hand.

1. **Future Work**

This study can be improved in a number of ways. First, our study was limited in effectiveness by including duplicate results in the analysis. These could be removed to generate more accurate models. Secondly, a significant improvement could be made by normalizing the score of the answers based on the score of the question to quantify popularity. We believe that score may be directly related to the number of views that a question receives and therefore may have little effect on whether an answer is accepted or not.

1. **Related Works**

A. Bosu, C. S. Corley et al. "Building reputation in StackOverflow: an empirical investigation", <i>Proceedings of the 10th Working Conference on Mining Software Repositories</i>, pp 89-92, 2013

C. Treude, O. Barzilay. M. Storey, "How do programmers ask and answer questions on the web?", <i>Proceedings of the 33rd International Conference on Software Engineering</i>, pp. 804-807, 2011

D. Movshovits-Attias, Y. Movshovits-Attias, P. Steenkiste, C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: StackOverflow", <i>Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on</i>, pp. 886-893, 2013

V. Baht, A. Gokhale et al. "Min(e)d Your Tags: Analysis of Question Response Time in Stack Overflow", <i>Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on</i>, pp. 328-335, 2014