# Sentiment Analysis for Homework Reviews

**Hive Implementation**

Letian Chang, Liuyuan Tan, Vince Campanale

May 5, 2016

# 1 Introduction to Problem (What?)

With this final project, we want to show that we have learned a reliable approach to analyzing the sentiment, the emotion of the author, for a given document. In order to test our approach, we worked with "Homework Review" data collected over the course of the semester. The homework reviews consisted of a minimum of fifty words and were written by each student at the end of each homework assignment. Each student was prompted to write how they felt about completing the homework immediately after reading it. Our goal is to accurately asses the sentiment the writer had at the time of writing their review. We used MapReduce and Hive to pre-process our data and to conduct sentiment analysis on the resulting processed data. We found that we were able to accurately assess the sentiment of the document 87.5 percent of the time. Through our approach iterations, we learned that n-grams are a more effective way of analyzing sentiment, because they incorporate words and phrases which alter meaning. For example, "This assigment was not fun" vs. "This assignment was fun." If we were to only count positive and negative words, both of these sentences would most likely return as postive due to the presence of the word "fun" and the likelihood of "not" being filtered as a stop word. However, using n-grams, we could accont for the "not" in front of "fun" and produce a much more reliable result.

# 2 Motivation to Solve this Problem (Why?)

Our motivation for this problem is to develop our skills in analyzing data effectively. We want to be able to apply our implementation to any document, not just a homework review, and accurately assess, to the best of our ability, what the actual sentiment of that document is. At a higher level, this is a quickly developing field which is very useful for review websites, and any business that relies on feedback from customers. For our purposes, this project is an exercise in bettering our development abilities and our understanding of the way Hive works. Sentiment analysis with Hive has major applications in business intelligence. For example, if a company wanted to ask "Why aren't customers buying our cellphones?" we could look at the problem from multiple angles. We could look at the concrete data certainly, but this gives us a limited view of the product itself. It limits our assessment of it to a view hard and unchangeable variables: price, competition, specs to name a few. However, with sentiment analysis, we can analyze the "behind-the-scenes" data, such as customer reviews which describe real users interactions with the product.

Sentiment analysis also plays a major part in political campaigns. Anybody can look at poll data and see in binary form "Do you support candidate x?" but what if we could further analyze specific people's opinions on candidate x's policies and debating style. Then we could get a more holistic view of the populus' support for candidate x and better manage their campaign.

The real-world applications of effective sentiment analysis techniques are numerous, there are too many to name all of them here, however at the base of all of those applica-

tions, a useful and reliable technique must exist. Our goal with this project, using our own homework reviews as a benchmark and testing sample, is to see how well we can analyze the sentiment of a document using Hive.

# 3 Description of Approach (How?)

Our approach to analyzing the sentiment of our given homework reviews was two-fold: first, we pre-processed the raw data so that it would be fit for accurate analysis; second, we did two forms of analysis on this data, which we will describe later in this section.

## 3.1 Filtering (using Pig)

First, we filtered. To do this, we used Pig in order to filter out all empty reviews and any reviews shorter than 50 words. We deemed these reviews as having insufficient material from which to deduce an accurate sentiment. After filtering these files, we deleted any reviews with shorter than 50 words. Then, from the remaining reviews, we removed all non-word characters (such as newline and tab characters), put the homework number at the beginning of each file, and separated this homework number from the actual content of the review with a tab character. In addition, we added a new line at the end of each review so they would be evenly spaced and easier to parse through.

## 3.2 Text Pre-processing (using MapReduce)

The next step in the processing portion of our project entailed using MapReduce to organize and further process the filtered data. We modified the filtered data into text files with two tab-delimited fields: "hw_number" and "review." We then got rid of all the stop words in these text files using the English.stop file we were given. We changed the stop words list slightly to encompass certain outliers and frequently used words that we found in our first round of analysis, which we did not deem as relevant to determining the sentiment of the author. For example, we removed "not" from the stop words list because it clearly impacted our N-grams, which will be discussed further in section 3.4. Finally, we changed all of our text data to lower-case letters, hoping to effectively prevent double counting or over-counting words. Once all of this was done, we were ready to begin our formal analysis!

## 3.3 Basic Analysis using Hive (Word Lists)

We started our analysis by simply using a word count approach. We did this using the pos-words.txt and neg-words.txt files provided in our SVN repository. We modified these lists slightly to more accurately reflect the nature of the content we were analyzing. For example, the term "problem" may be negative in most cases, but we found in many of the situations in the reviews, it was being used neutrally, simply to refer to which problem in the assignment the author was addressing. Then, we used Hive to load the reviews and word lists, compute the number of positive and negative words in each review. After that,

we devided the frequency of positive and negative words by the total number of reviews. We did this for each homework assignment, compared the positive words/reviews and negative words/reviews ratio, and found out which homework assignments got the most positive and negative reviews on average across all students. After finding these averages, we went on to find out which positive and negative words were used most frequently and listed the top five most frequently used words in each category. We will incude the detailed data and analysis in later sections.

### 3.4 More Advanced Analysis using Hive (N-grams)

A more informative way for us to conduct sentiment analysis would be to use N-grams, which are essentially strings of words 'N' long that allow us to account for differences such as "not easy" and "easy." We wrote a Hive script to compute N-grams for a given corpus of text documents, which was, in our case, the processed homework reviews. We wrote the script such that it could be run with two parameters: the first being the path to the documents we wanted to analyze (the processed homework reviews) and the second being the number of words we wanted in each N-gram. We retrieved the three most positive 2- and 3- grams and analyzed them to make sure that they made sense. After this trial run, we deemed 2-grams to be the most effective way to analyze the sentiment of the homework reviews. We used the N-gram approach to analyze the homework assigments from the entire class and our own homeworks. We evaluated how the class felt about each section of homeworks (first half vs. second half, and section by section) and determined which parts of the semester were viewed positively or negatively by the class as a whole. We then applied our approach to our own homework reviews to see how effectively it worked when compared with our own intuition when reading our reviews. Our results, discussed in the next section, were promising.

## 4 Results (Does Our Approach Work?)

### 4.1 Basic Analysis (Word Lists)

From cross-reference our positive and negative word lists with our homework review data, we found that the homework assignment which was percieved the most positive was homework number 1 and the homework assignement which was perceived the most negative was homework number 4. Detailed data could be reviewed in Figure 1 below.

We also found the number of positive and negative words used most frequently. The top five most frequently used positive words were: good, interesting, easy, pretty, easier; their respective frequencies were 143, 108, 92, 81, 47. The top five most frequently used negative words were: hard, difficult, long, confusing, frustrating; their respective frequencies were 118, 112, 70, 43, 40.

For the most part, these results make sense. The only parts where we raised an eyebrow were in the list of most positive words, "pretty" was probably being used as an appendage to another adjective such as "pretty easy" or "pretty hard" rather than the actual word "pretty" to describe the physical appearance of the homework assignment.

Figure 1: Positive/Negative Word Frequencies and Ratios Summary

| Homework # | Total Reviews | Positive Words Frequency | Negative Words Frequency | Avg. Positive Words (words/review) | Avg. Negative Words (words/review) |
|---|---|---|---|---|---|
| 1 | 81 | 243 | 175 | 3.00 | 2.16 |
| 2 | 74 | 170 | 111 | 2.30 | 1.50 |
| 3 | 74 | 166 | 121 | 2.24 | 1.64 |
| 4 | 80 | 155 | 188 | 1.94 | 2.35 |
| 5 | 80 | 190 | 153 | 2.38 | 1.91 |
| 6 | 75 | 179 | 158 | 2.39 | 2.11 |
| 7 | 74 | 203 | 106 | 2.74 | 1.43 |
| 8 | 74 | 154 | 135 | 2.08 | 1.82 |
| Theory (1-2) | 155 | 413 | 286 | 2.66 | 1.85 |
| MapReduce (3-6) | 309 | 690 | 620 | 2.23 | 2.01 |
| PIG (7-8) | 148 | 357 | 241 | 2.41 | 1.63 |
| First Half (1-6) | 464 | 1103 | 906 | 2.38 | 1.95 |
| Second Half (7-8) | 148 | 357 | 241 | 2.41 | 1.63 |
| Total | 612 | 1460 | 1147 | 2.39 | 1.87 |

Also, "easier" may have been a part of a larger phrase such as "could've been easier" which would actually count as a negative phrase. We anticipated these problems would be accounted for in our N-gram approach, discussed in the next section. The negative words, however, are all appropriate.

## 4.2 More Advanced Analysis (N-grams)

We used the lists positive.txt and negative.txt to determine whether a given n-gram was positive or negative. We found the three most frequent positive 2-grams from our analysis were "not hard", "not difficult", and "homework good." The three most frequent positive 3-grams were "homework not hard", "felt homework positive", "homework not difficult." These results are directly in line with what we predicted in the previous section. By removing "not" from the stop-words list and incorporating N-grams into our analysis, we were able to account for phrases that my consist of negative words but actually convey a positive sentiment.

In order to classify review data, we needed to come up with a classification method. In this particular case, we decided to find the ratio between positive and negative 2-grams, and set a threshold for classification. To find this threshold, we looked into our result data, and found that the ratio of positive to negative 2-grams was roughly 1.33 across all reviews, which means our students tend to use words on the positive words list, even though their true emotions might not exactly be positive. Therefore, we only deemed a review truly "positive" if it had a ratio of positive to negative 2-grams above 1.33. If the

ratio is below 1.33, we would classify it as "negative," and if the ratio is exactly 1.33, we would classify it as "neutral." If there were positive 2-grams but no negative ones, the ratio would be infinity , and, of course, we would classify it as "positive;" if there was no positive nor negative 2-grams, we'd say it's "neutral." We were able to break down each homeworks and determine which ones were viewed more positively and which were viewed more negatively (as shown in Figure 2).

Figure 2: Positive/Negative 2-Grams Frequencies and Ratios Summary

| Homework # | Positive 2-grams Frequency | Negative 2-grams Frequency | Positive-Negative Ratio on 2-grams | Classification |
|---|---|---|---|---|
| 1 | 248 | 162 | 1.53 | Positive |
| 2 | 168 | 106 | 1.58 | Positive |
| 3 | 162 | 119 | 1.36 | Positive |
| 4 | 153 | 180 | 0.85 | Negative |
| 5 | 190 | 142 | 1.34 | Positive |
| 6 | 177 | 156 | 1.13 | Negative |
| 7 | 202 | 97 | 2.08 | Positive |
| 8 | 150 | 130 | 1.15 | Negative |
| Theory (1-2) | 416 | 268 | 1.55 | Positive |
| MapReduce (3-6) | 682 | 597 | 1.14 | Negative |
| PIG (7-8) | 352 | 227 | 1.55 | Positive |
| First Half (1-6) | 1098 | 865 | 1.27 | Negative |
| Second Half (7-8) | 352 | 227 | 1.55 | Positive |
| Total | 1450 | 1092 | 1.33 | – |

When analyzing the 2-grams in the first half of the year reviews (hw1-6), we determined that the majority were positive, except for homeworks 4 and and 6 were distinctly negative. Homework 4 was especially lower and by far the most negatively reviewed homework of all. When comparing them to the second half of the year reviews (hw7-9), we found that the majority of the second half was relatively positive. Homework 7 was reviewed very positively. The differences between the homework assignments on theory (hw1 and hw2), MapReduce (hw3-6), and Pig (hw7-9) were quite interesting. We found that on average, the homeworks centered on theoretical concepts at the beginning of the semester received predominantly positive reviews. The homeworks centered on MapReduce received mostly negative reviews. The homework assignments made for Pig received quite positive reviews on average, even more so than the theoretical homework assignments at the beginning of the semester.

After all the processing and analysis above, we are interested in seeing whether our approach agrees with our own review data. We did this in two ways. First, we manually read our own reviews for each homework, labelled them "positive," "negative," or "neutral" based on how we feel as humans, then we compared these labels we assigned to our reviews to the classification our algorithm made for each homework review (shown in Figure 3). We found that we were able to assess our sentiment with 87.5% accuracy, which we consider very good.

Finally, we ran our own reviews through the exact same algorithm, classified them,

Figure 3: Review Labels and Accuracy of our Prediction Approach

| Homework # | Prediction | Letian | Vince | Liuyuan |
|---|---|---|---|---|
| 1 | Positive | Positive | Positive | Positive |
| 2 | Positive | Positive | Positive | Positive |
| 3 | Positive | Positive | Positive | Negative |
| 4 | Negative | Neutral | Negative | Negative |
| 5 | Positive | Positive | Positive | Positive |
| 6 | Negative | Negative | Negative | Negative |
| 7 | Positive | Positive | Positive | Positive |
| 8 | Negative | Negative | Positive | Negative |
| Accuracy | 0.875 | | | |

and compared the results with our own labels (shown in Figure 4). Some of our reviews didn't exceed 50 words, and were filtered out during pre-processing, hence the N/A in the table. Otherwise, the result came out very good as well, with an overall accuracy of 70% (N/A reviews not taken into consideration).

Figure 4: Review Labels and Accuracy of our Prediction Approach

| Homework # | Letian | | | Vince | | | Liuyuan | | |
|---|---|---|---|---|---|---|---|---|---|
| | Positive-Negative Ratio on 2-grams | Classification | Label | Positive-Negative Ratio on 2-grams | Classification | Label | Positive-Negative Ratio on 2-grams | Classification | Label |
| 1 | 1 | Negative | Positive | 0.5 | Negative | Positive | N/A | N/A | Positive |
| 2 | ∞ | Positive | Positive | 1.34 | Positive | Positive | N/A | N/A | Positive |
| 3 | 1.67 | Positive | Positive | ∞ | Positive | Positive | N/A | N/A | Negative |
| 4 | 1.33 | Neutral | Neutral | 0.5 | Negative | Negative | 1.5 | Positive | Negative |
| 5 | 1 | Negative | Positive | 1 | Negative | Positive | 4 | Positive | Positive |
| 6 | 1.25 | Negative | Negative | N/A | N/A | Negative | 1 | Negative | Negative |
| 7 | 1.25 | Negative | Positive | ∞ | Positive | Positive | ∞ | Positive | Positive |
| 8 | 0.5 | Negative | Negative | ∞ | Positive | Positive | 1 | Negative | Negative |
| Overall Accuracy | 0.7 | | | | | | | | |

## 5 Discussion (What's Next?)

While our analysis was reasonably effective, or effective "enough," we still believe there are several ways in which we could improve our approach. First of all, a larger dataset would greatly improve our results. Because we only had one semester worth of data to work with, and a significant portion of this data was filtered out, we ended up with a pretty small dataset to work with. This small sample size certainly skewed our results somewhat. With longer reviews and more students, we would have much more accurate predictions.

Second, a 2-gram is a relatively simple method. We believe that using a more complex and in-depth machine learning algorithms, we could have a more thorough analysis.

Finally, we think it would be useful to quantify the sentiments in the form of some sort of scale and then categorize each document into sections. For example, using a 1 to 5 scale, we could use 5:Great, 4:Good, 3:Average 2:Bad, 1:Terrible. This would allow

for more accurate and holistic interpretations of the assignments, and the introduction to more complex algorithms such as the cosine similarity, or Pearson correlation.

Overall, our approach works. It is a good foundation for sentiment analysis. However, taking further steps such as improving our algorithm, accumulating more data, and increasing the number of classifications, would certainly be beneficial.