Talita Anthonio
S2477106
Creating Data
Report 1
Master Digital Humanities

Using Crowdsourcing for Sentiment Analysis:

A Study on the Manual Annotation of Tweets about Clinton and Trump

06-11-16

Preface

This report was written for the course Creating Data (Master Digital Humanities). It covers a detailed description of a project in which I had to participate with a group of six other students. The purpose of this project was to manually annotate a large amount of Twitter data with the use of crowdsourcing. In order to succeed within the short time available, it was necessary to make a clear division of labour. My responsibilities in this project consisted of writing the guidelines and extracting the data for crowdsourcing. Overall, I am very satisfied with the collaboration in this group. Everyone participated equally and all members were motivated to help others when necessary.

1 Introduction

Twitter captures information about human behavior. Because of this, Twitter data can be useful to determine differences in language between groups, such as differences between female and male language (Armstrong and Gao, 2011). In addition, Twitter data can be useful to determine opinions about a certain entity. The most common method for such analysis is sentiment analysis (Medhata et al., 2014; Sylwester and Purver, 2015). This is often conducted by a program that automatically annotates the sentiment of tweets. In order to test such programs properly, accurate manually annotated data by humans is necessary. Yet it can be rather difficult to acquire such data. It can be hard to find participants who are willing to perform the annotation task, especially when it requires domain-specific knowledge (Alvaro et al., 2015). Moreover, elements such as the instructions can influence the accuracy of the data (Pustejovsky and Stubbs, 2012). Because of this, it is important to know which issues can be expected when sentiment analysis is conducted manually.
However, there is not much literature available about the issues encountered during manual annotation and in particular not with manual sentiment analysis. Therefore, the aim of this research is to provide a description of the issues that are encountered during the manual annotation of sentiment in tweets. As the annotation task concerned labeling a sentiment to a tweet, this research can be motivated by the following question:

RQ: Which issues are encountered when the sentiment of a tweet is manually annotated?

In order to answer this question, we designed a task on a crowdsourcing website called Crowdflower (`www.crowdflower.com`). The job of the annotators was to determine the sentiment of tweets about Hillary Clinton or Donald Trump. Tweets from June and August were obtained in order to determine how sentiment developed in these months. As Clinton and Trump are the current presidential candidates of the United States, this work can also be of interest for political purposes.

2 Related Research

In previous research, three labels are frequently used to classify the opinion towards subjects on Twitter data. There is a distinction between positive, negative and neutral sentiment (Sanghvi et al., 2015; Gurkhe, 2014; Spencer and Uchyigit, 2012; Kouloumpis et al., 2011). These three labels are also frequently used to determine the opinion of people towards politicians from the United States (Ringsquandl and Petkovic, 2013; Taddy, 2013). For instance, Ringsquandl and Petkovic (2013) used sentiment analysis of tweets to examine the sentiment towards a series of politicians of the Republican party. With the use of NLTK, they examined which words were frequently used to express a positive, negative or neutral sentiment. Their findings show which words and names are often used to express negative and positive sentiments towards politicians. For instance, tweets that mention Mitt Romney and President Obama are often used to express a negative sentiment. Ringsquandl and Petkovic (2013) did not use any manual annotation in their project, as the aim of their research was to provide insights in machine learning.
In contrast, Taddy (2013) used automatically annotated data as a primary source and manually annotated data as a secondary source. The manually annotated data was obtained with a crowdsourcing platform called AmazonMechanicalTurk. With an agreement score of 80 percent, their work showed that data with manual annotation of sentiments can be a reliable source as well.

Yet there has been research conducted in which various other labels were used. An example of such work is conducted by Sylwester and Purver (2015). In addition to positive and negative sentiment they used labels such as certainty, uncertainty, and feeling. This was due to the purpose of their research. Instead of measuring the opinion of people towards an entity, the sentiment analysis was conducted in order to provide insights in linguistic differences between democrats and republicans. Their results indicate that democrats express more negative sentiments by using swear words. Furthermore, their work shows that instead of using three categories of sentiments, more categories can be useful as well.
In summary, most research focuses on automatically annotated data, rather than manually. One exception is the study conducted by Taddy (2013). However, even in this study there is little information given about the manual annotation. Furthermore, although an agreement score of 80 percent might indicate that the sentiment task was doable, better insights could be achieved if there was descriptive data available about the task itself. Besides, this section shows that sentiment analysis is often conducted with the use of three labels: positive, negative and neutral. Because of this, it is questionable whether another label would positively affect the reliability of the data.

3 Data

3.1 Sampling

Our final data set was a sample of Twitter data previously collected by the University of Groningen. This data set includes tweets that refer to Hillary Clinton and Donald Trump. The tweets are dated from January 2016 until September 2016. Each file from this data set is named after the candidate, the number of the month and the time in hours. Thus, this meta data is not specified in the file itself. The file itself consists four columns. The first column indicates the label of the tweet, which is represented by the label Unknown (UNK). The second column shows the unique identifier of the tweet and the third shows the unique identifier of the user. The last column features the tweet itself. Every line consists of one tweet.
From this data set, we extracted 3,000 tweets from June and 3,000 from August. From every candidate, we collected 1500 tweets from July and 1500 tweets from August. In this way, the data set was suitable for the statement of purpose for the annotation task, which was formulated as follows:

We want to use labels to detect the sentiment of tweets in order to determine how sentiment has developed in June and August

In order to extract the sample, we implemented a script (see Appendix A, Image 1). This script executed four steps. Firstly, it extracted all the tweets from either Trump or Clinton from June at 8 pm. Secondly, it removed all the tweets that contained 'Clinton' and 'Trump'. Thirdly, it extracted the first 1500 lines of that text file. Fourthly, this file was divided into three files of 500 tweets in order to chunk the data into smaller pieces. All steps were executed four times in total.

3.2 Modification

Our primary intention was to upload smaller units of the data set on Crowdflower. This would allow us to test the design before uploading the whole data set. Due to time issues this was not possible. Therefore, our final data set included all 6,000 tweets in a single file. It was necessary to modify this file because the month and the name of the presidential candidate were not indicated in the file itself. We added two more columns, namely: the month, which was either 06 or 08, and the name of the candidate. Furthermore, we removed the 'UNK' column, as it was not possible to provide in-line annotation with Crowdflower. The result of this is visible in Image 2 from Appendix A.

3.3 Corpus Analytics

The tweets of our final data set included 77,118 words and 517,892 characters. With the use of NLTK, a frequency distribution of this data set was plotted. The script used for this plot is given in Appendix

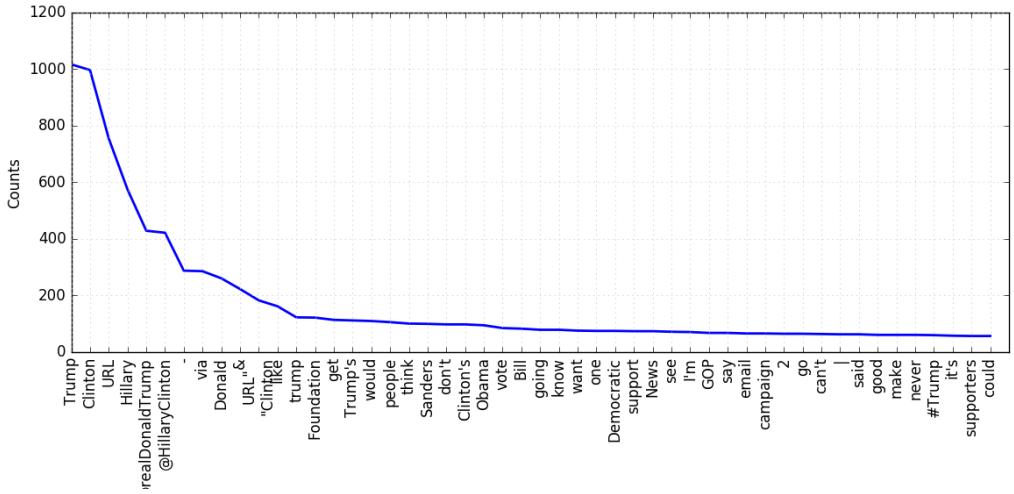

Figure 1: Plot for 50 most Frequent Words in Final Data set

A, Image 3. As is visible in the script, the frequency distribution does not include stop words. The final plot is visible in Figure 1.

It is remarkable that Trump is mentioned more frequently in the tweets (1016 times) than Clinton (996 times). In addition, more users tweet to Donald Trump (RealDonaldTrump, 428 times) than to HillaryClinton (HillaryClinton, 421 times) Furthermore the plot shows that the data set contains words used to express a sentiment. For instance, the word 'support' appears 73 times in this corpus. This word can be used to express either a negative sentiment, such as "I do not support Trump" or a positive sentiment. However, the data set also includes an amount of words that are not directly used to express a sentiment, such as 'see' and 'one'. Because of this, it might be the case that lots of tweets will be labeled with a neutral sentiment.

Moreover, other politicians are mentioned frequently in the data set, such as Obama (94 times) and Sanders (99 times). This might indicate that the data set includes tweets that express a sentiment towards other politicians, rather than Clinton or Trump.

Although this is section provides limited information about the data set, the results show that this set is suitable to determine the sentiment towards Clinton and Trump. Firstly, Clinton and Trump are mentioned frequently. Secondly, the data set contains words that could be used to express a sentiment. However, the question remains whether the data set also includes tweets which express a sentiment towards other politicians. Therefore, it was important to take this into account when formulating the instructions for the crowdsourcing job.

4 Method

As the final data set contained 6,000 tweets it was necessary to use crowdsourcing to annotate the sentiment of the tweets. The reliability of data obtained from crowdsourcing has often been questioned (Alvaro et al., 2015; Staiano and Guerini, 2014; de Winter et al., 2015; Pustejovsky and Stubbs, 2012). There are multiple elements that can negatively affect the accuracy of the annotation, such as the fee that the annotators receive (Alvaro et al., 2015) and the clarity of instructions (Pustejovsky and Stubbs, 2012). Therefore, the general approach of our method consisted of developing, testing and revising all the material we created. In total, we had three such stages prior to the final launch of the annotation job on Crowdflower.

In this section, the process that led us to the final material and the final Crowdflower settings is discussed in detail. Firstly, the establishment of the annotation model will be discussed. Secondly, the development of the guidelines and the resulting final version will be presented. Thirdly, the annotation rounds we used to test the material will be discussed. In the same section, the settings we used for the final Crowdflower

job will be presented.

4.1 Annotation Model

Two models were used in this project. The final model was a revision from the first model. In order to describe both models, the definition of Pustejovsky and Stubbs (2012) is used. They state that a model consists of three parts: the vocabulary of terms (T), the relations between these terms (R) and their interpretation (I). Thus, our first model can be described as follows:

- T = {Sentiment, positive, negative, neutral, irrelevant}
- R = {Sentiment ::= Positive | Negative | Neutral | Irrelevant}
- I = {Positive = "tweets that express a positive sentiment towards the presidential candidates", Negative ="tweets that express a negative sentiment towards the presidential candidates", Neutral ="tweets that express neither a positive sentiment or a negative sentiment", Irrelevant = "tweets that refer neither to Hillary Clinton or Donald Trump"}

After the second stage of annotation our model was revised. This resulted in our final model, which can be described as follows:

- T = {Sentiment, positive, negative, other}
- R = {Sentiment ::= Positive | Negative | Other}
- I = {Positive = "tweets that express a positive sentiment towards the presidential candidates", Negative ="tweets that express a negative sentiment towards the presidential candidates", other ="tweets that express neither a positive sentiment or a negative sentiment towards the presidential parameters"}

The reason why the model has been revised will be explained in Section 4.3, in which the stages of annotation are presented.

4.2 Guidelines

After we formulated our model, we wrote the annotation guidelines for the annotators. Once we finished the model and the guidelines, they were revised after it was tested. This process was repeated until there was a final version of the guidelines. In the end, we changed the guidelines two times. Firstly, after we revised our first model. Secondly, after the second stage of annotation.
The main challenge we encountered when writing the guidelines was how to write comprehensive guidelines in a low amount of words. If the text of the guidelines is too long, than there is a higher chance that the annotators will not read it (Pustejovsky and Stubbs, 2012), especially with non-experts. We tried to write the guidelines as short and comprehensible as possible.
The first version of the guidelines is visible in Appendix B. These guidelines were written for the first annotation stage. In this stage the annotation was done in the data files themselves. Therefore, this version of the guidelines has the most words because it was necessary to make sure that the annotators understood how the data file was structured.
The second version of the guidelines is visible in Appendix C. This was the first version after we revised the model. Thus, the major change in this version is that the neutral-label and the other-label are merged by one label. We did not write any new instructions for this label. Instead, we placed all the instructions from the neutral label and the irrelevant label in one column.
The last and final version of the guidelines is visible in Appendix D. This is the version we used for Crowd-flower. Therefore, the instructions of this version differ from the other versions. Instead of providing the annotation in a data file, the annotators provided the annotation with the use of multiple chose questions. Because of this, we had to change the instructions. Another change we made was the descriptions for the labels in order to make sure that the text was not too long.

4.3 Stages of Annotation

The previous section showed that there are multiple issues to consider when the guidelines of an annotation task are developed. One important issue was how to make sure that the guidelines are clear and short. This section will describe how we tried to deal with such issues and how that led to our final decisions.
There were three annotation stages prior to the final stage in which the data was annotated with Crowdflower. Thus, the first three stages of annotation can be viewed as a single process, aimed to develop the final data and design for crowd-sourcing. In this section, a description of each annotation stage is given.

4.3.1 First Stage of Annotation: Two Annotators

The aim of the first stage of annotation was to explore whether the guidelines were clearly written and whether the task was feasible with the given data set. The annotation was performed by two members from our group. The data they annotated was a sample from the original data set, extracted from January 8 pm. Every annotator received two files from the original data set. One file included 50 tweets from Clinton and the other file included 50 tweets from Trump. After they finished annotated, we calculated the agreement scores. These scores are given in Appendix E. As is visible, Cohen's Kappa for Clinton was low (0.26). Cohen's Kappa for Trump was moderate (0.54). We discussed with the annotators which problems they encountered. Overall, there was a confusion about the differences between the neutral and the irrelevant label. This mainly caused by one annotator, who did not read the guidelines properly. Because of this, we decided to not change the guidelines until the second stage of annotation, as it was not sure whether the difficulties were caused by the unclear written instructions. We hoped that the conclusions from the second stage would lead to more specific problems which we could solve with either changing the guidelines or the data set.

4.3.2 Second Stage of Annotation: Seven Annotators

The aim of the second stage of annotation was to annotate a larger amount of tweets (1000 tweets) by all members from our group. Every person from our group received approximately 143 tweets. Again, every person received two files. One file with tweets from Clinton (approximately 70) and one from Trump (approximately 70). We made sure that the data was distributed in such a way that the annotators from the previous round did not receive the same data set.
After everyone had completed their annotation task, we wanted to calculate the agreement scores. However, this was not possible because of time issues. In order to still get an overview of the problems, we discussed them briefly in our group. Again, the annotators encountered difficulties with discriminating between a neutral label and an irrelevant one. According tot the annotators, the description of the labels seemed almost the same. Because of this, we decided to revise the model. This revision implied merging the irrelevant and neutral label into one. We called this label 'other'. In order to make the guidelines fit the model, we revised the guidelines as well.

4.3.3 Third Stage of Annotation: Crowdflower

The third stage of annotation was the first time when we used Crowdflower. The settings we used for this round on Crowdflower are visible in Table 1. As is visible in Table 1, we excluded participants from the US and Mexico to prevent bias. As the data set was about US politics, there was a high chance that they would get a higher performance of the task because of prior knowledge.
In order to upload the data correctly on Crowdflower, we needed to replace all the comma's in the text from the tweets. Furthermore, we had to make sure that the files were encoded in UTF-8. Once these issues were solved, we uploaded a small amount of tweets (100) from the final data set, as this round was used as a test rather than a final round.
Once we uploaded the data, we developed the test questions. Test questions are used to guarantee the quality of the data and are presented during the annotation job. If the annotator makes a lot of mistakes, than all the work by the annotator will be removed. Thus it is important to accurately develop the test-questions. Because of this, everyone of our group took part in designing the questions. Furthermore, we only submitted a test-question if everyone agreed about its answer.

One we submitted the test-questions, it was finally possible to launch the job. After the annotation was finished, we evaluated the results. The conclusion of this evaluation was that we had to re-upload the data. This decision was based on two issues. Firstly, we noticed that we uploaded the wrong data set. Instead of the tweets from June, the data set contained tweets from July. Furthermore, it was remarkable that almost 70 percent of the tweets were labeled as 'other'. According to our guidelines, every tweet that contained the names Clinton and Trump should be labeled as 'other'. When we investigated the data, we counted a lot of tweets that contained both names. Because of this, we decided to exclude tweets that mentioned both 'Clinton' and 'Trump' for the next round of annotation. Furthermore, we decided to increase the payment with 0.01 cent. The payment was rated very badly (2.5 / 5). If the payment is too low, there is a higher chance that users on Crowdflower will not contribute in the project (Pustejovsky and Stubbs, 2012). As we only had one week left to gather annotated data it was necessary to change the payment.

Setting	**Value**
Payment	0.01 cent per 10 tweets
Annotators per tweet	3
Data	100 tweets from Clinton (June)
Guidelines	Version 3
Location	Non US and Non-Mexico
Test Questions:	13

Table 1: Crowdflower Settings For Third Annotation Round

4.3.4 Fourth and Final Stage of Annotation

The aim of the last stage of annotation was to annotate all 6000 tweets via Crowdflower. In this stage we used the data set described in section 3. This file did not include tweets with both names.
We encountered a lot of difficulties in uploading the data on Crowdflower. Firstly there were troubles with the required file extension. After we converted the files to the right file extension with the right encoding, there was another issue. As Crowdflower converts all the data to a json-format, double quotes inside the text of the tweets were identified as separate elements. Because of this, we had to modify our data set again. With the use of a python script (see Appendix F, Image 1), we put the text of the tweet in quotes. All double quotes inside the text of the tweets were replaced by fourth quotes. This made it finally possible to upload our data to crowd-flower.
The final settings for Crowdflower are visible in Table 2. Apart from the data and the amount of payment, we also made a change with regards to the test-questions. In the previous annotation round, we used 13 test-questions (see Table 1). As we used a larger amount of data for the final stage, we decided to use more test questions.
One day after we launched the job, we paused the job to evaluate the process. Because of this, the amount of judgements on the 8th of October was zero until 8 pm (see Appendix H, image 2). As we were satisfied with the current result, we decided to continue the job. The reason why we evaluated the job was because the job contained a large amount of data. The benefit of identifying issues soon is that it allows to re-upload the data. As there were no issues, we decided to continue the job. More information about the crowdsourcing process, such as the total cost and the number of trusted judgements are presented in Appendix G.

Setting	**Value**
Payment	0.02 cent per 10 tweets
Annotators per tweet	3
Data	6000 tweets
Guidelines	Version 3
All countries	
Test Questions:	20

Table 2: CrowdFlower Settings for Final Annotation Round

5 Results

5.1 Sentiment

All tweets have been annotated 3 times. Crowdflower always selected the label with the highest confidence score. Crowdflower uses the confidence score to represent the level of agreement. In Table 3, the results of the sentiment analysis per candidate is presented. The results show that most tweets were labeled as other. In Table 4, the Results per month are presented.

	Clinton	Trump	**Total**
Positive	339 11,3% %	429 14,3% %	768 12,8% %
Negative	1209 40,3% %	1243 41,3%	2452 40,8% %
Other	1452 48,4%	1328 44,3%	2780 46,3%
Total	3000	3000	6000

Table 3: Results per Candidate

		Positive	Negative	Other	Total
June	Clinton	216 14,4%	557 37,1	727 48,5	1500
	Trump	238 15,8%	593 39,5 %	725 44,6%	1500
August	Clinton	123 8,2 %	652 43,5%	725 48,3%	1500
	Trump	191 12,7%	650 43,3%	659 43,9	1500
	Total	786 12,8 %	2452 40,8%	2780 46,3%	6000

Table 4: Results per Month

Table 4 shows that sentiment has changed. For Clinton, the number of tweets that express a negative sentiment has increased by 6,4 percent. Furthermore, the number of tweets that express a negative sentiment towards Trump has increased with 3,8 percent. The number of tweets that express a positive sentiment has decreased with 3,1 percent. Noticeable is that there is a little change in the amount of tweets labeled as other. The number of tweets that express a neutral sentiment towards Clinton has only increased with 0,2 percent. Towards Trump, this percentage is increased with 0,2 percent.

5.2 Confidence Score

The confidence scores were obtained with Crowdflower. Crowdflower uses confidence scores to represent the level of agreement between annotators. The mean and other descriptive statistics measures of the confidence scores per tweet are visible in Table 5. The maximum number of agreement was 1.00. The lowest number of agreement was 0.3333. Table 5 shows that the mean of the confidence score was approximately 0.805.
The mean of June and August were almost identical. The mean of the confidence score for June was 0.806375 and the mean for August was 0.803278.

Metrics	Amount
N	6000
Mean	0.804826
Median	0.698300
Std. Deviation	0.1959979
Range	0.6667
Min	0.3333
Max	1.000

Table 5: Descriptive Statistics for confidence scores

6 Discussion

6.1 Difficult Cases and Issues

The mean of the confidence score might indicate that the task was doable for the annotators as it proves that annotators often agreed with each other. Furthermore, the distribution of the sentiments shows that the task was durable. However, when we look at the contributor satisfaction (Appendix G, Image 3),

then the overall task is rated with a score of 3.3 out of 5. Moreover, the ease of the task was rated with 3.2. In contrast with the confidence score, these ratings might indicate that the task was encountered as difficult.
The tweets with a low confidence score (confidence <0.4) are visible in Appendix H. This data set shows that there were difficulties with sarcastic tweets, such as the tweet: "Are you guys going to enjoy president trump already or what".
Furthermore, it seems that the data set included tweets that did not express a sentiment towards Clinton or Trump but towards another politician. An example of such a tweet is: "BarackObama funny how you have NOT used your 8 yr. position to better #black lives but you call realDonaldTrump unfit to serve America". The question was formulated in such a way that it only asked the annotators to label the sentiment towards Clinton or Trump. Because of this formulation, it might be the case that in cases in which there was no sentiment expressed towards Clinton or Trump, they did not know which label they had to chose.
Moreover, it seemed that there were difficulties with tweets that required background knowledge about the elections. We tended to solve this problem by mentioning context knowledge hashtags in the guidelines, such as #TrumpTrain and #ImWithHer. However, from the 339 tweets with a low confidence score 95 tweets contained such hashtags, including the ones we listed in the guidelines. Moreover, the most frequent hashtag in this sets were: #Trump2016 (4), #NeverHillary (4), #TrumpTrain (3), #Nevertrump (3). These were the hashtags we listed in the guidelines (see Appendix D). Two examples of these tweets were: "Anderson Cooper is a puss. #TrumpTrain" and "Cackling at the fact that more people think Trump is honest than HRC. Women even find Trump more honest. #ImWithHer". The fact that they had difficulties with these hashtags evokes two questions. Firstly whether this is due to either not reading the guidelines properly or to not understanding it. Secondly, whether this was due to lack of domain specific knowledge. Our first intention was to exclude all people from the United States and from Mexico, just as in round 3. Yet, after the annotation job was finished, we saw that they were included in the annotation. Because of this, it can be even possible that the high confidence score was obtained because there were a lot of Americans that participated in the task. This could also effect the distribution of the sentiment. Yet, it is rather difficult to answer these questions, because Crowdflower provides only quantitative results. Furthermore, if we were able to ask the annotators about their experience with the annotation task, then we would be able to provide more insights about the difficulties of the task itself.

6.2 Suggestions for Future Research

Further research would be needed to answer whether excluding Americans would lead to a lower confidence score. Furthermore, other differences in the design would be of interest for future research. For instance, future work could experiment with a higher payment. Besides, experiments could be conducted with manual sentiment analysis, but then on a different data set. Finally, the annotated data set could be used to evaluate the performance of programs that automatically assign the sentiment to tweets.

7 Conclusion

The aim of this research was to identify the issues which are encountered when the sentiment of a tweet is annotated manually. Although the quantitative results indicate that the agreement among annotators was high, the contributor satisfaction showed that the task was experienced as difficult. One very important issue is the design of the task. If the guidelines are not clear, then the annotators might have difficulties with assigning the label. Thus, it is important that the guidelines formulate how the researchers define the sentiment. In addition to the guidelines, the formulation of the question can be important as well. Another difficulty was how to deal with sarcasm. It seemed that the annotators encountered difficulties with assigning the sentiment to a sarcastic tweet. Finally, another issue was how to determine whether the difficulty of the task is influenced by domain specific knowledge.

References

Alvaro, N., Conway, M., Doan, S., Lofi, C., Overington, J., and Collier, N. (2015). Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of Biomedical Informatics*, 58:280–287.

Armstrong, C. L. and Gao, F. (2011). Gender, twitter and news content. *Journalism Studies*, 12(4):490–505.

de Winter, J., Kyriakidis, M., Dodou, D., and Happee, R. (2015). Using crowdflower to study the relationship between self-reported violations and traffic accidents. *Procedia Manufacturin*, 3:2518–2525.

Gurkhe, D. (2014). Effective sentiment analysis of social media datasets using naive bayesian classification. *International Journal of Computer Applications*, 99(13):1–4.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, pages 538–541. AAAI Press.

Medhata, W., Hassanb, A., and Korashyb, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093 – 1113.

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. O'Reilly Media: Sebastopol, CA.

Ringsquandl, M. and Petkovic, D. (2013). Analyzing political sentiment on twitter. *Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium*.

Sanghvi, A., Guruprasad, H., and Yogita, S. (2015). A study on sentiment analysis using tweeter data. *IJIRS - International Journal For Innovatie Research in Science & Technology*, 1(9):2349–6010.

Spencer, J. and Uchyigit, G. (2012). Sentimentor: Sentiment analysis on twitter data. In *In The 1st International Workshop on Sentiment Discovery from Affective Data*.

Staiano, J. and Guerini, M. (2014). Depechemood: a lexicon for emotion analysis from crowd annotated news. *Proceedings ACL*, 2:427 – 433.

Sylwester, K. and Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PLoS ONE*, 10(9):1 – 18.

Taddy, M. (2013). Measuring political sentiment on twitter: Factor optimal design for multinomial inverse regression. *Technometrics*, 55(5):415–425.

Appendix A – Dataset

```
s2477106@karora:~/Documents/Crowdflower$ cat getdata.sh
cat /net/corpora/creating_data/twitter-clinton-trump-2016/06/clinton-201606*:02.out | grep -v -i ".*clinton.*trump*" |
grep -v -i ".*trump.*clinton.*" | head -1500 > clinton06.txt
cat clinton06.txt | head -500 > file1_clinton06.csv
cat clinton06.txt | head -1000 | tail -500 > file2_clinton06.csv
cat clinton06.txt | tail -500 > file3_clinton06.csv

cat /net/corpora/creating_data/twitter-clinton-trump-2016/08/clinton-201608*:02.out | grep -v -i ".*clinton.*trump*" |
grep -v -i ".*trump.*clinton.*" | head -1500 > clinton08.txt
cat clinton08.txt | head -500 > file1_clinton08.csv
cat clinton08.txt | head -1000 | tail -500 > file2_clinton08.csv
cat clinton08.txt | tail -500 > file3_clinton08.csv

cat /net/corpora/creating_data/twitter-clinton-trump-2016/06/trump-201606*:02.out | grep -v -i ".*clinton.*trump*" | grep
-v -i ".*trump.*clinton.*" | head -1500 > trump06.txt
cat trump06.txt | head -500 > file1_trump06.csv
cat trump06.txt | head -1000 | tail -500 > file2_trump06.csv
cat trump06.txt | tail -500 > file3_trump06.csv

cat /net/corpora/creating_data/twitter-clinton-trump-2016/08/trump-201608*:02.out | grep -v -i ".*clinton.*trump*" | grep
-v -i ".*trump.*clinton.*" | head -1500 > trump08.txt
cat trump08.txt | head -500 > file1_trump08.csv
cat trump08.txt | head -1000 | tail -500 > file2_trump08.csv
cat trump08.txt | tail -500 > file3_trump08.csv
```

Image 1 – Shell script to extract sample

```
name,month,tweet_id,user_id,tweet
Clinton,06,737810654241947648,14529929,"2008 flashback: Clinton asks if Obama's so inevitable why can't he close the deal? URL"
Clinton,06,737807886332657664,1067120240,"3am phone call in the @HillaryClinton White House URL"
Clinton,06,737802711073050626,710971457203867650,"@AC360 @CNN Clinton LIED to benghazi families"
Clinton,06,737803507881771008,724458833586708480,"@AC360 @DrewGriffinCNN @ac360 Where's inside Clinton foundation"
Clinton,06,737808532452614145,2241717492,"@AC360 NO. if in doubt"
Clinton,06,737801876280770561,94283060,"@AC360 When will @HillaryClinton's lies matter to you & everyone else at @CNN?"
Clinton,06,737797581342834689,3911239745,"@AdamBensouda @JerryBrownGov @HillaryClinton @politico Yup. And my uninformed ass is def
initely casting a ballot on June 7th! xoxo"
Clinton,06,737810511904051200,1556525761,"@allen_clinton25 3-4 goals bruh not pages"
Clinton,06,737799866919878661,2390636869,"@Andrew4BW And I think Hillary Clinton to win the presidency."
Clinton,06,737802186667614208,66302096,"And why still no media investigations of quite interesting Clinton Fdn money trails? URL"
Clinton,06,737806992664756224,49464307,"Another Example Of #Maddow Using #HillaryClinton's Talking Points As Facts URL #CNN #NBC #
MSNBC URL"
Clinton,06,737796901853106176,68543202,"@BartMcCoy5 @benshapiro Sounds like .@benshapiro wants Clinton to win. He's proving himsel
f to be an irrelevant little snake. 🐍"
Clinton,06,737804258091892736,700512482318635009,"#berniesanders A few words on Clinton emails URL URL URL URL"
Clinton,06,737809954350891008,125128428,"Bill Clinton Pleading for Donations After 'Clinton Cash' Chases Donors Away URL"
Clinton,06,737805781416841216,4795853156,"@billclinton you never knew Daddy !! That is why you have been a serial philanderer!! A
philanderer is a punk. You Sir"
Clinton,06,737800328322646016,2706281318,"Bill for First Lady Funny Hillary Clinton Anti-Hillary URL"
Clinton,06,737797573067440128,3029152757,"Bravo. Now do the Clinton Foundation @brithume @SopanDeb @AP"
Clinton,06,737800164879044608,4735949221,"Breitbart News Updates. Cheryl Mills Testifies: Hillary Clinton’s Emails Were NOT Availa
ble For FOIA Requests URL"
Clinton,06,737806821340020737,635011613,"@BritainsBerning Bill Clinton was caught on an open mic telling Paul Ryan he wants to cut
 social security. The dem establishment is corrupt"
Clinton,06,737803470275678210,709107246827905024,"@BuybyFelicia @Benross75 @chuckwoolery @hale4jesus @nikkitur Reuters poll she's
up by 10 Rasmussen Clinton up NYT/CBS up by Clinton up 6"
Clinton,06,737804664595304448,275787081,"California Gov. Jerry Brown gives Clinton coveted nod as lead over surging Sanders slides
 URL URL"
Clinton,06,737809576997847040,278279975,"California primary: Clinton"
Clinton,06,737798474339516416,951824648,"Can lawyers interfere with a government investigation or are these just one of the many @
HillaryClinton perks? URL"
Clinton,06,737796066637156352,86796377,"Charles Ortel: Clinton Foundation Largest Charity Fraud Ever .... URL @MishGEA @Steen_Jako
bsen @MattSingh_ @DA_Stockman"
Clinton,06,737797266824564736,1223294035,"@chrislhayes @HillaryClinton @RalphNader Could you ask Nader if he thinks AL Gore and Bu
sh are the same? Wld like a thorough explanation."
Clinton,06,737801225312337920,131971803,"Clinton aide transcript released as part of email case URL"
Clinton,06,737801947370160128,154521153,"Clinton Ally Will Return Campaign Donations Amid a Federal Probe into Potential Ethics Vi
olations URL"
Clinton,06,737806170270896128,166350245,"Clinton Best ""Endorsement"" Yet #UAW💪#NJPrimary #PRPrimary & #CAPrimary was with BS N
OW WITH HC😎#VoteHillary ❤️URL"
Clinton,06,737801570193022976,28156710,"Clinton Camp Manager Silent on FBI Investigation Surrounding 'Bill Clinton's Best Friend'
URL #cnn #outfront #boston #la"
Clinton,06,737805605000204288,2858316326,"Clinton Contradicts IG Report That She Told Staffers Not To Talk About Her Email Account
 [VIDEO] URL"
Clinton,06,737806162775642113,724050768106500096,"#Clinton Do You Have News to Share? Get It Published. URL"
```

Image 2 – Final Dataset

```
import nltk
import mathplotlib

def main():
    text = open('tweets.txt').read()
    tokens = text.split(" ")
    stopwords = nltk.corpus.stopwords.words('english')
    new_tokens = [item for item in tokens if item.lower() not in
stopwords]
    freqdist = nltk.FreqDist(new_tokens)
    freqdist.plot(50)

main()
```

Image 3 – Python Script for Frequency Distribution

Appendix B – Guidelines Version 1

ANNOTATION GUIDELINES

OUR GOAL:
We want to use sentiment analysis in order to determine which of the two Presidential candidates (Hillary Clinton or Donald Trump) gets more negative tweets.

INSTRUCTIONS:

1. You have received a file which contains Tweets that mention the name Clinton or Trump.
2. Every line in this file is one Tweet.
3. The abbreviation 'UNK' does not belong to the Tweet. Your task is to change 'UNK' into the tag which represents the most suitable sentiment of a Tweet. Do this for every Tweet.
4. The sentiment of a Tweet can be:
    a. Positive **(POS)**
    b. Negative **(NEG)**
    c. Neutral **(MEH)**
5. Irrelevant (**IRR).** There might be tweets that do not refer to Donald Trump or Hillary Clinton but do contain the same names, these should be labeled as irrelevant (**IRR**).
6. If multiple users ('@username') are mentioned and Hillary Clinton or Donald Trump are also mentioned then the Tweets is considered a Trump/Clinton Tweet.
7. The table below helps you to determine the sentiment.

Sentiment	***Description***
Positive (Label: POS)	• Tweets that contain words referring to positive feelings about Hillary Clinton, words such as***: love, like, support, agree etc.*** • Tweets that contain hashtags referring to positive feelings about Hillary Clinton, hash tags such as ***#imwithher***. *(always consider hashtags in their context)* • Tweets that contain hash tags that refer to positive feelings about Donald Trump, hashtags such as: ***#makeamericagreatagain, #trumptrain, #alwaystrump***. *(always consider hashtags in their context)* • Tweets that copy a text about Hillary Clinton or Donald Trump from another source and positively support that text.
Negative (Label: NEG)	• Tweets that criticize, offend or make fun of Hillary Clinton or Donald Trump. • Tweets that contain words referring to negative feelings about Hillary Clinton or Donald Trump, words such as: ***hate, dislike, loathe, fuck, screw, sucks, idiot, stupid, etc.*** • Tweets that contain hashtags referring to negative feelings about Hillary Clinton, hash tags such as: ***#hillaryforprison***, ***#hillaryisacommiewhore, #killary, #neverhillary, #crookedhillary, #hillno***. *(always consider hashtags in their context)* • Tweets that contain hashtags referring to negative feelings about Donald Trump, hashtags such ***#nevertrump***. *(always consider hashtags in their context)* • Tweets that copy a text from another source about Hillary Clinton or Donald Trump from another source and negatively support that text.

Neutral (Label: MEH)	• Tweets that give an objective description about Hillary Clinton or Donald Trump. These tweets often do not contain words that express either support or critique, once these words are present as stated in the categories above, the tweet might belong to the POS or NEG category. An exception is when these words refer to other people using the words and the tweet refers back to those people. • Tweets that give an objective description about policy or stances that belong to Hillary Clinton or Donald Trump. These tweets often do not contain words that express either support or critique, once these words are present as stated in the categories above, the tweet might belong to the POS or NEG category. An exception is when these words refer to other people using the words and the tweet refers back to those people. • Tweets by news agencies often fall in this category. Beware though! News agencies such as FOX, NBC and other colored news agencies can sometimes express support or criticism of Hillary Clinton and Donald Trump.
Irrelevant (Label: IRR)	• Tweets that ONLY mention family members or other people with the same last name.

Appendix C – Guidelines version 2

ANNOTATION GUIDELINES

OUR GOAL:
We want to use sentiment analysis in order to determine which of the two Presidential candidates (Hillary Clinton or Donald Trump) gets more negative tweets.

INSTRUCTIONS:

1. You have received a file which contains Tweets that mention the name Clinton or Trump.
2. Every line in this file is one Tweet.
3. The abbreviation 'UNK' does not belong to the Tweet. Your task is to change 'UNK' into the tag which represents the most suitable sentiment of a Tweet. Do this for every Tweet.
4. The sentiment of a Tweet can be:
   A. Positive **(POS)**
   B. Negative **(NEG)**
   C. Others **(OTH)**
5. If multiple users ('@username') are mentioned and Hillary Clinton or Donald Trump are also mentioned then the Tweets is considered a Trump/Clinton Tweet.
6. The table below helps you to determine the sentiment.

Sentiment	***Description***
Positive (Label: POS)	• Tweets that contain words referring to positive feelings about Hillary Clinton, words such as***: love, like, support, agree etc.*** • Tweets that contain hashtags referring to positive feelings about Hillary Clinton, hash tags such as ***#imwithher***. *(always consider hashtags in their context)* • Tweets that contain hash tags that refer to positive feelings about Donald Trump, hashtags such as: ***#makeamericagreatagain, #trumptrain, #alwaystrump***. *(always consider hashtags in their context)* • Tweets that copy a text about Hillary Clinton or Donald Trump from another source and positively support that text.
Negative (Label: NEG)	• Tweets that criticize, offend or make fun of Hillary Clinton or Donald Trump. • Tweets that contain words referring to negative feelings about Hillary Clinton or Donald Trump, words such as: ***hate, dislike, loathe, fuck, screw, sucks, idiot, stupid, etc.*** • Tweets that contain hashtags referring to negative feelings about Hillary Clinton, hash tags such as: ***#hillaryforprison***, ***#hillaryisacommiewhore, #killary, #neverhillary, #crookedhillary, #hillno***. *(always consider hashtags in their context)* • Tweets that contain hashtags referring to negative feelings about Donald Trump, hashtags such ***#nevertrump***. *(always consider hashtags in their context)* • Tweets that copy a text from another source about Hillary Clinton or Donald Trump from another source and negatively support that text.

Others (Label: OTH)	• Tweets that give an objective description about Hillary Clinton or Donald Trump. These tweets often do not contain words that express either support or critique, once these words are present as stated in the categories above, the tweet might belong to the POS or NEG category. An exception is when these words refer to other people using the words and the tweet refers back to those people. • Tweets that give an objective description about policy or stances that belong to Hillary Clinton or Donald Trump. These tweets often do not contain words that express either support or critique, once these words are present as stated in the categories above, the tweet might belong to the POS or NEG category. An exception is when these words refer to other people using the words and the tweet refers back to those people. • Tweets by news agencies often fall in this category. Beware though! News agencies such as FOX, NBC and other colored news agencies can sometimes express support or criticism of Hillary Clinton and Donald Trump. • Tweets about family members that have the name Clinton or Trump. • Tweets that do not express any form of support of critique about Donald Trump or Hillary Clinton. • All Tweets that do not belong to either the Positive or Negative label.

Appendix D – Guidelines version 3

ANNOTATION GUIDELINES

INSTRUCTIONS:

1. You see in front of you ten tweets that mention the name Clinton or Trump.
2. Your task is to analyse the sentiment expressed in each individual tweet. Do this for every Tweet.
3. The sentiment of a Tweet can be:
   A.Positive
   B. Negative
   C.Other
4. If multiple users ('@username') are mentioned and Hillary Clinton or Donald Trump are also mentioned then the Tweets is considered a Trump/Clinton Tweet.
5. The table below helps you to determine the sentiment.

DEFINITION OF LABELS

Sentiment	***Description***
Positive (Label: POS)	• Tweets that contain words referring to positive feelings about Hillary Clinton, words such as***: love, like, support, agree etc.*** • Tweets that contain hashtags referring to positive feelings about Hillary Clinton, hash tags such as ***#imwithher***. *(always consider hashtags in their context)* • Tweets that contain hash tags that refer to positive feelings about Donald Trump, hashtags such as: ***#makeamericagreatagain, #trumptrain, #alwaystrump***. *(always consider hashtags in their context)*
Negative (Label: NEG)	• Tweets that criticize, offend or make fun of Hillary Clinton or Donald Trump. • Tweets that contain words referring to negative feelings about Hillary Clinton or Donald Trump, words such as: ***hate, dislike, loathe, fuck, screw, sucks, idiot, stupid, etc.*** • Tweets that contain hashtags referring to negative feelings about Hillary Clinton, hash tags such as: ***#hillaryforprison***, ***#hillaryisacommiewhore, #killary, #neverhillary, #crookedhillary, #hillno***. *(always consider hashtags in their context)* • Tweets that contain hashtags referring to negative feelings about Donald Trump, hashtags such ***#nevertrump***. *(always consider hashtags in their context)*
Others **(Label: OTH)**	• All Tweets that do not belong to either the Positive or Negative label. • All Tweets that mention both candidates (Trump & Clinton).

Appendix E – Agreement Scores Annotation Round 1

A	B	C	D	E	F	G	H	I
	B Positive	B Neutral	B Negative	B Irrelevant	amount per label	(the amount per label / the total amount (50))		
A Positive					0	0		
A Neutral	5	14	11	8	38	0,76		
A Negative			4		4	0,08		
A Irrelevant		1	1	6	8	0,16		
amount per label	5	15	16	14				
(the amount per	0,1	0,3	0,32	0,28				
A_0	(14 + 4 + 6) / 50 = 0,48							
A (Positive) x B (Positive) = 0 x 0,1 = 0								
A (Neutral) x B (Neutral) = 0,76 x 0,3 = 0,228								
A (Negative) x B (Negative) = 0,08 x 0,32 = 0,0256								
A (Irrelevant) x B (Irrelevant) = 0,16 x 0,28 = 0,0448								
A_e	0 + 0,228 + 0,0256 + 0,0448 = 0,2984							
Cohen's k	(0,48 - 0,2984)/(1-0,2984) = 0.25883694412							

Image 1 – Agreement Scores

Appendix F – Script

```
import sys
import codecs

lines = [line.strip() for line in codecs.open(sys.argv[1], encoding="utf-8").readlines()]
for l in lines:
    fields = l.split(",",4)
    tweet_id = fields[1]
    user_id = fields[2]
    tweet = fields[3].replace('"','""')
    print("{},{},\"{}\"".format(tweet_id,user_id,tweet))
```

Image 1 – The Python script used to modify the data

Appendix G – Final Results Crowdflower

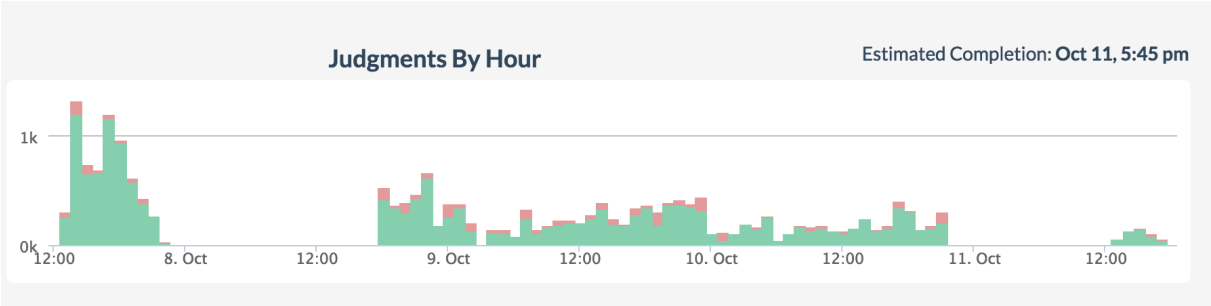

Image 1 – Judgements by hour

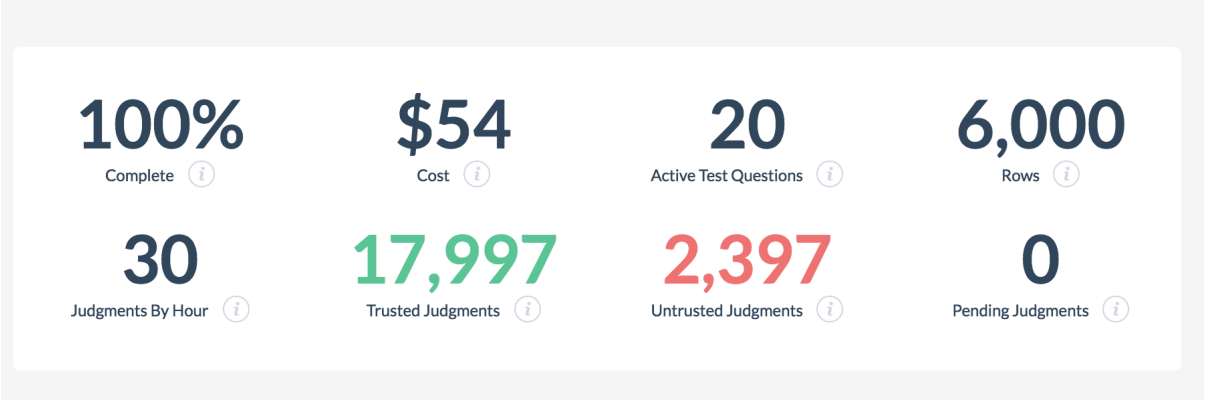

Image 2 – Final Results

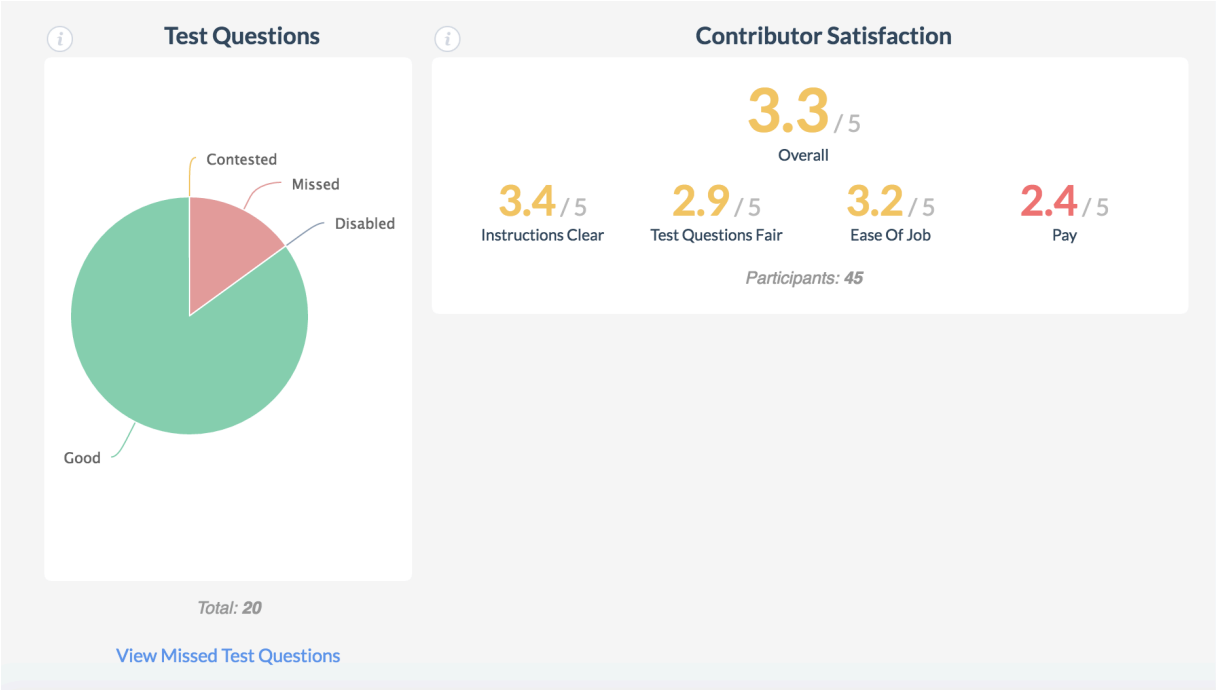

Image 3 – Contributor Satisfaction