Sentiment Analysis

Sa Adat Azam Saniat

1. Set Up

We first collect the data for the data sets "Weather Sentiment - AMT," "Sentiment popularity – AMT," and "Crowdsourced Amazon Sentiment." After transforming the raw data into CSV files, we use the pandas library to transform each data set into a data frame. We get these four data frames:

1. weather_sentiment_df from "Weather Sentiment - AMT"
2. sentiment_popularity_df from "Sentiment popularity – AMT"
3. is_book_df from "Crowdsourced Amazon Sentiment"
4. is_negative_df from "Crowdsourced Amazon Sentiment"

The data is retrieved and formatted using CrowdData

1.1 Overview of the Data Sets

- The Weather Sentiment data set consists of tasks that falls under multi-class classification tasks. Here the responses are encoded as followed: negative (0), neutral (1), positive (2), tweet not related to weather (3) and can't tell (4).
- The Sentiment Popularity data set is a binary-classification, meaning the responses were either negative (0), positive (1) or just "not sure".
- The Crowdsourced Amazon Sentiment data set is also a binary-classification which follows the same response encoding as mentioned above. However, the data set is split in two as there are two predicates. The two subtasks were to label if the review is on a book or not (isBook) and to label if the review was negative (isNegative).

1.2 Reader's Notes

Throughout the paper, the terms "prediction accuracy" and "consensus confidence" have been used interchangeably.

2. Initial Consensus Method

The consensus method is relatively simple. We review each unique 'taskID,' find the most common response, and append that in the 'consensus' column. We then calculate the consensus confidence:

$$consensusConfidence = \frac{totalNumberOfConsensusResponses}{totalNumberOfResponses}$$

We then add this in another column. We also keep track of the total number of workers for each 'taskID' as we will require it later.

3. Performance of Predictions

Analysis of the accuracy of the predicted labels with respect to three factors.

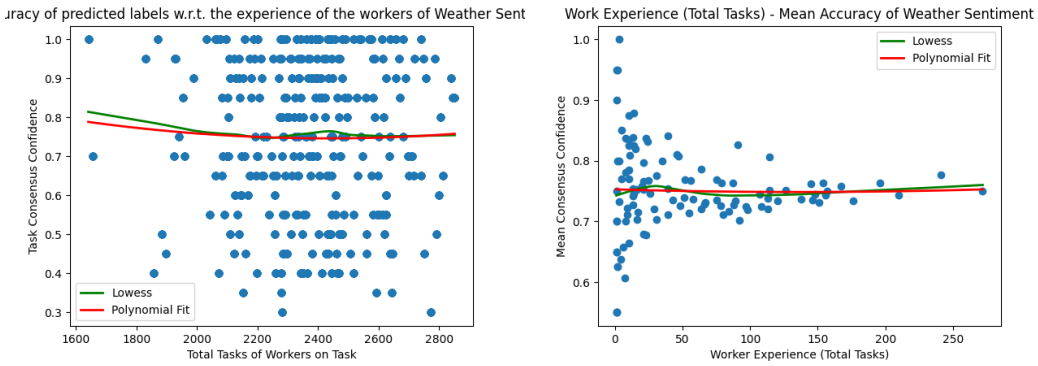
1. Total experience of the workers that worked on each task as available from the dataset
2. The threshold of the consensus (0.5 by default)
3. Time spent on each task

3.1 Experience of Workers

First Method: To set the analysis up, we first created a dataset of unique 'workerID's and caculated the 'totalTasks' they've completed. We then use this data to make a scatter plot with the best-fit LOESS graph. The horizontal axis represents the sum of total amount of tasks each worker that worked on a task. The vertical axis represents the consensus confidence of the respective task.

Second Method: We also ran another worker oriented analysis where we make a scatter plot each worker's experience level (calculated in total tasks completed) and the mean consensus confidence of that worker.

3.1.1 Weather Sentiment



(a) Method 1 (b) Method 2

Figure 1: Method 1 and Method 2 applied on analyzing the correlation between the accuracy of the predicted consensus with respect to the experience of the workers in the weather sentiment data set.

For the weather sentiment data set, we see here in Figure 1 (a) the scatter plot reveals a slight decrease in consensus confidence with an increase in the sum of total tasks at first, but then the trend flattens out. The best fit LOESS line and polynomial fit demonstrate a marginal decline in consensus confidence initially, with only a minimal reduction over time. This pattern could suggest that worker experience has a limited impact on consensus confidence for tasks, indicating that other factors may influence the accuracy of predictions.

In Figure 1 (b) the best fit lines in the scatter plot are relatively flat as well, indicating that worker experience has minimal impact on mean task accuracy. It is possible that other factors, such as training, feedback, or inherent skills, may play a more critical role in determining worker accuracy. Further research may be necessary to explore the factors that influence task accuracy and the degree to which worker experience contributes to overall accuracy for weather sentiment tasks.

3.1.2 Sentiment Popularity

(a) Method 1 (b) Method 2

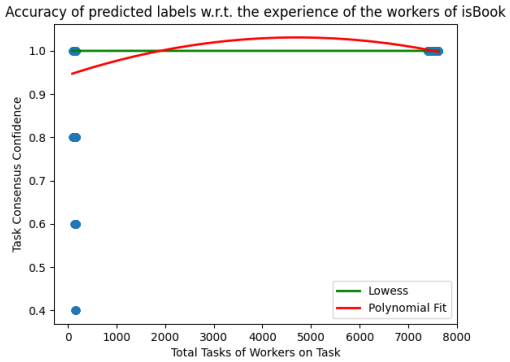
Figure 2: Method 1 and Method 2 applied on analyzing the correlation between the accuracy of the predicted consensus with respect to the experience of the workers in the sentiment popularity data set.

From Figure 2 (a) the best fit LOESS line on this graph shows a flat U-shape with most of the points concentrated around 0.9 to 1.0 task consensus confidence, indicating that they were mostly accurate. The polynomial fit is a bit more intense U-shaped, but still shows a similar trend. This suggests that the more experienced the worker, the more likely they are to accurately predict the task's label.

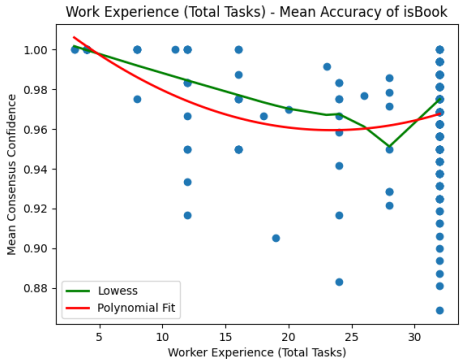
In Figure 2 (b) the best fit lines on this graph are super flat, almost forming an inverse U-shape, but still very flat. Most of the points are concentrated in the first quarter of the x-axis and then exponentially decrease, with only a few sparse points from the right half of the x-axis. This indicates that workers with fewer tasks tend to have higher accuracy, but this relationship flattens out as the worker gains more experience. Overall, these findings suggest that task consensus confidence and worker accuracy are affected by various factors, such as worker experience and task complexity, and that further investigation is needed to fully understand these relationships.

3.1.3 Crowdsourced Amazon Sentiment

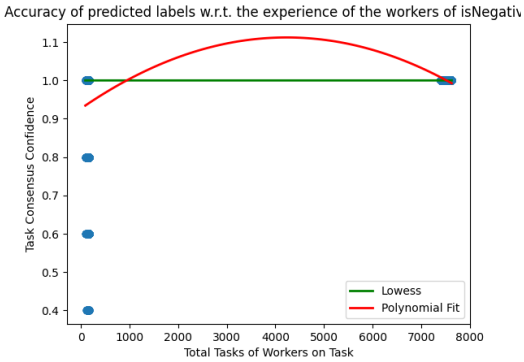
Figure 3's (a) and (c) are very similar because the subtasks 'isBook' and 'isNegative' are done by the same workers for each task. On the other hand, the task's consensus confidence



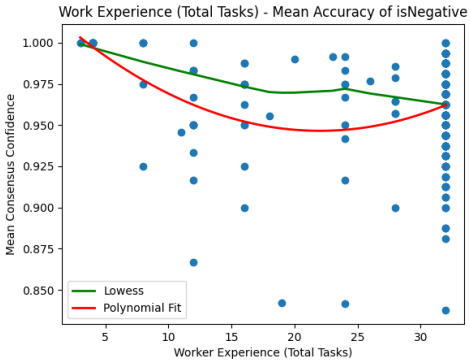
(a) Method 1 on isBook



(b) Method 2 on isBook



(c) Method 1 on isNegative



(d) Method 2 on isNegative

Figure 3: Method 1 and Method 2 applied on analyzing the correlation between the accuracy of the predicted consensus with respect to the experience of the workers in the crowdsourced amazon sentiment data set.

are very skewed towards 1.0 for the whole data set.

Figure 3's (b) and (d) shows a very clear negative relationship between a workers' experience level and their mean task accuracy. This is most probably because there are more workers that have better experience in this data set compared to the lesser experienced workers. This could have led to bringing the average of their aggregate mean accuracy to be brought down.

The findings are inconclusive for this data set.

3.2 Varying Consensus Thresholds

Here we look at the percentage of tasks that we can consider to be accurately labeled given our consensus with varying threshold levels.

3.2.1 Weather Sentiment

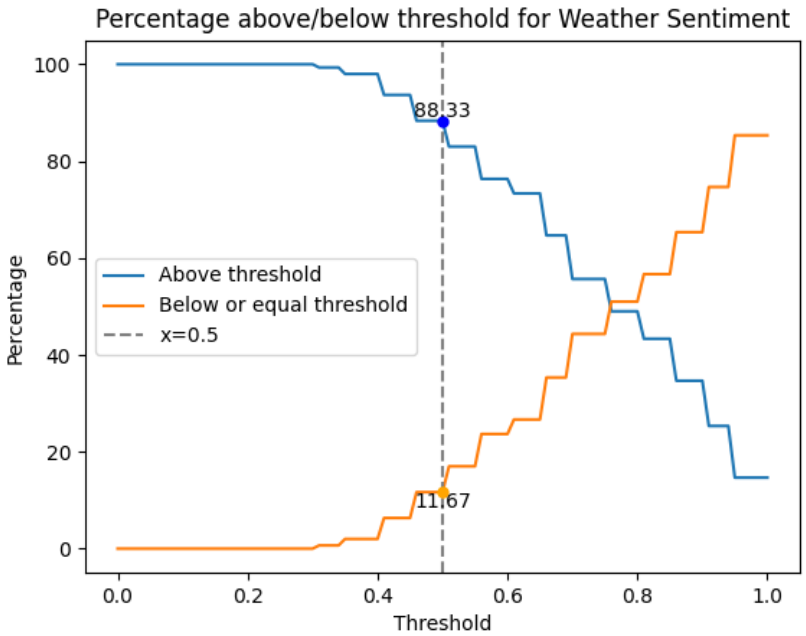


Figure 4: The percentage of tasks that are above and below varying threshold limits of consensus confidence or prediction accuracy of the tasks in the weather sentiment data set.

There is an 88.33% percent accuracy for this data set when the threshold was set to 50% (Figure 4). The workers' correctness was great, and the predictions were acceptable. However, there is a sharp decline in the amount of tasks that were accurate when the threshold was raised.

3.2.2 Sentiment Popularity

There was a 100% accuracy when at the default threshold of 50% for this data set (Figure 5). When the threshold was raised the accuracy only fell by small decrements – indicating the predictions for this data set was overall very accurate.

3.2.3 Crowdsourced Amazon Sentiment

The predictions for this task came out to be very accurate overall. At a threshold of 100% the accuracy was still 80% [Figure 6 (a) and (b)]

3.3 Time Spent on Each Task

This is the analysis on the average time spent on a task's correlation to the task's consensus confidence. The data was only available for the Weather Sentiment and the Popularity Sentiment data sets.

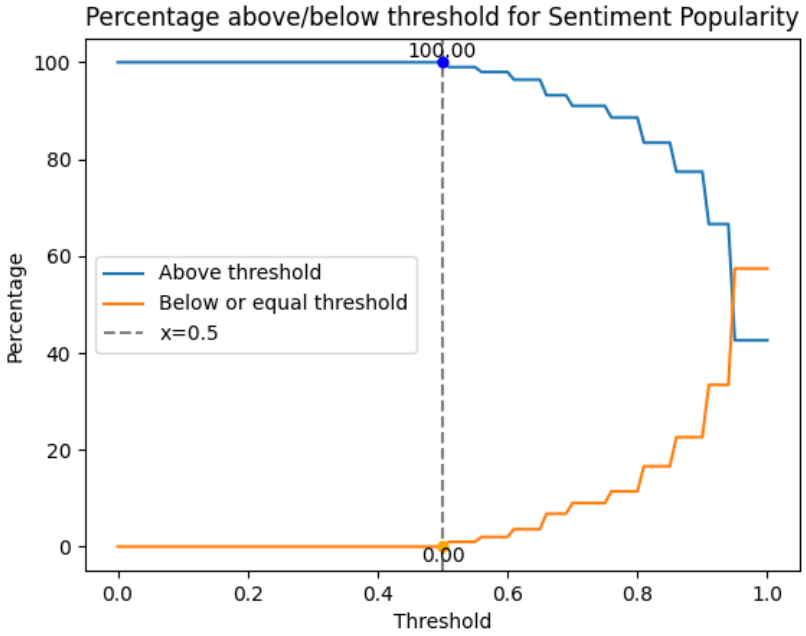
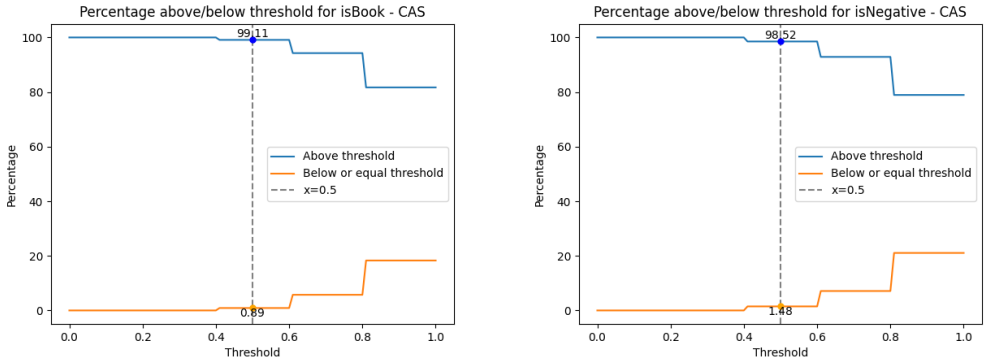


Figure 5: The percentage of tasks that are above and below varying threshold limits of consensus confidence or prediction accuracy of the tasks in the sentiment popularity data set.



(a) isBook (b) isNegative

Figure 6: The percentage of tasks that are above and below varying threshold limits of consensus confidence or prediction accuracy of the tasks in the isBook and isNegative data sets from Crowdsourced Amazon Sentiment.

3.3.1 Weather Sentiment

From Figure 7, the graph reveals a U-shaped relationship where most of the points are on the left side. The relationship appears to be negative until approximately 70 seconds, after

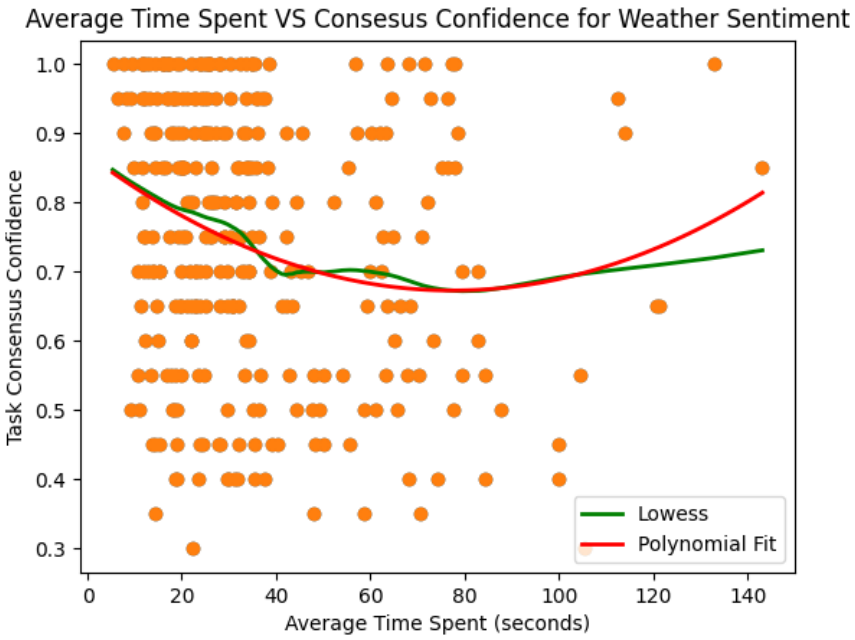


Figure 7: Correlation between average time spent per task and the task's prediction accuracy on the Weather Sentiment data set.

which it rises again.

One possible explanation for this relationship is that the workers may be rushing through the tasks in the beginning, which could lead to lower consensus confidence. As they spend more time on the task, they may be able to give more thorough and accurate responses, resulting in an increase in consensus confidence. However, if they spend too much time on the task, they may start to second-guess their responses and become fatigued, leading to a decrease in consensus confidence.

Another possible explanation could be that the difficulty of the task may vary, and some tasks may require more time and effort to complete accurately. Workers may be spending more time on the more difficult tasks, resulting in higher consensus confidence for those tasks.

In conclusion, the U-shaped relationship between the average time spent per task and the task's consensus confidence suggests that the time spent on a task may be a crucial factor in the accuracy of the workers' responses. It is important to further investigate the underlying factors contributing to this relationship, such as task difficulty and worker fatigue, to improve the overall quality of the responses.

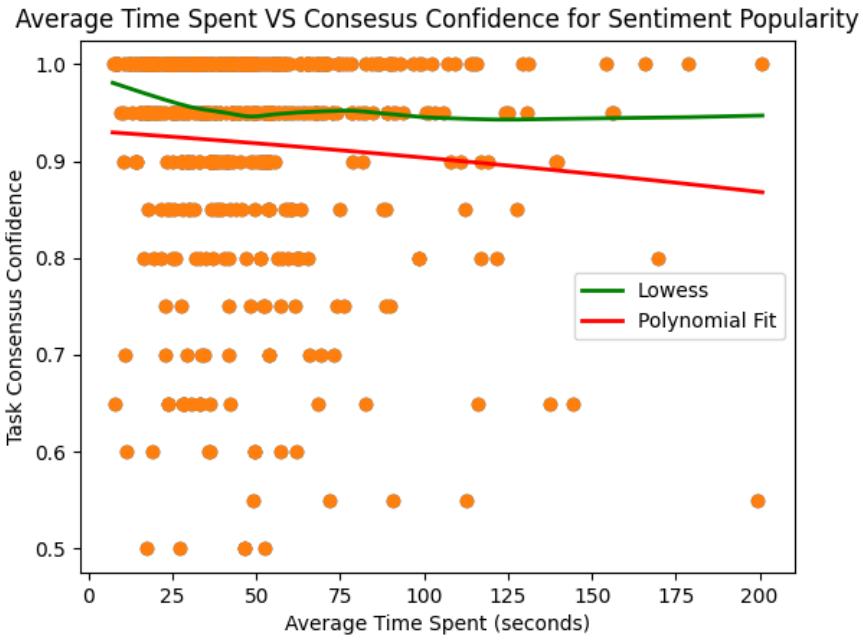


Figure 8: Correlation between average time spent per task and the task's prediction accuracy on the Sentiment Popularity data set.

3.3.2 Sentiment Popularity

The scatter plot in Figure 8 showed a negative relationship, with lower average times spent associated with higher consensus confidence. This relationship was however not very statistically significant.

These findings may be explained by several factors. Firstly, it's possible that workers who spend less time on a task are more confident in their initial impressions and less likely to be influenced by subsequent responses from other workers. Additionally, workers who spend less time on a task may be more selective in which tasks they choose to complete, preferring those that align with their expertise and experience. Finally, it's possible that workers who spend less time on a task are more motivated or efficient, completing tasks more quickly and with greater accuracy.

Overall, these findings suggest that time spent on a task may not always be indicative of the quality of work performed. Rather, workers who are able to quickly and accurately complete tasks may be highly valuable in crowdsourcing environments, and strategies for identifying and incentivizing these workers may be beneficial for ensuring high-quality results.

4. Findings

Looking at all the results from each of the data sets we can conclude a few things. For all three different tasks, it can be said that the total experience of all the workers do not influence the the prediction accuracy of the tasks – so tasks designers should not consider this as a factor when selecting workers. However, it can depend on the nature of the task and that would require analyzing much more datasets from a broader genre sample.

Moreover, it was seen that an individual workers' experience levels also did not give us conclusive information as the trend seemed to be mostly flat, or skewed due to nonuniform distribution.

The acceptability of our predictions seemed to perform very well when the consensus threshold was set to the default of 50% – with an average acceptable prediction percentage of 96.48%. However, the trend when the threshold was increased was mostly a sharp decline in the acceptable prediction percentage, where in Weather Sentiment it went as low as less than 20% when the threshold was set to 100%. On the other hand though the two predicates for Crowdsourced Amazon Sentiments performed really well and the the acceptable percentage was approximately 80% even when the threshold was set to 100%. Conclusively, task designers can set the threshold depending on how accurate they want the labels to be, consequently, how much they are willing to spend.

When comparing the difference between what we see in the prediction accuracy versus the average time spent on task graphs for the Weather Sentiment and Sentiment Popularity data sets – we see a two different trends. Weather Sentiment had an U-shaped curve while the Sentiment Popularity showed a relatively flat correlation. Both outcomes suggest different explanations but this maybe due to the what each task requires from the worker. Weather Sentiment workers had to work with 4 options while Sentiment Popularity workers only worked with 2. For multi class classifications like Weather Sentiment it could be true that the worker's reading efficiency may play a role as we can see that lower average times equate to a better task accuracy. This maybe also due to the length of the tweets. A big limitation is that we are using average time taken, as a result, outliers may skew the mean. If we truncate the average times after 90 seconds with the sparse the outliers, we'll come to see that both the curves tend to agree that more time taken equates to a poorer accuracy – hence worker acquisition teams should audit the workers for their reading and comprehension efficiency.

5. Results from other Data Sets

5.1 Blue Bird

The task of this data set is about identifying images that contains blue birds. Due to all the workers having done exactly 108 tasks exactly it was not informative to analyze the accuracy with respect to the worker's experiences. Time taken for the tasks were not provided for this data set.

5.1.1 Performance of Predictions on Varying Consensus Thresholds

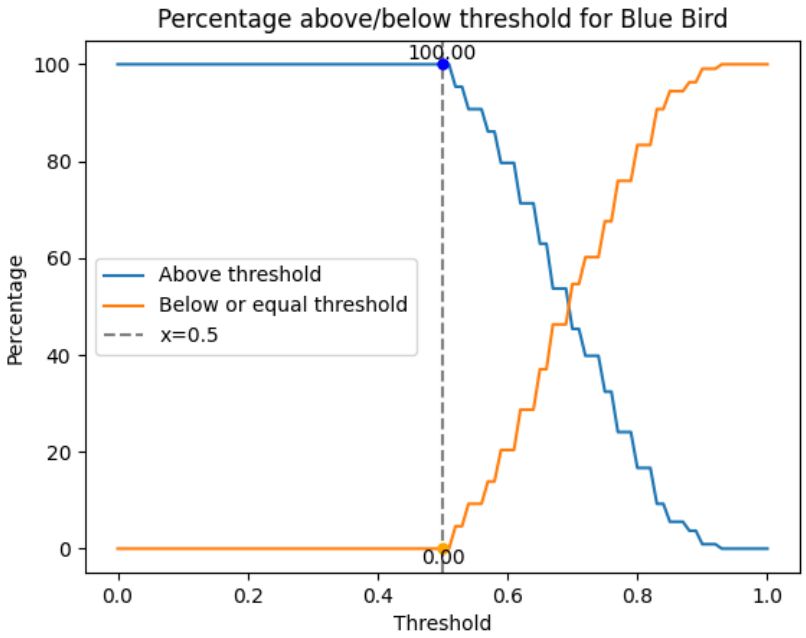


Figure 9: The percentage of tasks that are above and below varying threshold limits of consensus confidence or prediction accuracy of the tasks in the blue bird data set.

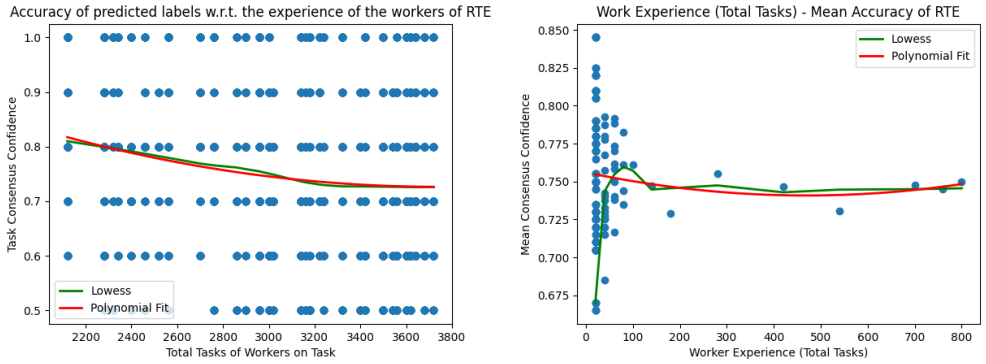
There is an 100% percent accuracy for this data set when the threshold was set to 50% (Figure 9). The workers' correctness was great, and the predictions were acceptable. However, there is a sharp decline in the amount of tasks that were accurate when the threshold was raised.

5.2 Recognizing Textual Entailment

This dataset contains the individual worker judgments and the related ground truths about identifying whether a given Hypothesis sentence is implied by the information in the given text. Time spent data was not provided.

5.2.1 Performance of Predictions on Experience of Workers

In Figure 10 (a) the best fit lines for both LOESS and polynomial regression had a slightly negative gradient. This finding suggests that as the sum of total tasks for workers increases, their accuracy in providing consensus confidence for a given task decreases. One possible explanation for this negative relationship could be that workers with higher levels of experience may become overconfident in their abilities, leading them to make more errors. Additionally, workers with more experience may have more diverse backgrounds, leading to a greater degree of subjectivity in their assessments.



(a) Method 1 (b) Method 2

Figure 10: Method 1 and Method 2 applied on analyzing the correlation between the accuracy of the predicted consensus with respect to the experience of the workers in the recognizing textual entailment data set.

On the other hand, in method two [Figure 10 (b)], we found a clear positive correlation between the number of tasks completed by a worker and their mean task accuracy. Granted, we don't count the outliers after 0-100 tasks. This trend suggests that workers may gain valuable experience and knowledge as they complete more tasks, allowing them to perform more accurately and efficiently over time. This result has important implications for task design and worker training, as it suggests that workers' experience level is a critical factor to consider in optimizing task performance and overall quality.

5.2.2 Performance of Predictions on Varying Consensus Thresholds

There is an 100% percent accuracy for this data set when the threshold was set to 50% (Figure 11). It decrements right after when the threshold is increased.

6. About Code

The whole work folder has been zipped containing all the figures and data. The main executable is:

```
work.py
```

Please see next page for Figure 11.

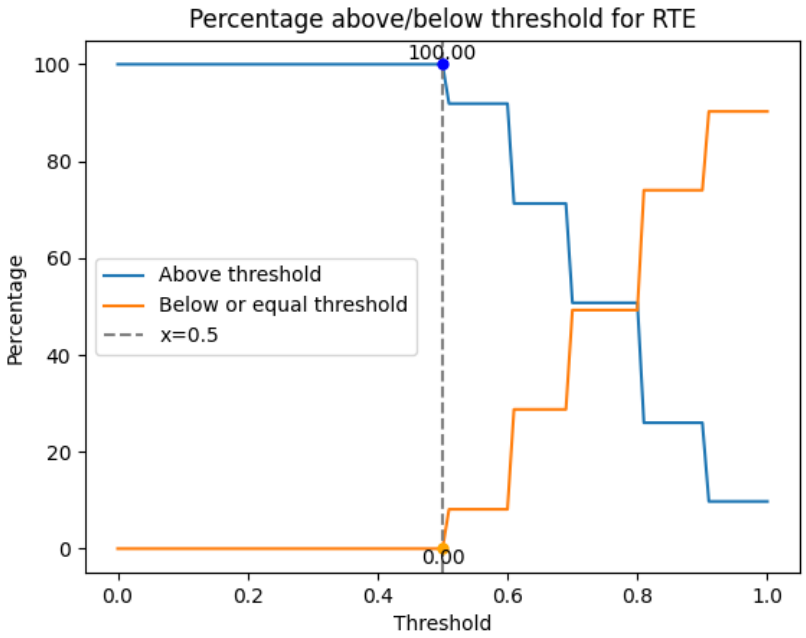


Figure 11: The percentage of tasks that are above and below varying threshold limits of consensus confidence or prediction accuracy of the tasks in the recognizing textual entailment data set.