Inter-Rater Reliability for NLP Machine Learning Model Training

Data Mine Students: Josiah Gottfried, Rishabh Kulkarni, Anish Thangavelu

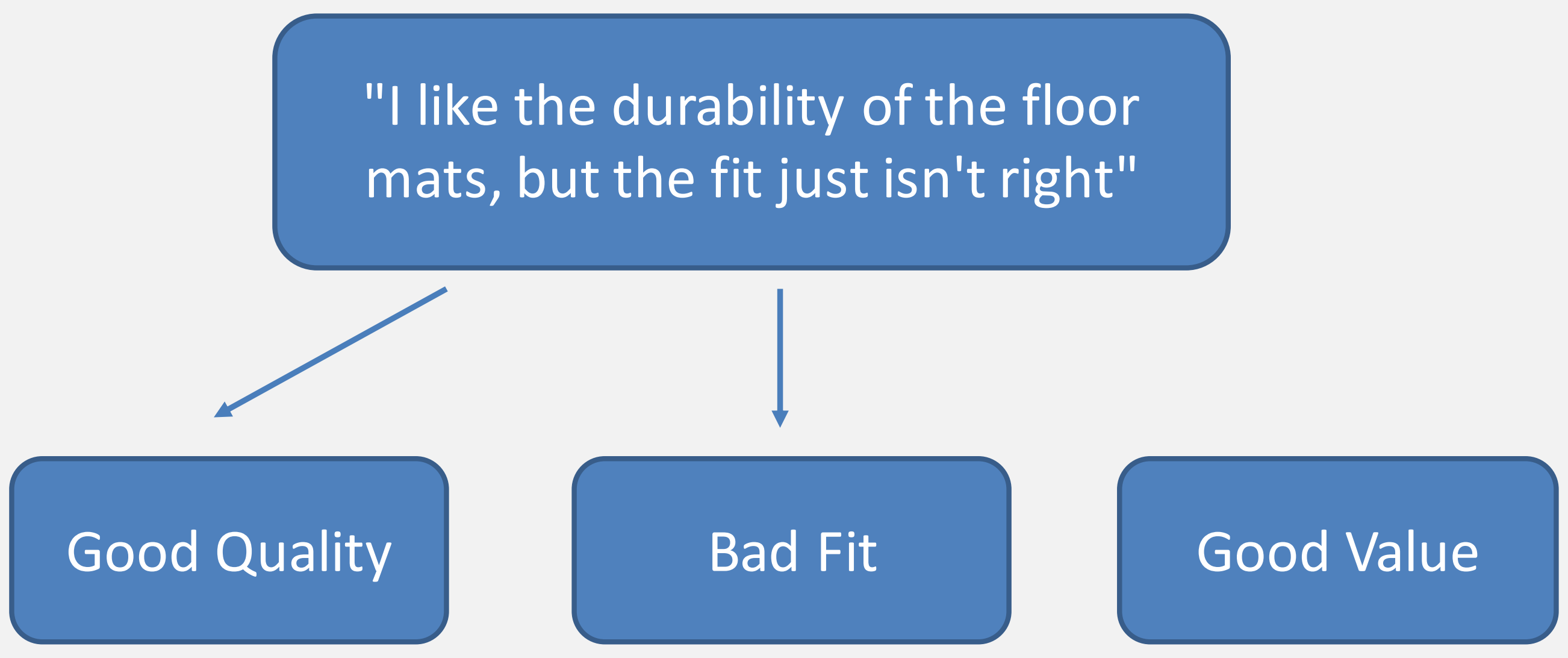
Data Mine Teaching Assistant: Min Lu

Data Science Capstone Students: Ben Moorman, Akanksha Bathini, Bharat Iyer, Arpit Dhawan, Eesha Deepak, Khushi Shah, Annie Wong

Project Motivation

- Customer feedback can reveal valuable information about a product
- How can an NLP model reliably analyze 100,000 customer reviews?
  - Training set quality is crucial (garbage in, garbage out)
  - Label definitions should be objective
- Our task: create a system for annotating a reliable training set
  - We worked with Amazon reviews on car floor mats
  - Manually annotated 5000 customer reviews
  - Measured agreement between reviewers with inter-rater reliability (IRR) coefficient

Figure 1. Annotation example

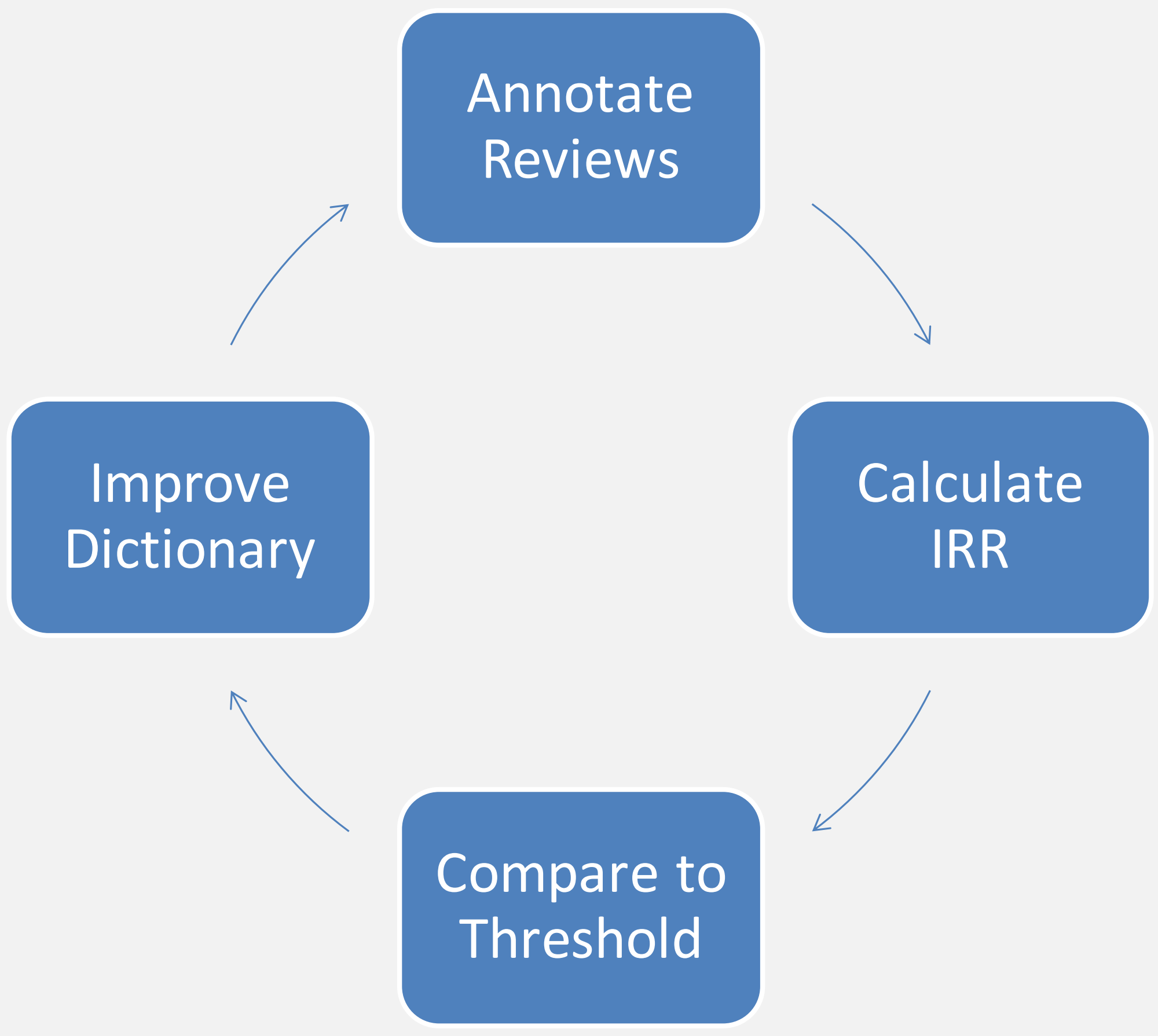

"I like the durability of the floor mats, but the fit just isn't right"
Good Quality
Bad Fit
Good Value

IRR Process

IRR Process Flow:

- Create a labelling dictionary
  - Decided on six labels and initialized definitions
- Annotate Reviews
  - Utilized the dictionary to make labeling decisions
- Calculate IRR
  - Processed labels using a Krippendorff's Alpha script to calculate IRR
- Compare to Threshold
  - Compared the calculated IRR to our minimum threshold of 0.7
  - If the scores were under the threshold, we would perform another cycle to improve them
- Improve Dictionary
  - Analyzed reviews with highest annotation variance
  - Clarified dictionary definitions to handle these cases

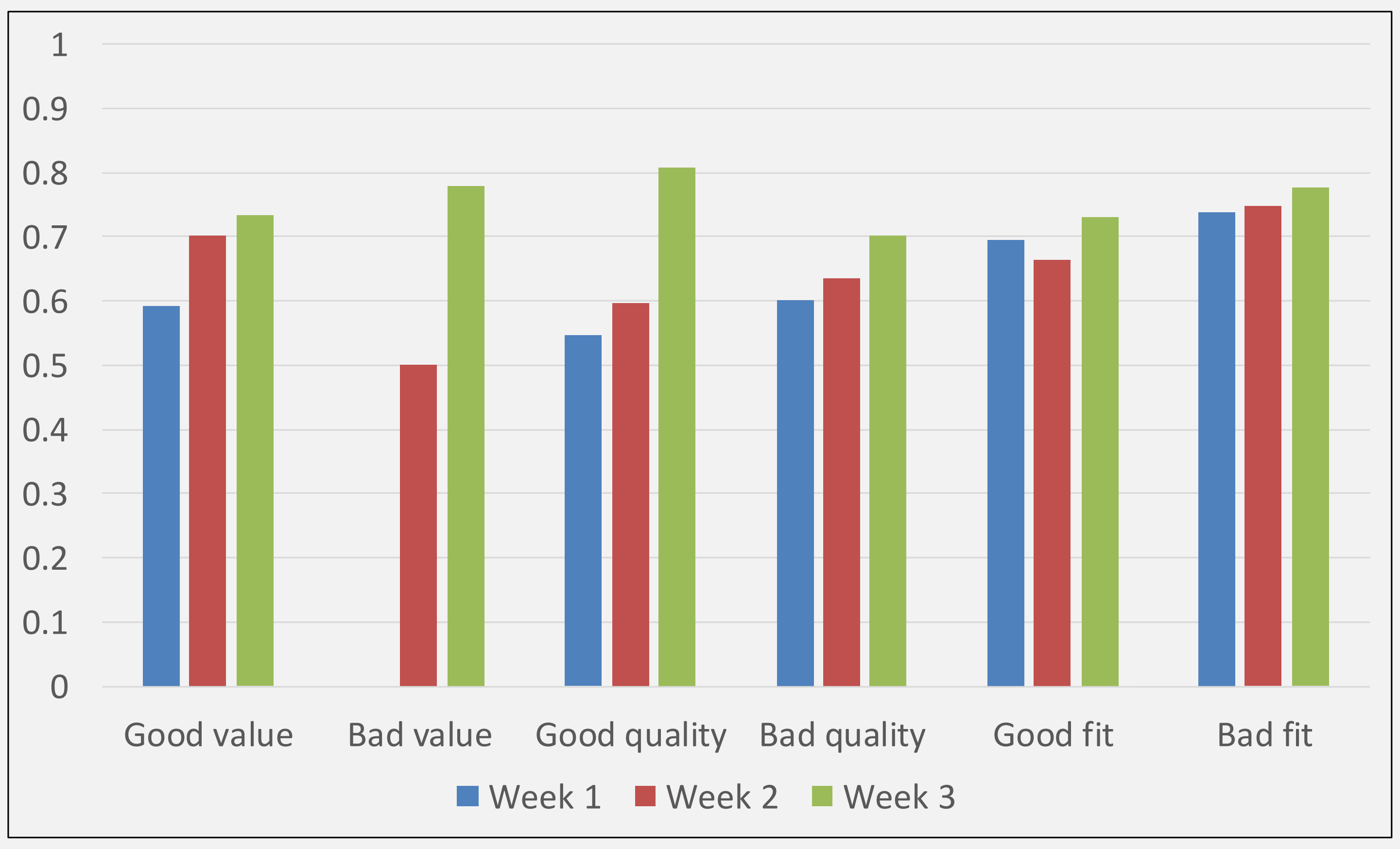
Figure 2. IRR process flow



Labelling Results

- Krippendorff's alpha represents the difference between actual agreement and expected agreement
  - Alpha is given by $\alpha = 1 - \frac{D_{observed}}{D_{expected}}$
  - $D_{observed}$ is actual disagreement and $D_{expected}$ is expected disagreement
- Results analysis
  - Alpha coefficient began at an average of 0.63 in the first week
  - As the dictionary became clearer, the coefficient increased
  - Alpha coefficient had reached an average of 0.75 by the third week
- Takeaways
  - Clarifying the dictionary successfully improved reviewer agreement
  - Due to poor choice in original labels, subjectivity made it difficult to achieve a high alpha coefficient

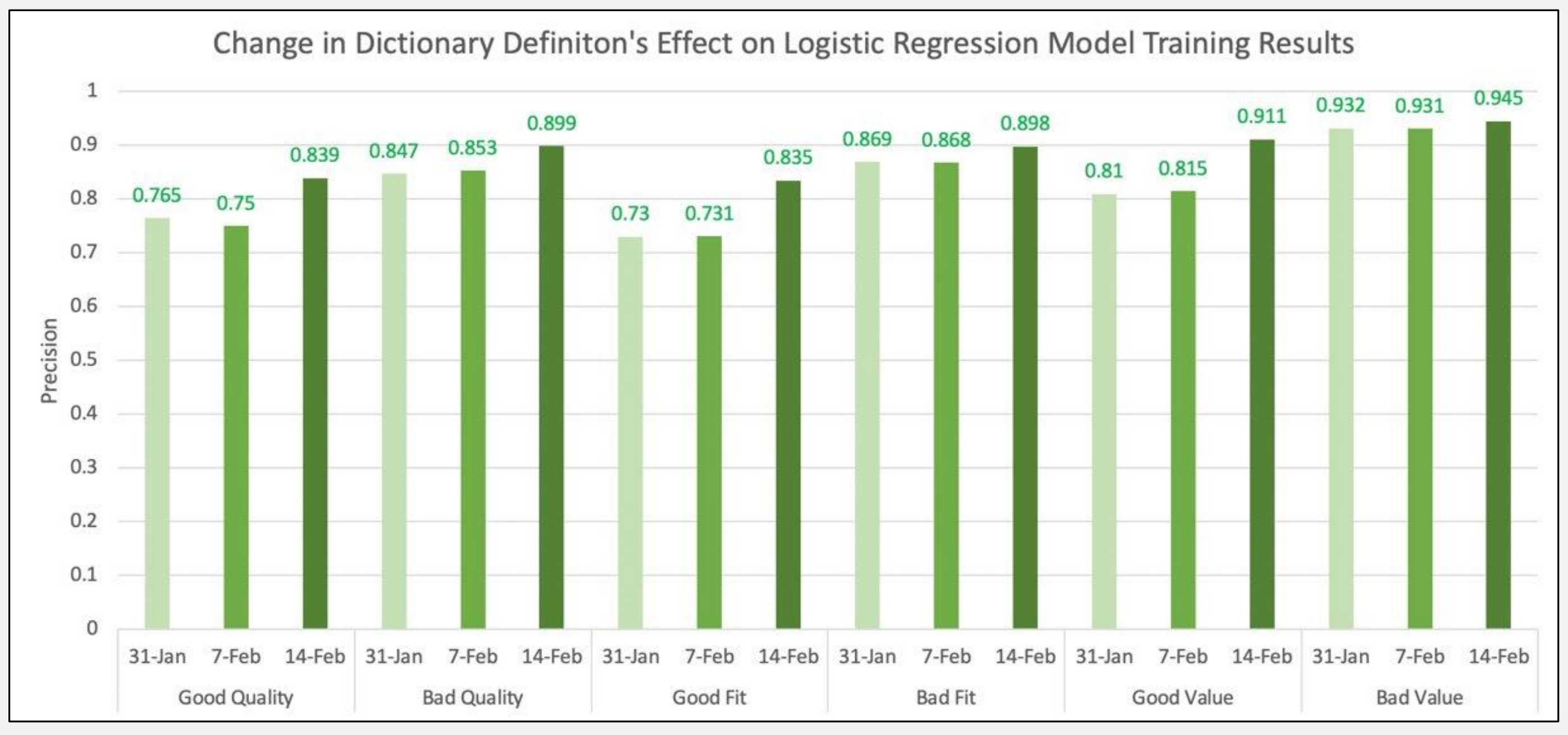
Figure 3. Krippendorff's alpha coefficient



Natural Language Processing Machine Learning Model

- **Hypothesis:** As our annotation dictionary improved, we expect to see improvements in the predictive model performance.
- **Conclusion:** We see that there's some improvement in the performance of predictive models as our IRR score improves. However, as we were unable to achieve our target IRR score of 0.8 in the end, we are unable to conclude that the improvement is statistically significant.
- **Future Direction:** Experiment further with fine tuning the annotation dictionary until an IRR score of above 0.8 is achieved and test for statistical significance in the changes of model performance.

Figure 4. Performance of logistic regression model over time



Acknowledgements

Corporate Partner Mentor: Blair Winograd