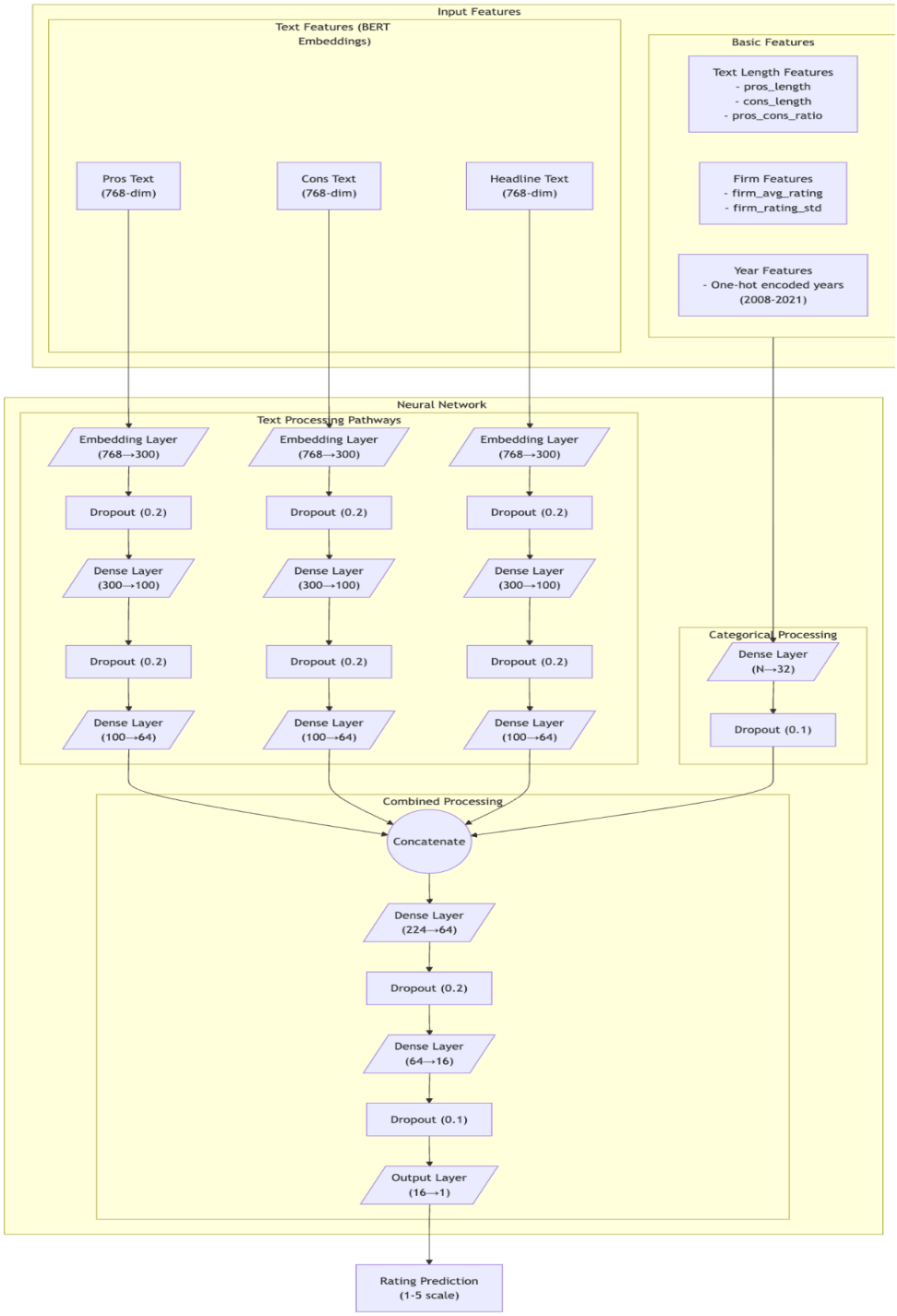
Anonymous name: zesty

MSE in training data: 0.4445
R2 in training data: 0.6388
Type of model: Deep Learning (DistilBERT Embeddings + Neural Network)

Q2:



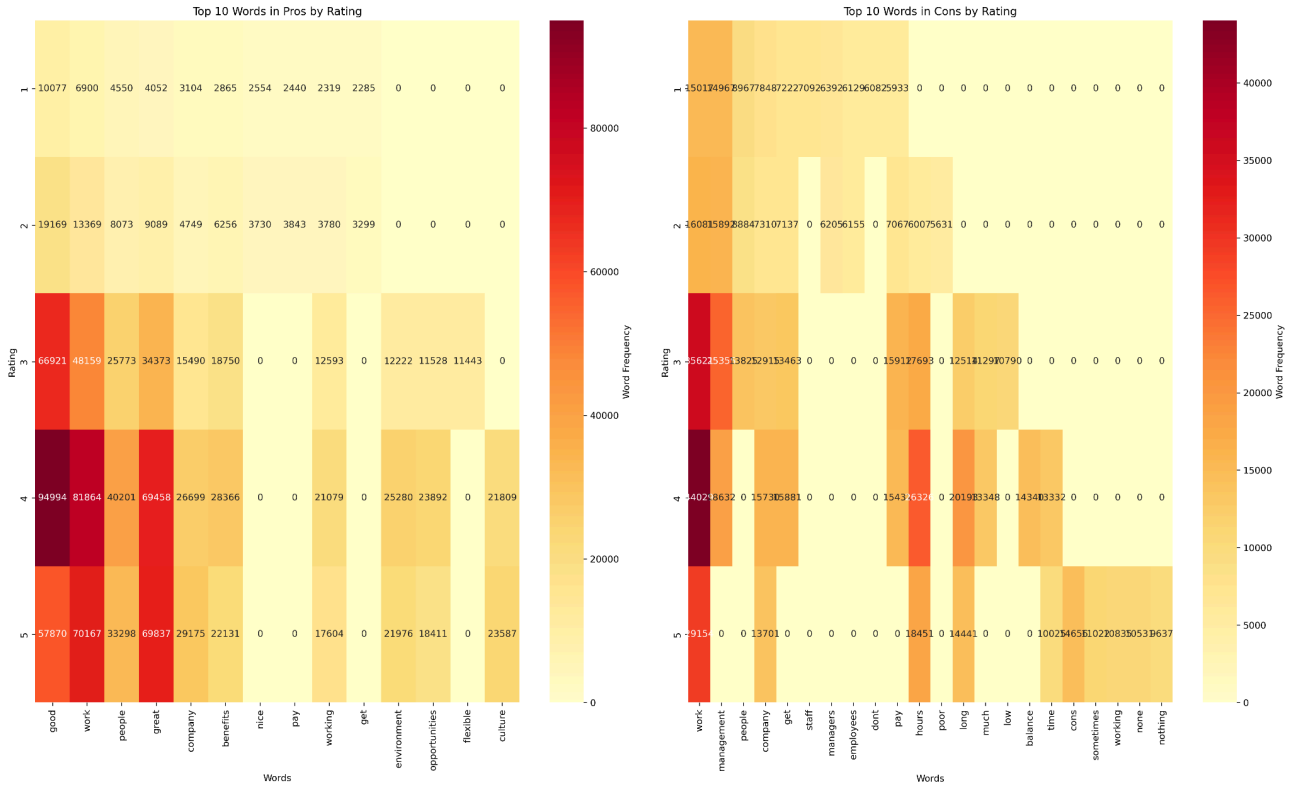
This diagram shows:

1. **Input Features**:
    - Text Features (processed with DistilBERT):
        - Pros text (768-dimensional embeddings)
        - Cons text (768-dimensional embeddings)
        - Headline text (768-dimensional embeddings)
    - Basic Features:
        - Text length metrics
        - Text sentiment metrics
        - Firm statistics
        - Year information
2. **Model Architecture**:
    - Three parallel pathways for text processing
    - Separate pathway for categorical/numerical features
    - Progressive dimension reduction
    - Multiple dropout layers for regularization
    - Final layers combining all features
3. **Feature Processing**:
    - Text embeddings go through multiple dense layers with dropout
    - Categorical features get compressed to 32 dimensions
    - All pathways are concatenated before final processing
4. **Output**:
    - Final sigmoid activation scaled to 1-5 range for rating prediction

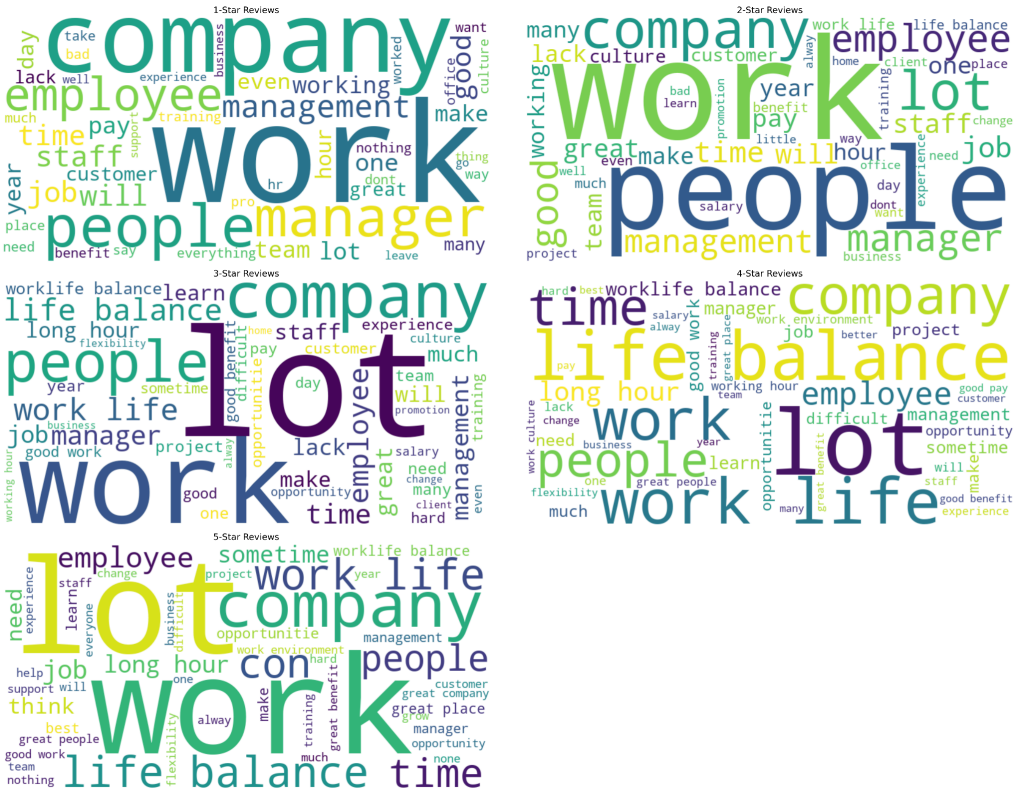
Q3:

a)

Word Frequency Heatmaps in the Top 10 Most Common Words in Pros & Cons Across Different Ratings (1 to 5 Stars)



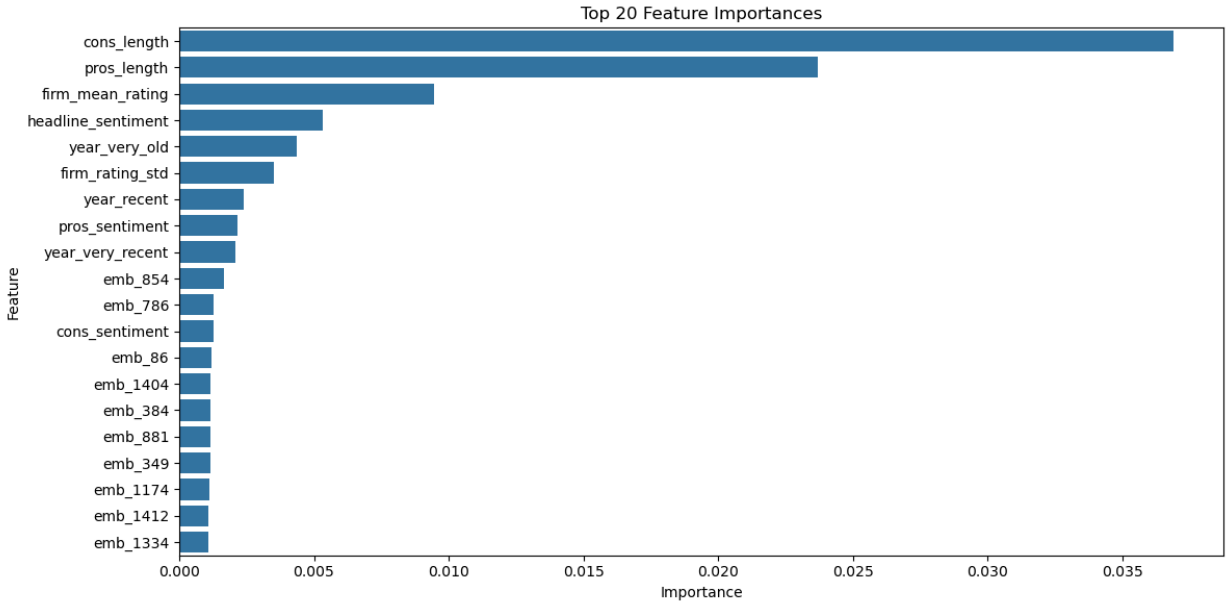
Word Cloud for each review (1 - 5)



- More common words are larger

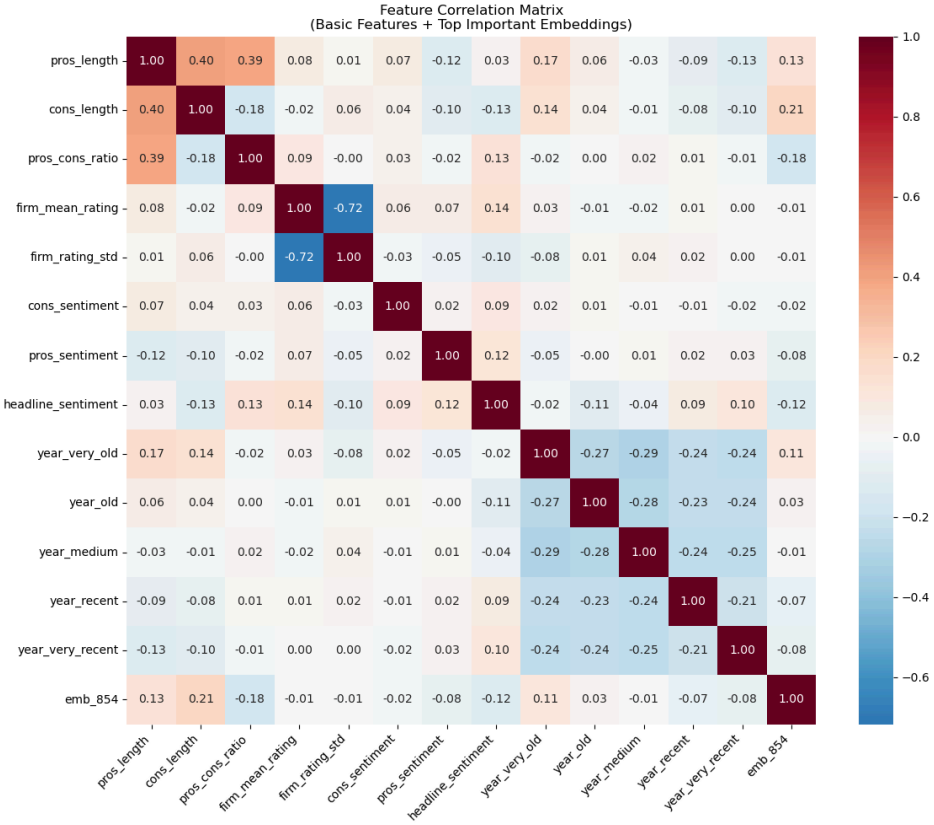
b)

Feature importance



- Text length features (cons_length, pros_length) are the most important

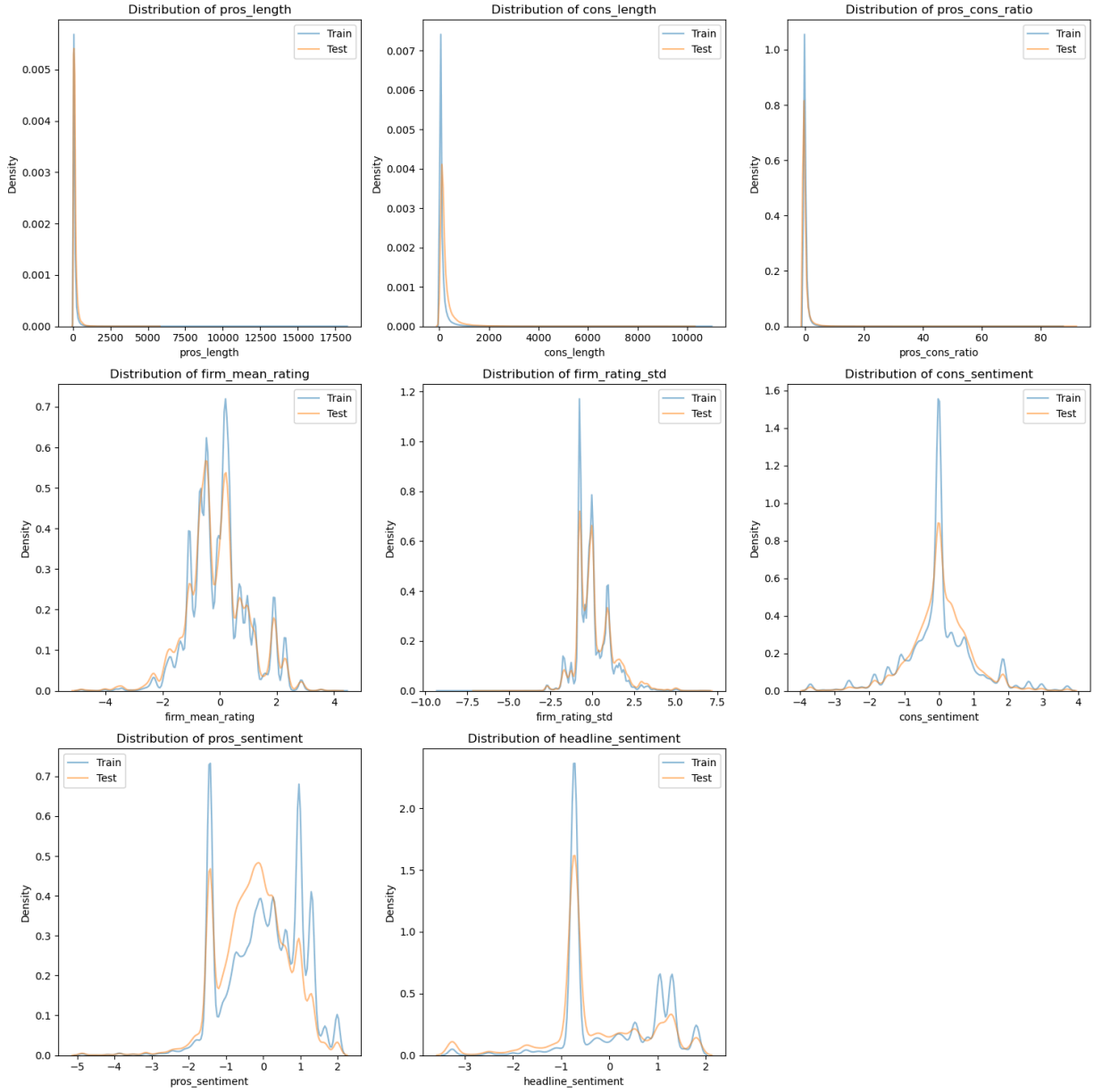
Feature correlation



- Strong negative correlation (-0.72) between firm_mean_rating and firm_rating_std
- Moderate positive correlation (0.40) between pros_length and cons_length
- Year category variables are negatively correlated with each other (expected since they're mutually exclusive)
- Most embedding features show weak correlations with other features
- Most features show low correlation, suggesting they provide independent information

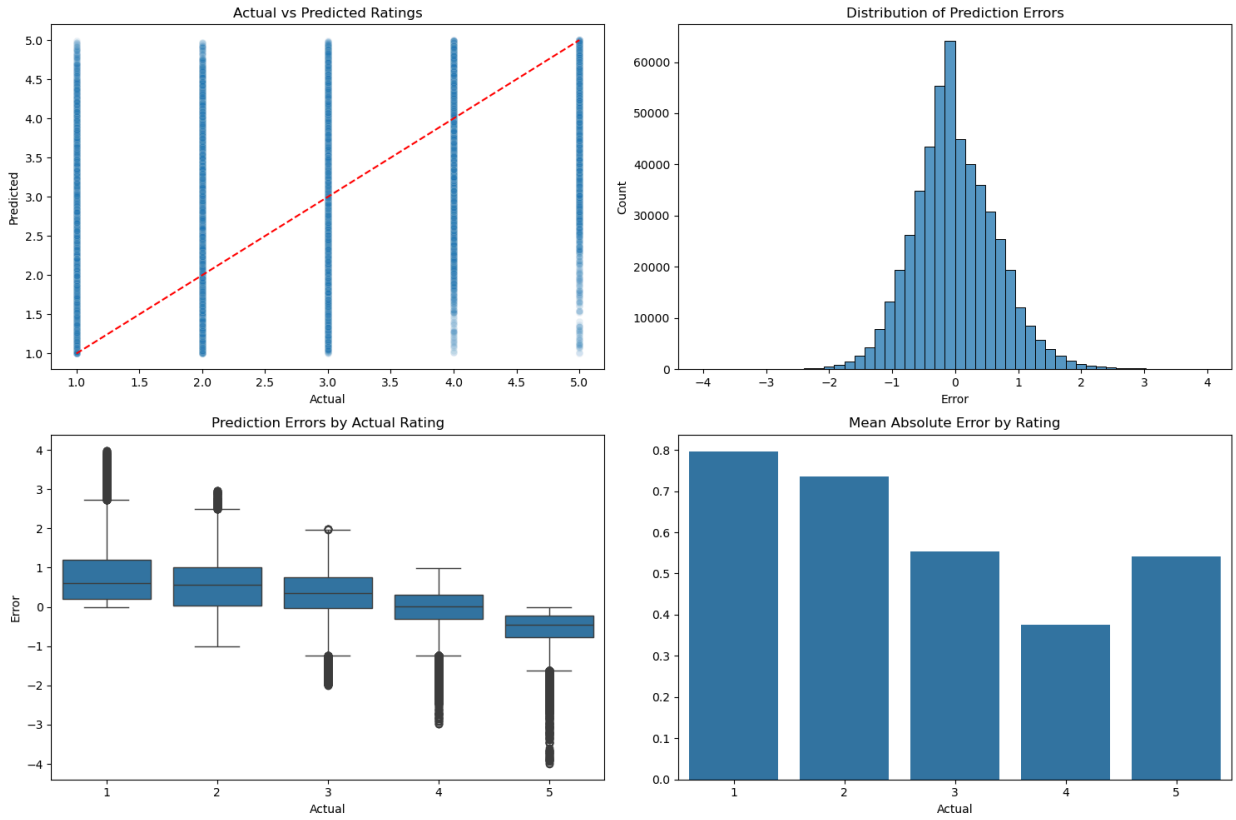
Distribution of each feature in the training data versus in the test data

- There would be too many graphs to plot the distribution of all embeddings (~1200), so I've skipped them here



- Overall, train and test distributions are fairly similar

How large and common are prediction errors at different points of the distribution of the response variable in the training data



- Error distribution is roughly normal but slightly skewed
- Larger errors for extreme ratings (1 and 5 stars)
- Mean Absolute Error is highest for 1-star ratings (~0.8) and lowest for 4-star ratings (~0.4)
- Model tends to overpredict low ratings and underpredict high ratings (regression to the mean)

Q4:

Methodology:

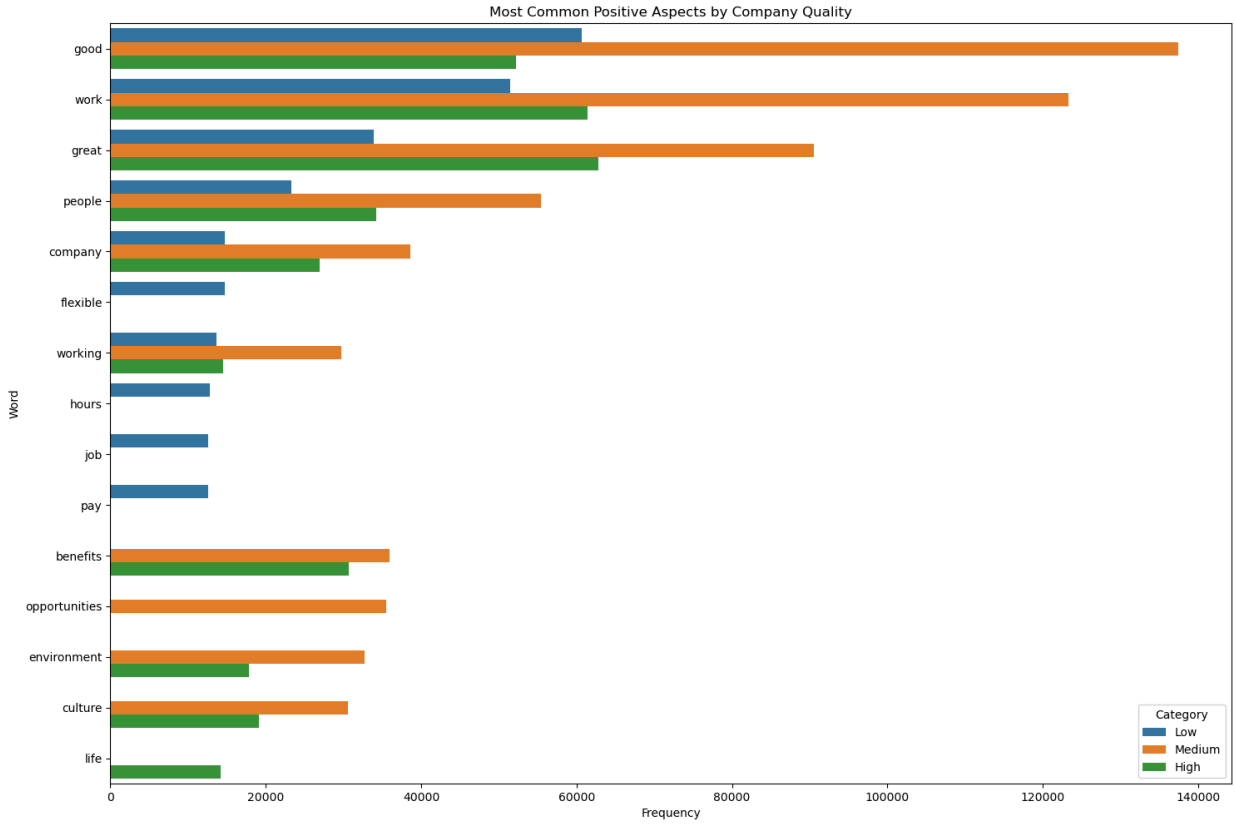
- Calculate average rating for each firm
- Group firms into categories (e.g., Low: <3, Medium: 3-4, High: >4)
- Keep the original review text (pros and cons) for each group
- For each quality group (Low/Medium/High):
    - Extract key phrases and topics from pros/cons using Natural Language Processing
    - Calculate frequency of these topics within each group
    - Compare topic distributions across groups

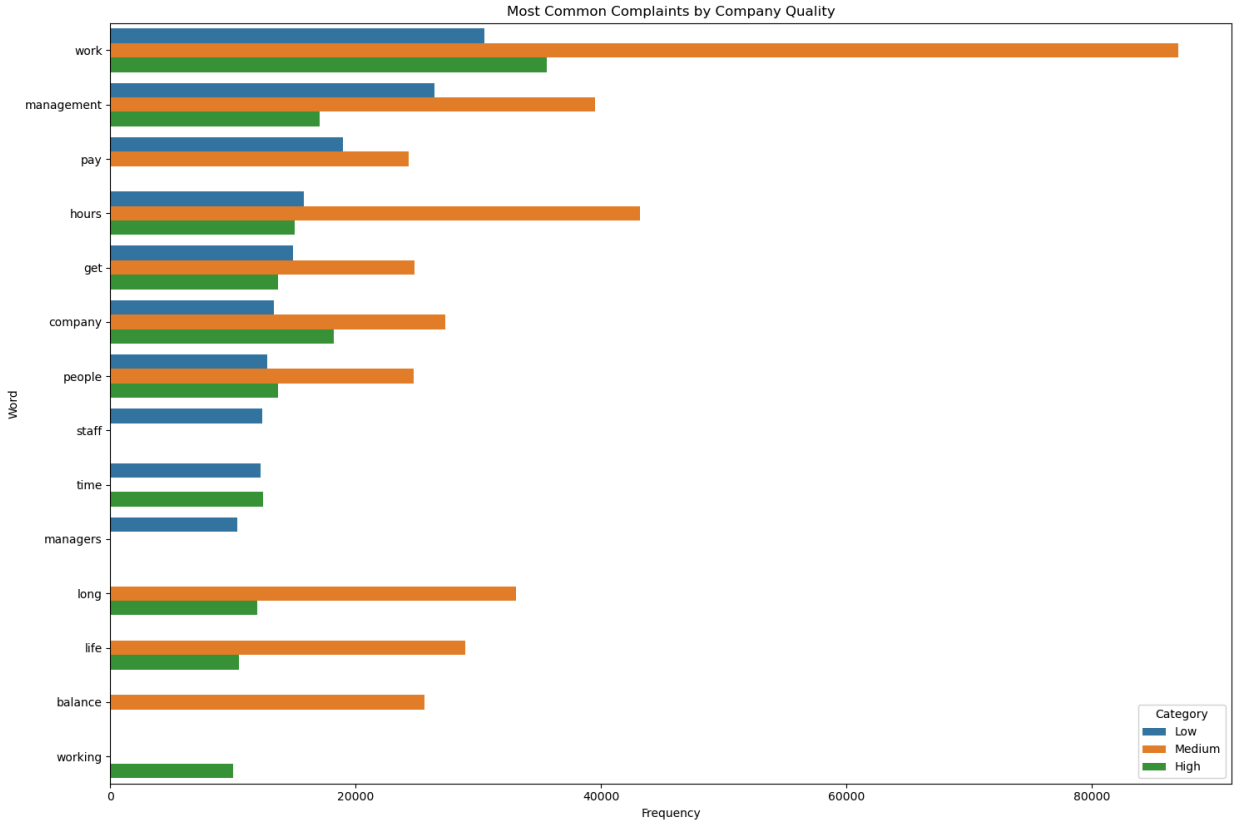
    - Identify what distinguishes each group (e.g., what do high-rated firms get praised for vs. what do low-rated firms get complained about)

Results:

- Low-rated companies:
    - Average rating: 3.39
    - Number of companies: 138
    - Top pros: good, work, great, people, company
    - Top cons: work, management, pay, hours, get
- Medium-rated companies:
    - Average rating: 3.72
    - Number of companies: 137
    - Top pros: good, work, great, people, company
    - Top cons: work, hours, management, long, life
- High-rated companies:
    - Average rating: 4.14
    - Number of companies: 138
    - Top pros: great, work, good, people, benefits
    - Top cons: work, company, management, hours, get

Visualization of the most common positive and negative aspects by company quality:





Analyzing the results:

- Common Patterns Across Quality Levels:
    - Work-related terms appear prominently in both pros and cons across all levels
    - Management appears as a con across all levels
    - People/culture aspects are consistently important in pros
- Key Differences and Actionable Insights:
    - Low-rated Companies (Avg 3.39):
        - Issues: More frequent complaints about pay and basic working conditions
        - Actionable Steps:
            - Focus on improving basic compensation structures
            - Address fundamental work-hour issues
            - Invest in management training (high frequency of management complaints)
    - Medium-rated Companies (Avg 3.72):
        - Issues: Work-life balance emerges ("long", "life" in cons)
        - Actionable Steps:
            - Implement better work-life balance policies
            - Address working hours flexibility
            - Review workload distribution
    - High-rated Companies (Avg 4.14):
        - Strengths: "Benefits" appears in top pros (unique to high-rated)
        - More balanced complaints (lower frequency of cons overall)
        - Actionable Steps:
            - Maintain strong benefits packages
            - Continue focus on people/culture
            - Fine-tune management practices
- Causality Discussion:
    - While this analysis reveals correlations between certain factors and company ratings, we cannot definitively claim causality because:
        - Confounding variables: Companies with better benefits might also have better management, making it hard to isolate the impact
        - Reverse causality: Successful companies might be able to afford better benefits, rather than benefits causing success
        - Selection bias: Reviews might come disproportionately from certain types of employees
        - Temporal aspects: We don't know if improvements in these areas actually led to rating improvements