Annotator Agreement: A Dual Perspective on Sentiments and Job Classifications

Rony Macwan
College of Information Science
University of Arizona
ronymacwan@arizona.edu

Abstract

Human annotators are critical to creating high-quality datasets in natural language processing (NLP), but their consistency in labeling can significantly impact model performance. This study investigates inter-annotator agreement across two classification tasks: sentiment analysis and job role categorization. Using a dataset annotated independently by two annotators, I assess agreement levels using various metrics, including Accuracy, Cohen's Kappa (K), and Confusion Matrices. These metrics provide a comprehensive evaluation of labeling consistency and highlight areas of agreement and disagreement. The analysis reveals the reliability of human annotations for each task and provides insights into the nature and extent of discrepancies between annotators. This work emphasizes the importance of consistent and accurate annotation practices in NLP workflows to ensure the reliability of labeled datasets.

1 Introduction

In the field of natural language processing (NLP), human annotation is a crucial step for generating labeled datasets that enable machine learning models to make accurate predictions. However, the consistency and reliability of human annotators can vary which can influence the quality of the labeled data. This project aims to compare the performance of two annotators in identifying sentiments versus identifying job roles, based on job descriptions. Sentiment analysis, which involves categorizing text as positive, negative, or neutral, requires an understanding of emotional tone, while job role classification demands identifying specific career titles from descriptive text. By evaluating the inter-annotator agreement on these two distinct tasks, I aim to explore whether annotators are more consistent in identifying sentiment or in categorizing job roles. This comparison is important because it provides insights into the challenges faced by human annotators when labeling complex text and highlights the factors that may affect annotation reliability. Through this study, I also examine various agreement metrics such as Cohen's Kappa, Agreement Scores, and Accuracy which will offer a quantitative understanding of annotator alignment and the potential discrepancies in their classifications. The goal of this research is to shed light on the importance of accurate annotation practices, contributing to the improvement of NLP tasks that rely on human-generated labels.

2 Methods

To create the dataset for this study, I first manually wrote sentences for three sentiment categories: Positive, Negative, and Neutral. Each sentence was carefully crafted to represent the tone and emotional context associated with these categories. Next, I wrote 100 job descriptions for three distinct roles: Data Scientist, Data Engineer, and ML Engineer. To introduce more variation into the dataset and ensure a diverse range of job descriptions, I prompted ChatGPT 4.0 to generate 40 additional job descriptions. These descriptions were created without specifying which role each description was meant for. This provided a mix of job descriptions that required classification into one of the three roles. This approach was intended to add variety and simulate a real-world scenario where job descriptions may be more ambiguous or less structured. Once the dataset was prepared, I began annotating the data, classifying the sentences according to sentiment (Positive, Negative, Neutral) and categorizing the job descriptions into their respective roles (Data Scientist, Data Engineer, ML Engineer). To ensure consistency and reduce bias in the annotation process, I created a set of annotation guidelines. These guidelines were then shared with a second annotator, my friend, who was tasked with annotating the same dataset according to the established guidelines. This process allowed for the comparison of the annotations made by two independent annotators. This practice provides the foundation for evaluating the inter-annotator agreement on both sentiment classification and job role categorization. For the sentiment classification task, I utilized a zero-shot classification approach using the "facebook/bart-large-mnli" model. This method allowed us to classify text into predefined sentiment categories, Positive, Negative, and Neutral, without requiring task-specific training. To enhance the model's ability to detect Neutral sentiments, I expanded the label definition to include terms like "factual" and "objective." This provided additional context for the classifier to distinguish Neutral content effectively. The process involved scoring each sentence against the candidate labels and identifying the most probable class based on confidence scores. Additionally, I incorporated a secondary check to detect Neutral sentences by examining the closeness of scores between competing labels and verifying whether the sentence exhibited objective characteristics. This two-pronged approach aimed to reduce misclassification of Neutral sentences and improve overall sentiment differentiation.

3 Results

Cohen's Kappa: The Cohen's Kappa results show a moderate agreement of 0.79 for the Sentiment category, indicating that the two annotators were generally consistent in identifying sentiments as Positive, Negative, or Neutral (Table 1). However, the moderate score suggests some degree of disagreement, likely due to the inherent subjectivity in sentiment classification. In contrast, the Job Role category shows a strong agreement of 0.86, indicating that the annotators were highly consistent in classifying job roles such as Data Scientist, ML Engineer, and Data Engineer (Figure 1). This higher Kappa score may reflect clearer, more objective distinctions between job roles compared to the potentially more nuanced task of sentiment identification, where personal interpretation and context can influence the outcome.

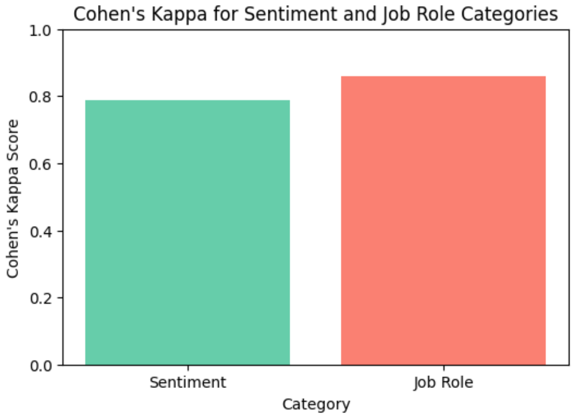

Figure 1: Bar Chart for Cohen's Kappa

Category	Cohen's Kappa
Sentiment Category	0.79
Job Role Category	0.86

Table 1: Cohen's Kappa Scores

Observed Agreement Percentage: (Table 2) For the Sentiment category, the highest agreement was observed for the Negative sentiment, with an agreement of 82.69%. This was closely followed by Positive sentiment, which had an agreement of 80.00%. In contrast, Neutral sentiment showed a significantly lower agreement of 52.94%, suggesting that both annotators had more difficulty reaching consensus on neutral statements. For the Job Role category (Table 3), the agreement percentages were higher across all values compared to the Sentiment category. The highest agreement was seen for Data Scientist, with 93.22%, reflecting a strong consensus between the two annotators. ML Engineer followed closely with an agreement of 78.43%, while Data Engineer showed the lowest agreement at 74.42%. (Figure 2) These trends indicate that job-related roles tend to have more consistent annotation, with Data Scientist standing out as the most clearly defined role among both the annotators.

Value	Agreement (%)
Positive	80.00
Negative	82.69
Neutral	52.94

Table 2: Agreement Scores for Sentiments

Value	Agreement (%)
Data Scientist	93.22
Data Engineer	74.42
ML Engineer	78.43

Table 3: Agreement Scores for Job Roles

Accuracy: (Table 4) Assuming that Annotation1 (the author's annotation) represents the ground truth, the accuracy of the annotation process was calculated by comparing Annotation2 against this reference. For the Sentiment Category, the accuracy was 87.27%, indicating that Annotation2

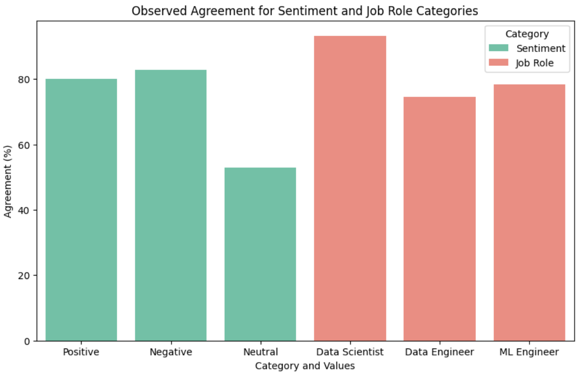

Figure 2: Observed Agreement

closely aligns with the ground truth in most cases. Similarly, the Job Role Category achieved an accuracy of 90.71%, reflecting a high level of agreement between the two annotators in labeling job roles.

Category	Accuracy (%)
Sentiment	87.27
Job Role	90.71

Table 4: Accuracy Scores for Categories

Confusion Matrices: (Table 5) Similarly, with Annotation1 as ground truth, the confusion matrix shows that in the Sentiment category, Annotation1 and Annotation2 show strong agreement in the Positive and Negative categories, with 44 and 43 matches respectively. However, there is some confusion with Neutral, where there are 5 instances of Annotation1: Neutral being misclassified as Annotation2: Positive or Negative. Overall, the matrix indicates high consistency in sentiment classification, with only a few misclassifications.

Annotation1/Annotation2	Positive	Negative	Neutral
Positive	44	6	0
Negative	0	43	3
Neutral	5	0	9

Table 5: Confusion Matrix for Sentiment Category

(Table 6) For the Job Role category, Annotation1 and Annotation2 align well for the Data Scientist and ML Engineer roles, with 55 and 40 matches, respectively. The Data Engineer role shows more variability, with 9 instances of misclassification between Data Engineer and ML Engineer. Overall, the agreement between the two annotators is strong.

Precision, Recall, F1 Score: (Table 7) For the Sentiment Category, the model performs best on the Negative class, with a high recall of 93% and an F1-score of 91%, indicating strong identification of negative sentiments. The Positive class also shows solid performance, with an F1-score of 89%. However, the Neutral class has lower precision (75%) and recall (64%), suggesting challenges in distinguishing neutral sentiments. Overall accuracy is 87%. (Table 8) For the Job Role Category, Data Scientist exhibits excellent metrics, with an F1-score of 96%. ML Engineer shows high precision (100%) but comparatively lower recall (78%), indicating some false negatives. Data Engineer achieves perfect recall (100%) but has the lowest precision (74%), reflecting false positives. The overall accuracy is impressive at 91%.

Annotation1/Annotation2	Data Scientist	Data Engineer	ML Engineer
Data Scientist	55	2	0
Data Engineer	0	32	0
ML Engineer	2	9	40

Table 6: Confusion Matrix for Job Role Category

Sentiment Category	Precision	Recall	F1-Score	Support
Positive	0.90	0.88	0.89	50.00
Negative	0.88	0.93	0.91	46.00
Neutral	0.75	0.64	0.69	14.00

Table 7: Classification Metrics for Sentiment Category

Job Role Category	Precision	Recall	F1-Score	Support
Data Scientist	0.96	0.96	0.96	57.00
Data Engineer	0.74	1.00	0.85	32.00
ML Engineer	1.00	0.78	0.88	51.00

Table 8: Classification Metrics for Job Role Category

Zero-Shot Classification: The zero-shot sentiment classification achieved an accuracy of 80.7% on the training dataset and 68.2% on the testing dataset. (Table 9, 10) The model demonstrates strong performance for the Negative sentiment in both training and testing, with high precision, recall, and F1-scores (training: 0.93; testing: 0.88). However, it struggles with the Neutral sentiment, achieving an F1-score of 0.00 in training and 0.40 in testing, indicating difficulty in distinguishing neutral statements (Figure 3). The Positive sentiment shows moderate performance, with consistent F1-scores (training: 0.82; testing: 0.67), though slightly reduced recall in testing suggests challenges in capturing all relevant positive cases. These trends highlight the model's effectiveness in polarized sentiment categories while underscoring the need for improvements in neutral classification (Figure 4).

Class	Precision	Recall	F1-Score
Negative	0.90	0.95	0.93
Neutral	0.00	0.00	0.00
Positive	0.81	0.83	0.82

Table 9: Training Classification Report

Class	Precision	Recall	F1-Score
Negative	0.78	1.00	0.88
Neutral	0.50	0.33	0.40
Positive	0.67	0.67	0.67

Table 10: Testing Classification Report

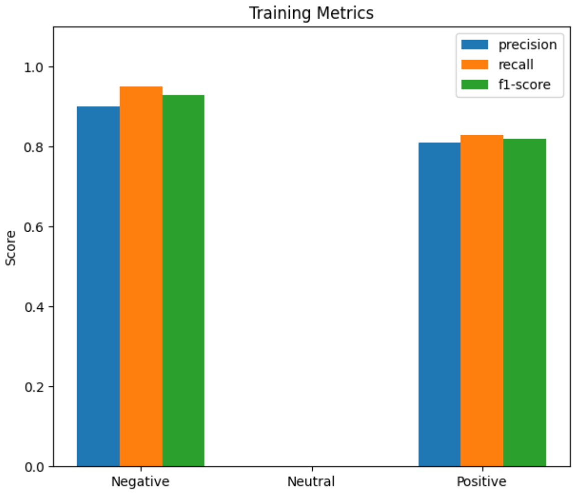

Figure 3: Classification for Training Data

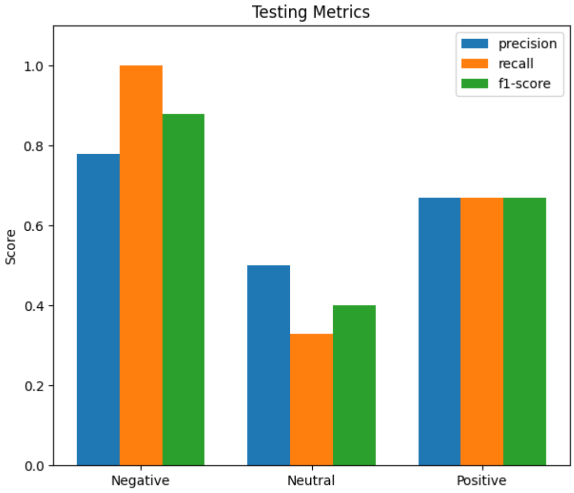

Figure 4: Classification for Testing Data

4 Conclusion

The Cohen's Kappa scores were 0.79 for sentiment classification, indicating substantial agreement, and 0.86 for job roles, reflecting almost perfect agreement. Agreement was notably higher for positive and negative sentiments compared to neutral, likely due to the inherent ambiguity in identifying neutrality within subjective or mixed contexts. For job role classification, the highest agreement was observed in the Data Scientist role. The relatively lower agreement for neutral sentiment may stem from the challenge of distinguishing neutral statements from those that are subtly opinionated or factual. This suggests that neutrality is context-dependent and harder to consistently classify. In contrast, job role classification achieved higher agreement due to both annotators' strong technical backgrounds and familiarity with job roles in the tech industry. This shared domain expertise likely contributed to more precise categorization. These findings highlight the importance of domain knowledge in annotation tasks. Future work could focus on enhancing the detection of neutral sentiment by refining the definitions or incorporating external contextual data, such as the broader discourse surrounding a statement. Additionally, using more advanced models, such as fine-tuned transformer-based architectures, could help improve classification accuracy, particularly for challenging categories. Moreover, further research could explore the impact of domain-specific training on sentiment and job role classification, considering the nuances of different industries. This study's methodology can be applied to various domains, such as customer feedback analysis, job market trends, and automated content categorization, where accurate sentiment and role classification play a critical role in understanding consumer or employee sentiments and optimizing business strategies.