Performance Analysis of Zero-Shot Classification Models for Sentiment Annotation

Shashwat Shrivastava

Abstract

This paper compares and evaluates the effectiveness of two NLP models, RoBERTa and BART, in annotating sentiment for a dataset of textual statements. The study compares their performance with human annotations, using metrics such as accuracy, Cohen's Kappa, Classification report and confusion matrix. Results reveal the models' strengths and limitations, including overclassification of Positive sentiment and poor performance in detecting Neutral sentiment. The findings offer insights into the applicability of zero-shot models for real-world sentiment analysis tasks.

1 Introduction

Sentiment analysis is a key component of NLP applications, from analyzing customer feedback to tracking trends on social media. This project explores the performance of two zero-shot classification models—RoBERTa and BART—in automatically annotating sentiments for a dataset of random statements. By comparing the models' annotations with human-generated labels, this study seeks to determine their reliability in categorizing statements as Positive, Negative, or Neutral. The goal was to test the reliability of zero-shot classification models in sentiment annotation tasks and analyze their performance using metrics such as accuracy, Cohen's Kappa, Classification report and confusion matrix.

2 Methodology

2.1 Data Collection and Human Annotation

1. The dataset consisted of randomly collected 150 textual statements (referred to as "Tweets" in the dataset) stored in an Excel file. 2. Dataset was first Annotated manually using spreadsheets. Each statement was classified into one of three sentiment categories: Positive, Negative, or Neutral. The statements in the manually annotated dataset were labelled as 50 positives, 50 negatives and 50 neutral.

2.2 Annotation using Zero-Shot Models

Two pre-trained zero-shot classification models—RoBERTa (roberta-large-mnli) and BART (facebook/bart-large-mnli)—were used. They were used because they can perform well without being pre-trained with the specific data or without fine-tuning. Both models employed the Hugging Face 'pipeline' function for zero-shot classification, categorizing statements into Positive, Negative, or Neutral. Outputs from each model were stored for comparison with human annotations.

2.3 Evaluation Metrices

Model performance was evaluated using the following metrics:
1. **Accuracy/ Percentage Agreement**: Measures the proportion of correct classifications compared to human annotations.
2. **Cohen's Kappa**: Assesses agreement between model predictions and human annotations, adjusting for chance.
3. **Confusion Matrix**: Visualizes classification performance, showing misclassification patterns among sentiment categories.
4. **Classification Report:** Performance overview of the model.

3 Results and Analysis

The evaluation metrics for both RoBERTa and BART models include accuracy/ percentage agreement, Cohen's Kappa, precision, recall, and F1-score. The results are summarized below:

3.0.1 BART Model:

```
BART Accuracy:  0.74
Cohen's Kappa for BART model:  0.61
BART Performance:
              precision    recall  f1-score   support

    Negative       0.83      1.00      0.91        50
     Neutral       1.00      0.22      0.36        50
    Positive       0.63      1.00      0.78        50

    accuracy                           0.74       150
   macro avg       0.82      0.74      0.68       150
weighted avg       0.82      0.74      0.68       150
```

Figure 1: Performance overview of BART model when compared with human-annotator.

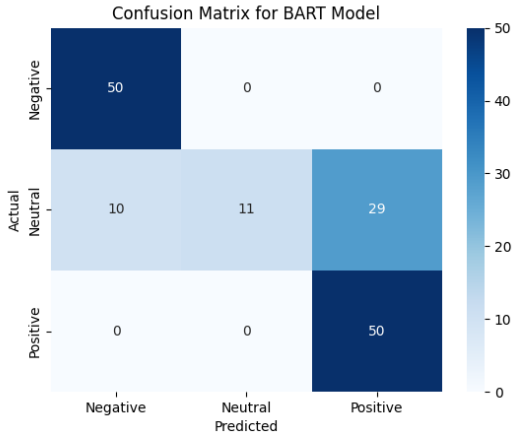

Figure 2: Confusion matrix for BART Model.

1. Figure 1 shows the percentage agreement (accuracy), Classification Report and Interannotator agreement between the model and the human-annotator.

2. The model labeled 79 statements as positive, 60 statements as negativenand 11 as neutral as shown in the confusion matrix (Figure2).

3. The model had a 74% agreement with Human-annotator (accuracy).

4. Cohen's Kappa had a value of 0.61 showing Moderate level of agreement with the human annotator and thus the data generated could only be 35-63% reliable.

3.0.2 RoBERTa Model:

```
RoBERTa Accuracy:  0.7
Cohen's Kappa for RoBERTa model:  0.55
RoBERTa Performance:
              precision    recall  f1-score   support

    Negative       0.83      1.00      0.91        50
     Neutral       1.00      0.10      0.18        50
    Positive       0.59      1.00      0.74        50

    accuracy                           0.70       150
   macro avg       0.81      0.70      0.61       150
weighted avg       0.81      0.70      0.61       150
```

Figure 3: Performance overview of RoBERTa model when compared with human-annotator.

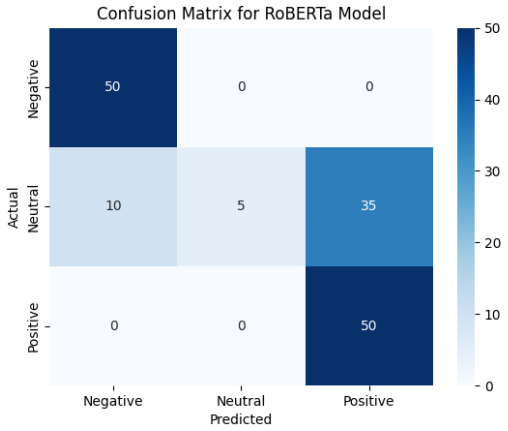

Figure 4: Confusion matrix for RoBERTa Model.

1. Figure 3 shows the percentage agreement (accuracy), Classification Report and Interannotator agreement between the model and the human-annotator.

2. The model labeled 85 statements as positive, 60 statements as negative and 5 as neutral as shown in the confusion matrix (Figure4).

3. The model had a 70% agreement with Human-annotator (accuracy).

4. Cohen's Kappa had a value of 0.55 showing Weak- Moderate level of agreement with the human annotator and thus the data generated could only be 15-35% reliable.

4 Key Insights

1. BART model was slightly better performing , as compared to RoBERTa model due to its higher accuracy.

2. For the "Positive" and "Negative" classes, both models obtained high recall (1.00), demonstrating their accuracy in recognizing these labels.

   However, BART's precision was higher (0.63) than RoBERTa's (0.59) for the "Positive" Label, indicating a more balanced prediction for this class.

3. Both models had considerable trouble in correctly recognizing "Neutral" statements: only 11 statements were successfully identified as

Neutral by RoBERTa and BART accurately identified only 5 statements as neutral.

4. Both models majorly misclassified "Neutral" statements as "Positive." Thus, showing a tendency to overpredict the “Positive” Class.

5. Cohen’s Kappa Score for both the models indicate a moderate level of agreement between human annotations and model predictions, with BART surpassing RoBERTa by a little margin.

5 Conclusion

This project compared the performance of two pretrained zero-shot classification models (BART and RoBERTa) when compared with the human- annotated data.While RoBERTa and BART performed well in detecting Positive and Negative sentiments, their inability to classify Neutral sentiments accurately underscores the need for further fine-tuning. These insights can guide the deployment of such models in real-world applications.