Multi-Label Industry Classification Experimental Report I

Stanislav Kharchenko
twotensor.com
daniel@twotensor.com

22 August 2023

1 Introduction

We aim to refine a multi-label classification methodology to achieve more accurate industry/sector categorization based on textual project descriptions. In the context of multi-label classification, it is feasible for a single project description to correspond to multiple industry sectors. This introduces intricate challenges, not only in terms of adapting the language model but also in identifying suitable metrics to evaluate its efficacy.

In this report, we offer a comprehensive evaluation of the capabilities of untuned GPT-3, focusing particularly on the GPT-3 Turbo 0301 variant. We use the following quantitative measurements to assess its performance:

- **Subset Accuracy**: Measures the proportion of instances where the entire set of predicted labels for an instance exactly matches the true set of labels. A higher accuracy is better because it means fewer incorrect labels.

- **Hamming Loss**: Measures the fraction of labels that are incorrectly predicted, i.e., the fraction of the wrong labels to the total number of labels. It's suitable for imbalanced datasets. A lower Hamming Loss is better because it means fewer incorrect labels.

- **Jaccard Similarity**: Measures the similarity between the predicted set of labels and the true set for each instance, indicating how closely the model's predictions align with the actual outcomes. A higher Jaccard Similarity is better, as it indicates greater overlap between the predicted and true labels.

- **Precision**: Indicates the proportion of correctly predicted positive observations to the total predicted positives.
- **Recall**: Determines the proportion of correctly predicted positive observations to the total actual positives.

1.1 Data Processing

We processed a random sample of project descriptions through the unmodified GPT-3 model. The ensuing responses were generated employing the subsequent prompt which involved classification into one or more industry/sector based on project descriptions.

2 Data Description

2.1 Dataset

We used a random sample of approximately 200 projects manually tagged by a human expert. True labels have the following distribution in this dataset.

3 Performance

3.1 Metrics

Metrics and their target values are tabulated below:

Metric	Value	Target Value
Subset Accuracy	0.1710	1
Hamming Loss	0.0662	0
Jaccard Similarity	0.4809	1
Precision	0.5692	1
Recall	0.6896	1

Table 1: Performance Metrics

3.2 Label ROC-AUC

Note: A higher AUC value indicates better performance.

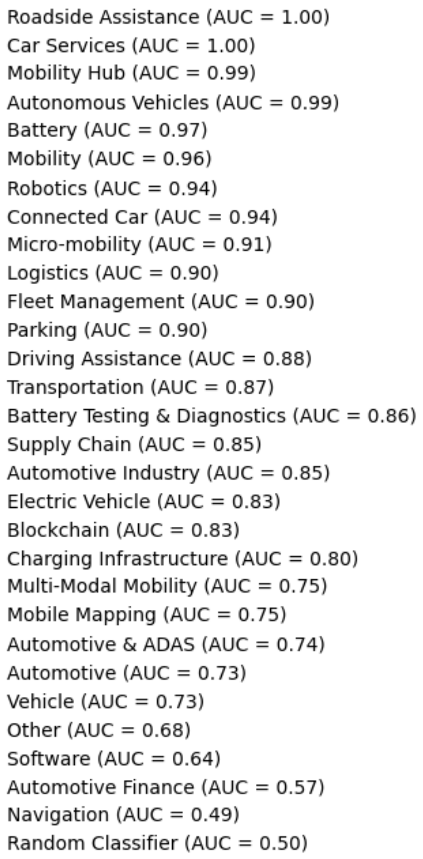

Figure 1: ROC-AUC Curve for Each Tag

3.3 Predicted Label Distribution

Predicted labels are distributed in the following way.

4 Analysis

From the analysis of the empirical results, several salient observations can be made:

4.1 Key Points

1. The model demonstrates a greater predilection for the misclassification of projects as opposed to the omission of tags. This tendency is substantiated by a significantly elevated Recall, which implies a lower rate of false negatives, and a generally augmented false positive rate (for an in-depth analysis, consult the Appendix). These observations are further buttressed by a markedly low Hamming Loss of 0.06 and a moderate Jaccard Similarity coefficient of 0.48. A Hamming Loss of this magnitude signifies that roughly 6% of the labels are incorrectly assigned. Conversely, a Jaccard Similarity coefficient nearing 0.48 suggests that the intersection between the predicted and actual labels for each instance is approximately 50%. These metrics collectively imply that although the model chiefly produces accurate label predictions, it concurrently yields a substantial volume of incorrect labels. An exception to this pattern is manifest in the "Software" tag, which, both in predicted and actual instances, recurs most frequently within the dataset. This tag demonstrates an elevated false negative rate compared to its false positive rate, suggesting that the observed patterns for other labels may be contingent on the sample size.

2. The performance of the model varies considerably across distinct tags. As explicated in Section 3.2, the Area Under the Curve (AUC) values differ depending on specific tags. The dataset's imbalanced nature is noteworthy; some tags appear more frequently than others, leading to an increased vulnerability to misclassification for these prevalent tags. For example, the tags 'Software' (AUC = 0.64), 'Other' (AUC = 0.68), 'Logistics' (AUC = 0.90), and 'Charging Infrastructure' (AUC = 0.80) are significantly present in the dataset, yet their performance metrics are conspicuously divergent. Interestingly, 'Logistics,' with 30 observations, is more common than 'Charging Infrastructure,' which has 26 observations. Despite this, the model exhibits a higher performance metric for 'Logistics,' indicating that sector-specific challenges may exist for both machine learning models and human experts.

3. Overall, the model's performance can be characterized as adequate but suboptimal. A diminished Subset Accuracy may initially seem deleterious; however, this metric chiefly reflects the frequency with which the predicted tag set precisely aligns with the true tag set. Given the intri-

cate complexity of multi-label classification, such results are not unanticipated and can be considered within acceptable boundaries. Nevertheless, the model's propensity for generating erroneous tags serves as an indictment of the current approach and necessitates further scrutiny.

5 References

- https://arxiv.org/abs/2307.03172