Pay Code Classification using BERT-based Natural Language Processing

F. Rodriguez Angel - University of Antioquia

Context of Application

In payroll operations, specialists often need to quickly identify and classify pay codes within a paycheck. These pay codes, which can vary significantly across vendors and systems, fall into five categories: **benefit**, **income**, **taxes**, **employee retirement contribution**, **employer retirement contribution**, and **loan**. The current process for manually classifying pay codes is time-consuming, taking between 30 seconds and 8 minutes for each pay code. This delay presents inefficiencies, especially when accuracy is critical for audits, benefit processing, or other payroll-related tasks. To address this, we propose developing a machine learning model capable of automating the classification of pay codes, helping payroll operators reduce their workload and increase productivity. This model will use Natural Language Processing (NLP) to predict the correct class for each pay code based on the pay code string itself and an optional description field.

Machine Learning Objective

The main objective of this project is to accurately predict the classification of pay codes using their string labels and descriptions. We aim to achieve this through **Transfer Learning** by leveraging optimized BERT-based models, such as **DistilBERT** or **ALBERT**, which are known for their efficiency in processing text data. These models will help the system understand patterns in pay code nomenclature and any associated descriptions to assign them to the correct class.

Dataset

The dataset consists of approximately **1300 pay codes**, collected from **13 US-based payroll provider software vendors**. The pay codes and descriptions were obtained through public APIs. Each pay code belongs to one of the five classes, and the distribution of the classes is as follows:

- Benefit: 16%
- Income: 33%
- Taxes: 22%
- Employee retirement contribution: 13%
- Employer retirement contribution: 10%
- Loan: 6%

Each record in the dataset contains two fields:

- **Paycode**: A string identifier (e.g., **401K**, **DENTALINS**, **LOAN1**).
- **Description**: A text field explaining the pay code (which may be empty).

Given the imbalance in class representation, where the majority of pay codes belong to the "Income" class, we need to ensure our model is evaluated using appropriate metrics that reflect this imbalance.

Performance Metrics

The primary machine learning metric for evaluating the model will be the **F1 Score**. This is a suitable metric for our problem due to the imbalanced distribution of classes, as it combines precision and recall to ensure that both false positives and false negatives are properly accounted for. Additionally, the business metric we aim to improve is the **throughput of payroll operators**, specifically targeting a **16% increase in efficiency** in paycode classification during their routine tasks.

Methodology

Our approach to solving this problem draws inspiration from similar applications in the financial industry, where Natural Language Processing and deep learning have been successfully applied to classify transaction data. For example, **JP Morgan Chase** has implemented BERT-based models to categorize unstructured banking transaction descriptions with high accuracy, which closely parallels our goal of categorizing unstructured pay code strings [1].

Additionally, projects such as **AutoNLP for Bank Transaction Classification** demonstrate the effectiveness of using BERT models to process short financial texts, which further supports the relevance of applying this approach to our dataset [2]. For scalability and efficiency, **weakly supervised learning** has been employed in large-scale transaction classification systems, a technique that may become useful for us as we refine our approach and grow our dataset [3].

Lastly, ensuring the transparency and explainability of model predictions can be crucial in high-stakes financial processes. **Hybrid models** that combine traditional RNNs with deep learning, such as those used in **Deep Learning for Hybrid Transaction Classification**, offer ways to increase classification accuracy while maintaining interpretability through tools like SHAP and LIME [4].

References

1. Sigmoidal. (n.d.). Natural Language Processing in Banking - Current Applications. Emerj Artificial Intelligence Research.
2. mgrella. (2023). AutoNLP for Bank Transaction Classification. Hugging Face.
3. Mukherjee, S., & Bardonski, M. (2023). Scalable and Weakly Supervised Bank Transaction Classification. arXiv, DOI: 10.48550/arXiv.2305.18430.
4. Grella, M. (2023). Deep Learning Enhancing Banking Services: A Hybrid Transaction Classification Approach. Journal of Big Data, DOI: 10.1186/s40537-023-00633-8.