1. Journals and Abstract Counts

I chose two journals for this assignment:

Personnel Psychology (ISSN: 1744-6570)

Entrepreneurship Theory and Practice (ISSN: 1540-6520)

Using the CrossRef API, I collected abstracts published since January 1, 2020. After filtering for entries that included abstracts, I ended up with:

Personnel Psychology: 147 abstracts

Entrepreneurship Theory and Practice: 273 abstracts

2. Removing Copyright Statements

The abstracts downloaded from CrossRef actually don't include copyright statements or publisher information. But I still used a GPT-based approach (via the OpenAI API) to remove any extraneous text. Specifically, I sent each abstract to the GPT model with a system prompt instructing it to keep only the research content and remove all copyright or publisher notices.

3. Obtaining Embedding Vectors

I used two approaches to obtain embedding vectors for each abstract:

TF-IDF Vectorization:

I applied the TfidfVectorizer from scikit-learn, restricting the vocabulary to a maximum of 5,000 features. Each abstract was converted into a TF-IDF vector of size up to 5,000 (depending on the final vocabulary).

OpenAI Embeddings:

I used OpenAI’s text-embedding-ada-002 model (with the new openai.embeddings.create interface) to generate an embedding for each abstract. I stored these embeddings in a NumPy array for subsequent modeling.

4. Supervised Machine Learning Models

I trained two supervised learning models on both TF-IDF features and OpenAI Embeddings:

Logistic Regression

Random Forest

4.1 Hyperparameter Tuning

I used GridSearchCV with 5-fold cross-validation to optimize the following hyperparameters:

Logistic Regression:

Regularization parameter C (searched over [0.01, 0.1, 1, 10, 100])

Random Forest:

Number of trees n\_estimators (searched over [50, 100, 200])

Maximum depth max\_depth (searched over [None, 10, 20, 30])

The best parameters found were:

TF-IDF + Logistic Regression: C = 100

TF-IDF + Random Forest: n\_estimators = 200, max\_depth = None

OpenAI + Logistic Regression: C = 100

OpenAI + Random Forest: n\_estimators = 200, max\_depth = 10

4.2 Model Performance

After refitting each model with its best hyperparameters, I measured the following metrics on a hold-out test set:

Accuracy

Precision

Recall

F1 Score

MCC (Matthews Correlation Coefficient)

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **MCC** |
| --- | --- | --- | --- | --- | --- |
| **TF-IDF + Logistic Regression** | 0.9048 | 0.9074 | 1.0000 | 0.9515 | 0.8819 |
| **TF-IDF + Random Forest** | 0.9048 | 0.8596 | 1.0000 | 0.9245 | 0.8143 |
| **OpenAI Embeddings + Logistic Regression** | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| **OpenAI Embeddings + Random Forest** | 0.9643 | 0.9423 | 1.0000 | 0.9703 | 0.9282 |

4.2.1 Comparison of TF-IDF vs. OpenAI Embeddings

Overall, the OpenAI embedding-based models performed better than the TF-IDF-based models on this classification task, achieving higher accuracy, precision, recall, and F1 scores. The best-performing model was:

OpenAI Embeddings + Random Forest (C=100)

5. Manual Coding of 20 Journal of Management Abstracts

I also selected 20 random abstracts from the Journal of Management and manually labeled each abstract as ob or ent. I then ran the OpenAI Embeddings + Logistic Regression model (trained on the two main journals) to predict the labels of these 20 abstracts. Below is the confusion matrix comparing my manual coding to the model’s predictions:

Confusion Matrix:

[[6 0]

[5 9]]

Row 1: True label = ob

6 correct predictions as ob

0 misclassified as ent

Row 2: True label = ent

5 misclassified as ob

9 correct predictions as ent

The overall accuracy on these 20 samples was 0.75 (75%).

The model did perfectly on precision for “ent” (1.00), but it misclassified 5 out of 14 “ent” abstracts as “ob,” which lowered the recall for “ent” to 0.64.

5.1 Assessment of Model’s Performance

The model did not miss any “ob” abstracts (recall = 1.00 for ob), but it tended to occasionally predict “ob” when the true label was “ent.” As there are only 20 papers and the articles I found in JOM have some articles that are not strictly ob or entrepreneuship, I think the model is in general fine.

5.2 Other Observations

The Journal of Management abstracts might have somewhat different language usage or content emphasis compared to the two journals used for training. This domain shift could explain why the model struggled on some “ent” abstracts. In addition, some articles among the 20 ones are more like strategy articles. That’s why I think it’s difficult for the model to evaluate very accurately. Increasing the training data or including more varied journals in the training set might further improve generalization.