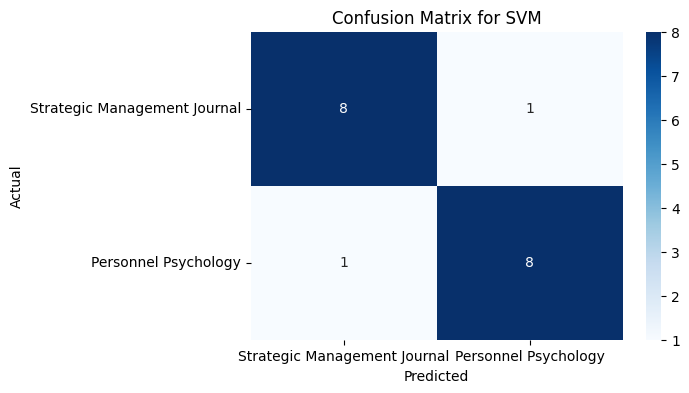
Carys Zhao - Summative Assignment Results

1. I chose Strategic Management Journal (SMJ) to represent the Strategy subdiscipline and Personnel Psychology (PPsych) to represent the Organizational Behavior subdiscipline. I downloaded 635 abstracts from SMJ and 314 abstracts from PPsych.
2. I defined a clean\_abstract function that includes a clause for removing common copyright phrases based on five different patterns. In the same function, I also included a clause to remove any rows without actual abstracts and a clause to remove prefixes such as “abstract” and “research summary”.
3. I applied two approaches for obtaining the embedding vectors for abstracts. The first is tf-idf, which is using TfidfVectorizer() from sklearn to convert abstracts into numerical representations.The second is using openAI through personal API key. I applied the gpt-4o-mini model and called get\_openai\_embedding(text), which sends each abstract to openAI’s API and returns a dense vector representation. These embedding vectors are then used to train machine learning models.
4. I trained two supervised machine learning models, which are Random Forest (RF) Classifier and Support Vector Machine (SVM), trained both by tf-idf and openAI. For RF trained by tf-idf, accuracy is .91, weighted average precision is .92, and weighted average recall is .91. For SVM trained by tf-idf, accuracy is .96, weighted average precision is .96, and weighted average recall is .96. For RF trained by openAI, accuracy is .91, weighted average precision is .92, and weighted average recall is .91. For SVM trained by openAI, accuracy is .97, weighted average precision is .97, and weighted average recall is .97. Taking together, the SVM model trained by openAI is the best performing model.



After excluding the two papers that did not provide abstracts, a total of 18 JOM papers were coded. The overall accuracy is .89. Both of precision and recall scores are also .89. Based on thee confusion matrix, the model predictions contain minimal misclassifications. The misclassifications are likely due to the ambiguity of the subject, with one abstract mentioning multiple fields and the other abstract about status, which can span across subdisciplines. Involving more abstracts in training could further refine performance.