INFO – 6148
Natural Language Processing
Garvit Sakhuja

Resume Categorization using Spacy

Introduction

This project aims at categorizing resumes as per their corresponding industries. Resumes are classified into 25 categories using Machine Learning models. Resume Categorization is a crucial task in HR Management, as it provides efficient candidate screening and recruitment by organizing resumes into predefined categories. This project utilizes spacy for text pre-processing and augmenty (library based on spacy) for augmenting resume texts for improving model's performance.

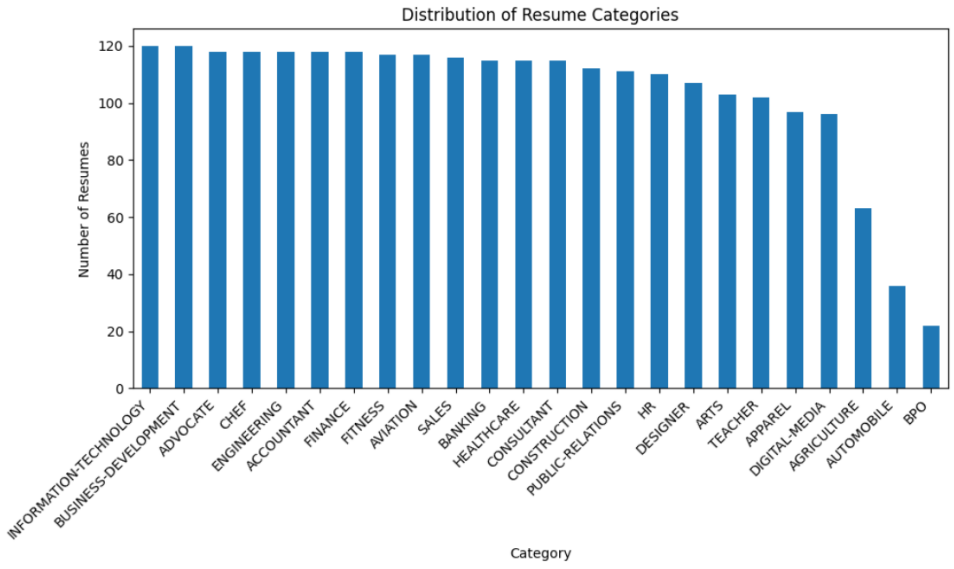
Dataset

The dataset used for this project is a Resume Dataset obtained from Kaggle. It contains the following columns:

1. ID: A unique identifier for each resume entry.
2. Resume_str: Contains the raw text of the content in Resume.
3. Resume_html: Provides an html version of the resume text.
4. Category: This column specifies category label for each resume.

For our purpose, we only require Resume_str and its category for classification task using machine learning techniques.

Resume Dataset



Data Preprocessing

For preprocessing of textual data, I took the following steps:

1. Lowercasing the text
2. Removing urls, emails and special characters.
3. Punctuation removed from the text.
4. Removing stopwords
5. Lemmatization of text
6. Tokenization of text

Text Augmentation was performed to enhance model performance, augmenty library for synonym replacement was used for this.

Text features were extracted by vectorizing the text. TF-IDF and Spacy vectorization was used to represent each resume as a vector.

Model Training and Evaluation

After vectorizing, the data is split into training and testing sets.

Models used

1. Random Forest Classifier
2. Support Vector Machine (SVM)

These models were tested with 2 hyperparameters, and each hyperparameter was tested with 2 different values. I used GridSearchCV to select the best parameter values for both the models.

Hyperparameters explored in Random Forest Model:

i. 'n_estimators': [50, 100]
ii. 'max_depth': [None, 10]

Hyperparameters explored in SVM Model:

i. 'C': [0.1, 1.0],
ii. 'kernel': ['linear', 'rbf']

TF-IDF Vectors – Model Evaluation

```
Random Forest Best Parameters: {'max_depth': None, 'n_estimators': 100}
Random Forest Accuracy: 0.862555720653789

Classification Report:
                        precision    recall  f1-score   support

            ACCOUNTANT       0.96      0.95      0.96        79
              ADVOCATE       0.84      0.93      0.88        68
           AGRICULTURE       0.90      0.82      0.86        34
              APPAREL        0.90      0.80      0.84        54
                 ARTS        1.00      0.10      0.18        31
            AUTOMOBILE       0.94      0.74      0.83        23
              AVIATION       0.86      0.73      0.79        33
               BANKING       0.88      0.32      0.47        44
                   BPO       1.00      1.00      1.00         9
  BUSINESS-DEVELOPMENT       0.82      0.95      0.88        63
                  CHEF        0.92      0.97      0.95        71
          CONSTRUCTION        0.83      0.96      0.89        50
            CONSULTANT        0.94      0.70      0.81        88
              DESIGNER        0.96      0.94      0.95        69
         DIGITAL-MEDIA        0.82      0.96      0.89        53
           ENGINEERING        0.88      0.96      0.92        73
               FINANCE        0.83      0.91      0.87        70
               FITNESS        0.93      0.93      0.93        69
            HEALTHCARE        0.83      1.00      0.91        65
                    HR        0.88      0.97      0.92        68
INFORMATION-TECHNOLOGY        0.79      0.97      0.87        74
      PUBLIC-RELATIONS        0.67      0.37      0.48        27
                 SALES        0.79      0.87      0.83        75
               TEACHER        0.76      0.96      0.85        56

              accuracy                             0.86      1346
             macro avg        0.87      0.83      0.82      1346
          weighted avg        0.87      0.86      0.85      1346
```

```
SVM Best Parameters: {'C': 1.0, 'kernel': 'rbf'}
SVM Accuracy: 0.8239227340267459

Classification Report:
                        precision    recall  f1-score   support

            ACCOUNTANT       0.91      0.95      0.93        79
              ADVOCATE       0.72      0.85      0.78        68
           AGRICULTURE       0.82      0.91      0.86        34
              APPAREL        0.78      0.70      0.74        54
                 ARTS        1.00      0.19      0.32        31
            AUTOMOBILE       1.00      0.39      0.56        23
              AVIATION       0.92      0.67      0.77        33
               BANKING       0.67      0.50      0.57        44
                   BPO       1.00      0.67      0.80         9
  BUSINESS-DEVELOPMENT       0.84      0.90      0.87        63
                  CHEF        0.94      0.94      0.94        71
          CONSTRUCTION        0.80      0.94      0.86        50
            CONSULTANT        0.86      0.64      0.73        88
              DESIGNER        0.97      0.86      0.91        69
         DIGITAL-MEDIA        0.78      0.87      0.82        53
           ENGINEERING        0.84      0.93      0.88        73
               FINANCE        0.89      0.90      0.89        70
               FITNESS        0.88      0.86      0.87        69
            HEALTHCARE        0.76      0.88      0.81        65
                    HR        0.87      0.97      0.92        68
INFORMATION-TECHNOLOGY        0.73      0.93      0.82        74
      PUBLIC-RELATIONS        0.79      0.56      0.65        27
                 SALES        0.70      0.83      0.76        75
               TEACHER        0.80      0.91      0.85        56

              accuracy                             0.82      1346
             macro avg        0.84      0.78      0.79      1346
          weighted avg        0.83      0.82      0.82      1346
```

Spacy Vectors – Model Evaluation

```
Random Forest Best Parameters: {'max_depth': None, 'n_estimators': 100}
Random Forest Accuracy: 0.7726597325408618

Classification Report:
                        precision    recall  f1-score   support

            ACCOUNTANT       0.81      0.91      0.86        79
              ADVOCATE       0.73      0.79      0.76        68
           AGRICULTURE       0.87      0.79      0.83        34
              APPAREL        0.83      0.72      0.77        54
                ARTS         1.00      0.16      0.28        31
            AUTOMOBILE       1.00      0.65      0.79        23
              AVIATION       0.78      0.42      0.55        33
               BANKING       0.60      0.14      0.22        44
                  BPO        0.90      1.00      0.95         9
  BUSINESS-DEVELOPMENT       0.66      0.92      0.77        63
                 CHEF        0.93      0.96      0.94        71
          CONSTRUCTION       0.66      0.88      0.75        50
            CONSULTANT       0.90      0.53      0.67        88
              DESIGNER       0.95      0.78      0.86        69
         DIGITAL-MEDIA       0.79      0.83      0.81        53
           ENGINEERING       0.71      0.85      0.78        73
               FINANCE       0.75      0.86      0.80        70
               FITNESS       0.77      0.87      0.82        69
            HEALTHCARE       0.78      0.91      0.84        65
                    HR       0.75      0.88      0.81        68
INFORMATION-TECHNOLOGY       0.72      0.88      0.79        74
      PUBLIC-RELATIONS       0.54      0.48      0.51        27
                 SALES       0.73      0.77      0.75        75
               TEACHER       0.78      0.84      0.81        56

              accuracy                           0.77      1346
             macro avg       0.79      0.74      0.74      1346
          weighted avg       0.78      0.77      0.76      1346
```

```
SVM Best Parameters: {'C': 1.0, 'kernel': 'linear'}
SVM Accuracy: 0.8142644873699851

Classification Report:
                        precision    recall  f1-score   support

            ACCOUNTANT       0.93      0.86      0.89        79
              ADVOCATE       0.71      0.87      0.78        68
           AGRICULTURE       0.67      0.94      0.78        34
              APPAREL        0.80      0.80      0.80        54
                ARTS         0.67      0.26      0.37        31
            AUTOMOBILE       0.58      0.83      0.68        23
              AVIATION       0.81      0.64      0.71        33
               BANKING       0.64      0.41      0.50        44
                  BPO        0.69      1.00      0.82         9
  BUSINESS-DEVELOPMENT       0.78      0.86      0.82        63
                 CHEF        0.90      0.97      0.93        71
          CONSTRUCTION       0.82      0.90      0.86        50
            CONSULTANT       0.84      0.60      0.70        88
              DESIGNER       0.92      0.84      0.88        69
         DIGITAL-MEDIA       0.80      0.96      0.87        53
           ENGINEERING       0.82      0.93      0.87        73
               FINANCE       0.78      0.89      0.83        70
               FITNESS       0.86      0.87      0.86        69
            HEALTHCARE       0.85      0.85      0.85        65
                    HR       0.90      0.91      0.91        68
INFORMATION-TECHNOLOGY       0.85      0.86      0.86        74
      PUBLIC-RELATIONS       0.80      0.44      0.57        27
                 SALES       0.82      0.77      0.79        75
               TEACHER       0.83      0.86      0.84        56

              accuracy                           0.81      1346
             macro avg       0.79      0.80      0.78      1346
          weighted avg       0.82      0.81      0.81      1346
```

Observations and Conclusion

Text augmentation of the Resume content improved model's performance significantly.

We trained 2 models for Resume Categorization: Random Forest and Support Vector Machine (SVM). We used two vectorization methods: TF-IDF and Spacy vectors. After training the models are evaluated using accuracy, classification reports and confusion matrix for both tf-idf and spacy vectorization.

When comparing the results for Random Forest and SVM for TF-IDF vectors, we can see Random Forest outperforming SVM in terms of accuracy.

- RF Accuracy - 86.2%
- SVM Accuracy - 82.4%

When comparing the results of the 2 models using spacy vectors, SVM outperforms Random Forest model here.

- RF Accuracy - 77.2%
- SVM Accuracy - 81.4%

Hence, we can say Random Forest is the best model for Resume Categorization using tf-idf vectorization giving an accuracy of 86.2%.