SOSC 4300 FINAL PROJECT

Fake Online Job Posting Detection

Team HaHa
Hui Ka Ming (20506919)
Lam Wing Yi (20510178)
Law Po Yi (20510673)
Lo Wing Ching (20521892)

TABLE OF CONTENTS

About the Project 01

Background 02

Data 03

04 Methodology

05 Preliminary Results

06 Conclusions

Recruitment Fraud

INTRODUCTION

- Fake recruitment posts, emails, URLs
- Offer fake job opportunity to job-seekers
- Taking money from job seekers
- Getting personal data
- Stealing personal identity

(Aurecon, n.d.)

JOB POST

BACKGROUND

$6 trillion

Cost of cybercrime damages

Over 600M

2030 Growth in Job Generation

- Employment scam is getting serious

  (Alghamdi & Alharby, 2019)
- Violation of reputed company

  (Dutta & Bandyopadhyay, 2020)
- Reliable Source

  Vidros, Kolias, Kambourakis, & Akoglu (2017) added many features of ORF to the public dataset (EMSCAD)

ABOUT THE PROJECT

Text Analytics

Model 1

- Logistics Regression

Model 2

- Multinomial Naive Bayes Classifier

Model 3

- SVC

Model 4

- Random Forest

Using Machine Learning based classification techniques to

Avoid fraudulent for job in the internet.

Objectives

1— Enhance accuracy of the model through pre-processing data

2— Apply feature selection techniques which assist to reduce dimensionality

3—Build a reliable model to detect ads with highest accuracy

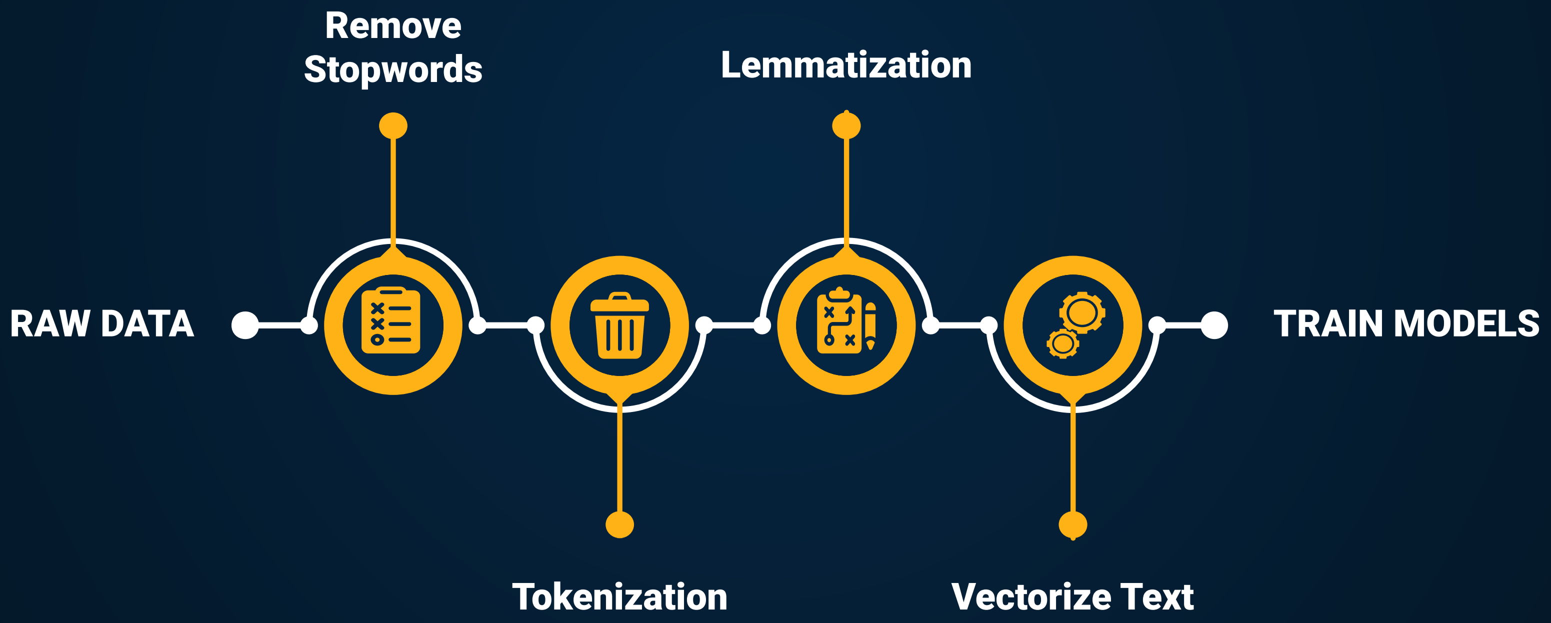
Find the best classification algorithm used for detecting Online Recruitment Fraud

DATA

- Employment Scam Aegean Dataset (EMSCAD)
- Published by the University of the Aegean
- Consist of 178800 online job advertisements
- Including title, location of job, company profile, description of the job etc.
- Columns “fraudulent” > legitimate job=0 and advertisements=1

METHODOLOGY

Pre-Processing & Cleaning



RAW DATA

Remove Stopwords

Tokenization

Lemmatization

Vectorize Text

TRAIN MODELS

METHODOLOGY

Train-Test Split

TRAIN DATASET

- Fit the machine learning model
- Split percentage: 80%

TEST DATASET

- Evaluate the fit machine learning model
- Split percentage: 20%

★ Estimate the performance of machine learning algorithms

METHODOLOGY

Text Analysis

LOGISTIC REGRESSION

works best on binary classification problems to examine the association of independent variable(s) with one dichotomous dependent variable.

MULTINOMIAL NAIVE BAYES CLASSIFIER

describe the probability of observing counts among a number of categories and most appropriate for features that represent counts or count rate

SVC

analyze data used for classification and regression analysis and helpful in text and hypertext categorization

RANDOM FOREST CLASSIFIER

construct a multitude of decision trees and output mean prediction of the individual trees

PRELIMINARY RESULTS

LOGISTIC REGRESSION

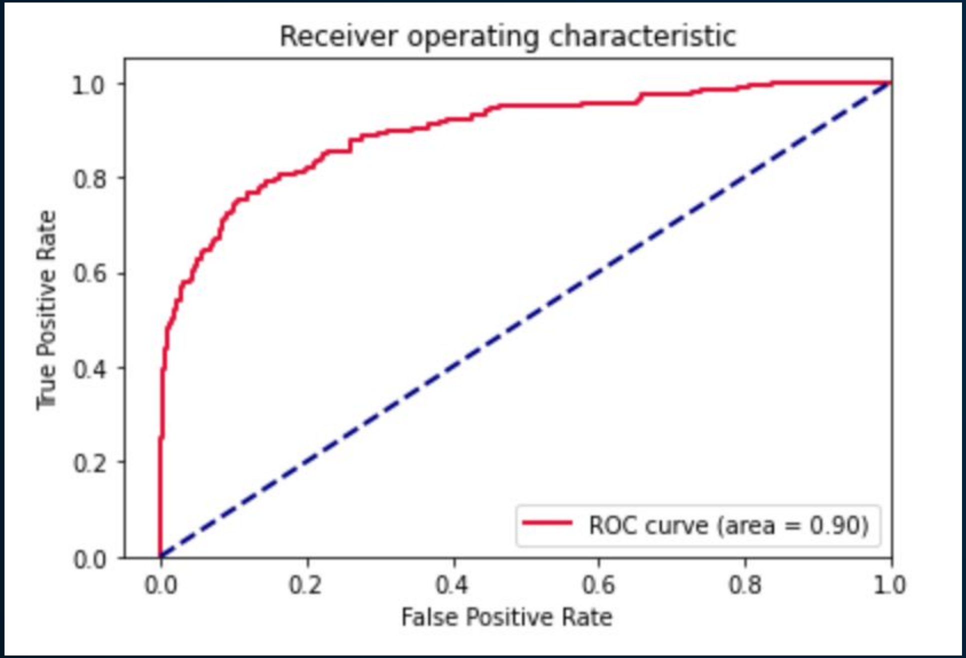
```
Classification Report of Logistic Regression:

              precision    recall  f1-score   support

           0       0.96      1.00      0.98      3324
           1       1.00      0.30      0.46       183

    accuracy                           0.96      3507
   macro avg       0.98      0.65      0.72      3507
weighted avg       0.96      0.96      0.95      3507
```

Accuracy= 0.963501

Weighted F1-score= 0.95
(best value at 1;
worst value at 0)

PRELIMINARY RESULTS

LOGISTIC REGRESSION



← **ROC Curve**

AUC= 0.897131
(larger AUC, better prediction performance)

PRELIMINARY RESULTS

MULTINOMIAL NAIVE BAYES CLASSIFIER

Bag-of-words model:

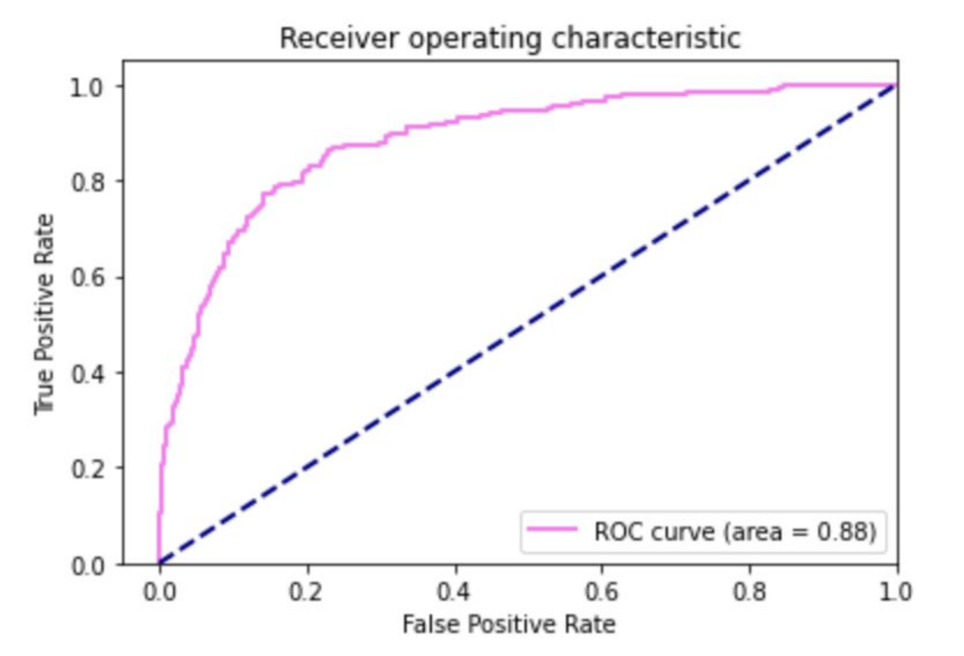
```
Classification Report of Multinomial Naive Bayes
Classifier:

              precision    recall  f1-score   support

           0       0.95      1.00      0.97      3324
           1       0.88      0.08      0.14       183

    accuracy                           0.95      3507
   macro avg       0.91      0.54      0.56      3507
weighted avg       0.95      0.95      0.93      3507
```

Accuracy= 0.94407

Weighted F1-score= 0.93
(best value at 1;
worst value at 0)

PRELIMINARY RESULTS

MULTINOMIAL NAIVE BAYES CLASSIFIER

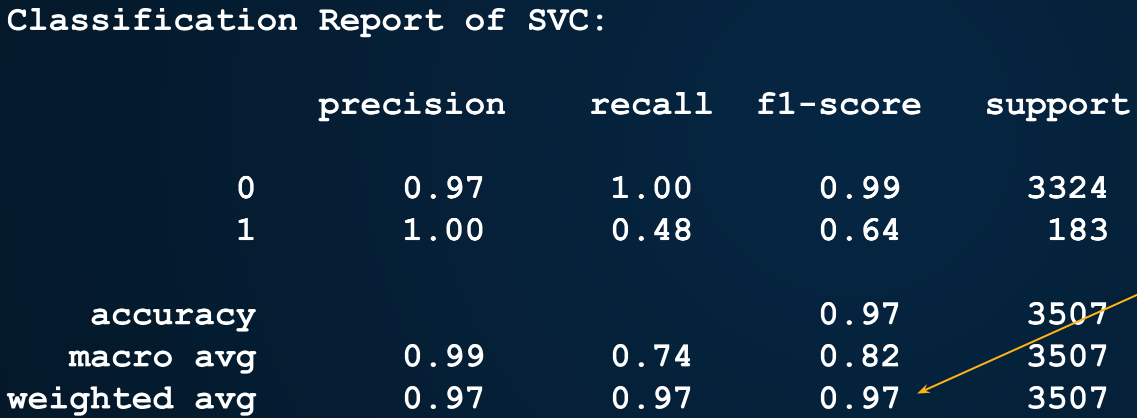


ROC Curve

AUC=0.883728

PRELIMINARY RESULTS

SVC



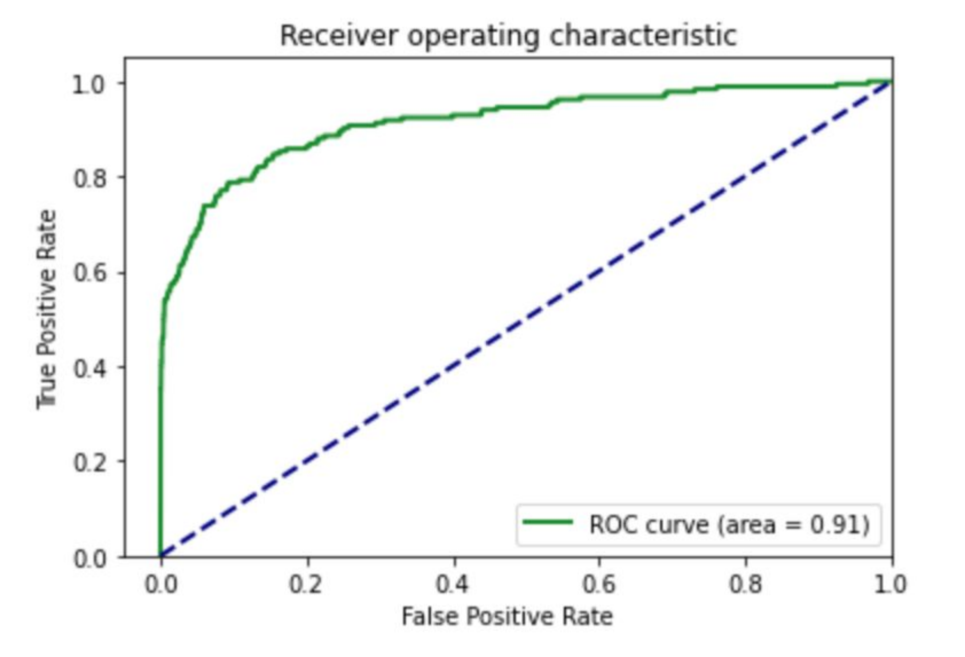
```
Classification Report of SVC:

              precision    recall  f1-score   support

           0       0.97      1.00      0.99      3324
           1       1.00      0.48      0.64       183

    accuracy                           0.97      3507
   macro avg       0.99      0.74      0.82      3507
weighted avg       0.97      0.97      0.97      3507
```

Accuracy= 0.972626

Weighted F1-score= 0.97 (best value at 1; worst value at 0)

PRELIMINARY RESULTS

SVC

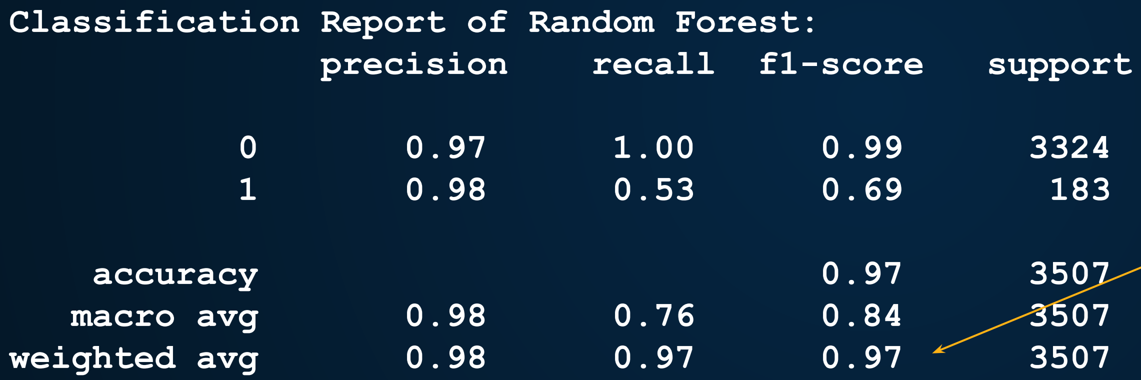


← ROC Curve

AUC=0.914577

PRELIMINARY RESULTS

RANDOM FOREST CLASSIFIER



```
Classification Report of Random Forest:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99      3324
           1       0.98      0.53      0.69       183

    accuracy                           0.97      3507
   macro avg       0.98      0.76      0.84      3507
weighted avg       0.98      0.97      0.97      3507
```

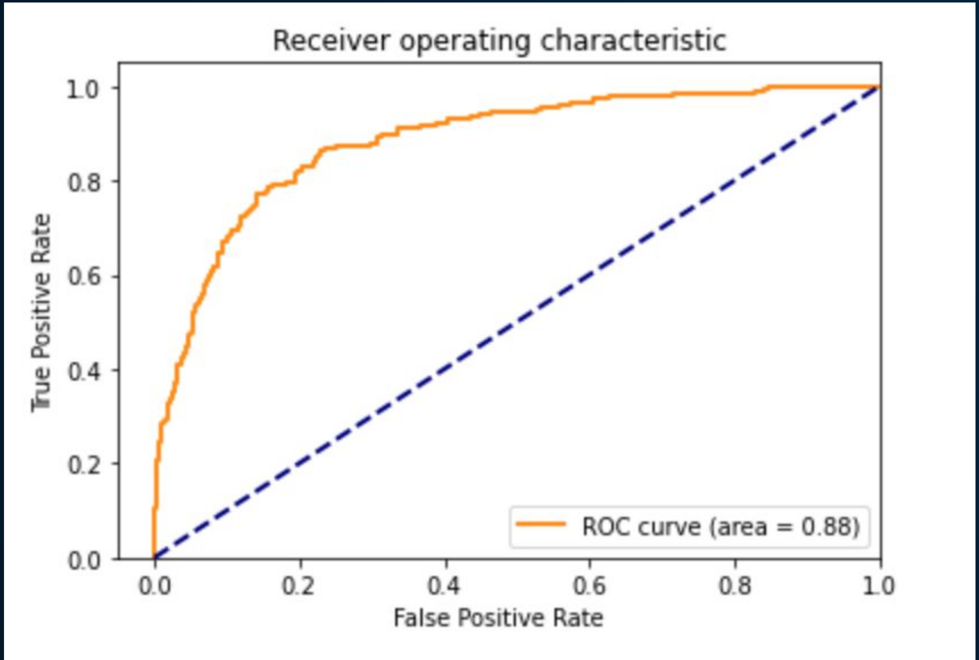
Accuracy=0.97490

Weighted F1-score=0.97
(best value at 1;
worst value at 0)

AUC=0.9028123

PRELIMINARY RESULTS

RANDOM FOREST CLASSIFIER



← ROC Curve

AUC=0.8807784

CONCLUSIONS

- **Predictive Research**
  - Find the best algorithm to detect fake job posting online
- **Four models**
  - used (1) Logistic Regression, (2) Multinomial Naive Bayes Classifier, (3) SVC, and (4) Random Forest
  - Further compare the predictive performance by evaluating the accuracy, f1 score, AUC of each model
- **Future Discussion**
  - Method for detecting topics from unlabelled data (topic modelling, clustering, similarity)

THANKS!

Q&A SECTION