CS839 Project Stage 1 Report

Member Infomation

- Guangtong (Bruce) Bai (`guangtong.bai@wisc.edu`)
- Dan (Joanna) Jiang (`dan.jiang@wisc.edu`)
- Changtian (Chante) Sun (`changtiansun@cs.wisc.edu`)

Entity Description

Person names are marked up using a pair of curly braces, i.e. `{person}` .

For a person name consisting of several parts, only the longest name will be marked.

Example 1

> When he meets {Marla} ({Helena Bonham Carter}), another fake attendee of support groups, his life seems to become a little more bearable. However when he associates himself with {Tyler} ({Brad Pitt}) he is dragged into an underground fight club and soap making scheme.

Example 2

> {William Wallace} is a Scottish rebel who leads an uprising against the cruel English ruler {Edward} the Longshanks, who wishes to inherit the crown of Scotland for himself.

Number of Mentions

- Total number of mentions: *2377*
- Number of mentions in set I: *1366*
- Number of mentions in set J: *770*

Model Selection with Cross Validation (Classifier M)

For our task, we consider the following models:

Classifier	Class Name in `scikit-learn` Package
Linear Regression	`RidgeClassifier`
Logistic Regression	`LogisticRegression`
SVM	`LinearSVC`
Naive Bayes	`GaussianNB`
Decision Tree	`DecisionTreeClassifier`
Random Forest	`RandomForestClassifier`

After conducting Cross Validation on set I, we selected `RandomForestClassifier` *the first time*, as it obtained the highest F1 score. All its scores are:

- Precision: *75.94%*
- Recall: *76.50%*
- F1: *75.97%*

For reference, below are the detailed outputs of each fold and the final averaged result:

```
================================================================
Metrics                      Precision(%)   Recall(%)      F1(%)
----------------------------Fold 1-------------------------------
RidgeClassifier              79.81          48.82          60.58
LogisticRegression           78.26          74.12          76.13
LinearSVC                    77.85          72.35          75.00
GaussianNB                   45.43          93.53          61.15
DecisionTreeClassifier       75.95          70.59          73.17
RandomForestClassifier       75.93          72.35          74.10
----------------------------Fold 2-------------------------------
RidgeClassifier              89.61          48.25          62.73
LogisticRegression           85.59          70.63          77.39
LinearSVC                    85.34          69.23          76.45
GaussianNB                   51.13          95.10          66.50
DecisionTreeClassifier       81.51          67.83          74.05
RandomForestClassifier       80.92          74.13          77.37
----------------------------Fold 3-------------------------------
RidgeClassifier              74.77          51.61          61.07
LogisticRegression           71.35          78.71          74.85
LinearSVC                    71.76          78.71          75.08
GaussianNB                   40.63          99.35          57.68
DecisionTreeClassifier       71.26          76.77          73.91
```

```
RandomForestClassifier        67.38          81.29          73.68
------------------------------Fold 4--------------------------------
RidgeClassifier               67.07          39.01          49.33
LogisticRegression            66.67          69.50          68.06
LinearSVC                     65.97          67.38          66.67
GaussianNB                    37.30          100.00         54.34
DecisionTreeClassifier        62.87          74.47          68.18
RandomForestClassifier        64.12          77.30          70.10
------------------------------Fold 5--------------------------------
RidgeClassifier               85.87          55.24          67.23
LogisticRegression            88.33          74.13          80.61
LinearSVC                     89.92          74.83          81.68
GaussianNB                    49.13          98.60          65.58
DecisionTreeClassifier        80.58          78.32          79.43
RandomForestClassifier        81.16          78.32          79.72
------------------------------Fold 6--------------------------------
RidgeClassifier               70.59          39.67          50.79
LogisticRegression            72.97          66.94          69.83
LinearSVC                     72.97          66.94          69.83
GaussianNB                    37.42          95.87          53.83
DecisionTreeClassifier        69.12          77.69          73.15
RandomForestClassifier        71.90          71.90          71.90
------------------------------Fold 7--------------------------------
RidgeClassifier               92.73          46.79          62.20
LogisticRegression            81.05          70.64          75.49
LinearSVC                     81.63          73.39          77.29
GaussianNB                    39.25          95.41          55.61
DecisionTreeClassifier        82.61          69.72          75.62
RandomForestClassifier        81.25          71.56          76.10
------------------------------Fold 8--------------------------------
RidgeClassifier               84.27          67.57          75.00
LogisticRegression            85.58          80.18          82.79
LinearSVC                     84.26          81.98          83.11
GaussianNB                    40.07          100.00         57.22
DecisionTreeClassifier        73.28          86.49          79.34
RandomForestClassifier        75.97          88.29          81.67
------------------------------Fold 9--------------------------------
RidgeClassifier               89.06          44.53          59.38
LogisticRegression            85.29          67.97          75.65
LinearSVC                     85.29          67.97          75.65
GaussianNB                    45.04          99.22          61.95
DecisionTreeClassifier        76.42          73.44          74.90
RandomForestClassifier        77.42          75.00          76.19
------------------------------Fold 10-------------------------------
RidgeClassifier               89.09          33.33          48.51
```

```
LogisticRegression            82.93          69.39          75.56
LinearSVC                     83.61          69.39          75.84
GaussianNB                    51.99          97.96          67.92
DecisionTreeClassifier        82.26          69.39          75.28
RandomForestClassifier        83.33          74.83          78.85
--------------------------Mean Score-----------------------------
RidgeClassifier               82.29          47.48          59.68
LogisticRegression            79.80          72.22          75.64
LinearSVC                     79.86          72.22          75.66
GaussianNB                    43.74          97.50          60.18
DecisionTreeClassifier        75.59          74.47          74.70
RandomForestClassifier        75.94          76.50          75.97
================================================================
```

Model Evaluation before Post-processing (Classifier X)

We finally settled with **`RandomForestClassifier`**. The scores it obtained on set J are:

- Precision: *93.22%*
- Recall: *89.22%*
- F1: *91.17%*

For reference, here is the performance of all the models:

```
================================================================
Metrics                       Precision(%)   Recall(%)      F1(%)
----------------------------------------------------------------
RidgeClassifier               82.81          48.18          60.92
LogisticRegression            91.18          90.00          90.59
LinearSVC                     90.48          90.13          90.31
GaussianNB                    61.46          97.14          75.29
DecisionTreeClassifier        89.99          87.53          88.74
RandomForestClassifier        93.22          89.22          91.17
================================================================
```

Ruled-based Post-processing

As our classifier X has already met the requirements for Precision and Recall, we did not have any ruled-based post-processing.

Final Scores (Classifier Y)

Without any post-processing, our Classifier Y is just Classifier X. The final scores we achieved are:

- Precision: *93.22%*
- Recall: *89.22%*
- F1: *91.17%*