CISC7107 Data Mining and Decision Support Systems

Assignment #3

**Time-series Forecasting**

**Text Mining**

Student Name: Justice Abban

**Text Mining**

**Spam Text Message Classification**

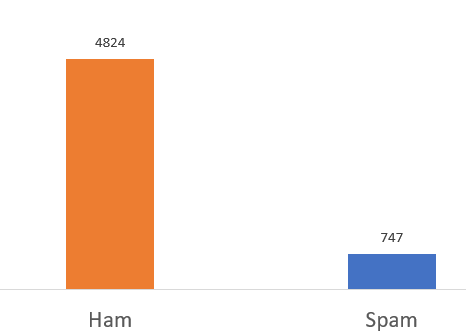
1. Data Introduction

The dataset is retrieved from Kraggle.com, which classified message into “Spam” and “Ham”.

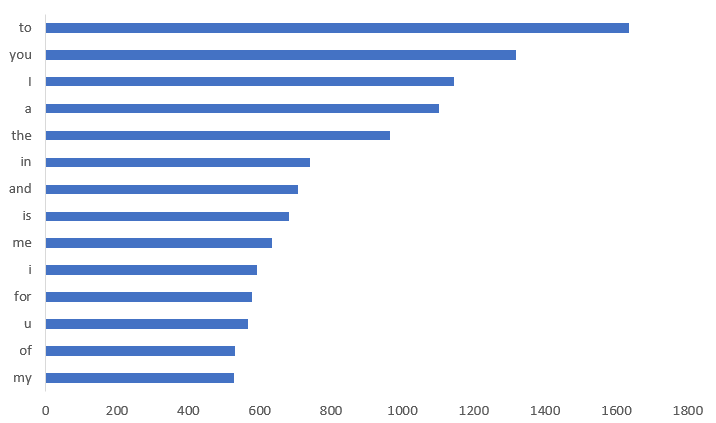
1. Method

Using *StringToWordVector* to split message content and J48 algorithm to build a model, as well as testing data by using filterclassifier algorithm.

1. Data Preview



Total 5571 samples from the dataset, 4824 Ham and 747 Spam. Spam is the target.



The most common words in messages sorted by frequency.

1. Performance Evaluation



Using J48 algorithm and StringToWordVector, the accuracy of “Spam” is 82.9% and the overall accuracy is 97.42%. After optimization, the accuracy increases from 82.9% to 89% for “Spam”. The overall accuracy and Kappa also improve. On the other hand, the result for testing data does not change.

**Apple Twitter Sentiment Texts**

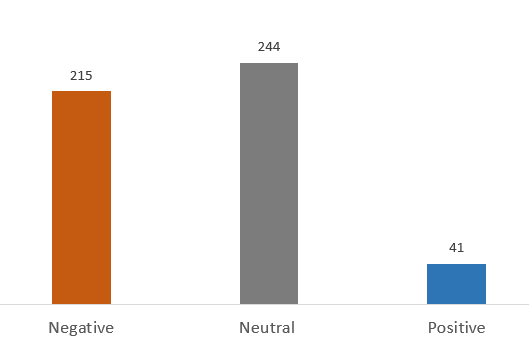
1. Data Introduction

The dataset is retrieved from Kraggle.com. The data shows some twits about Apple products. The classification includes three types, positive, neutral and negative twits.

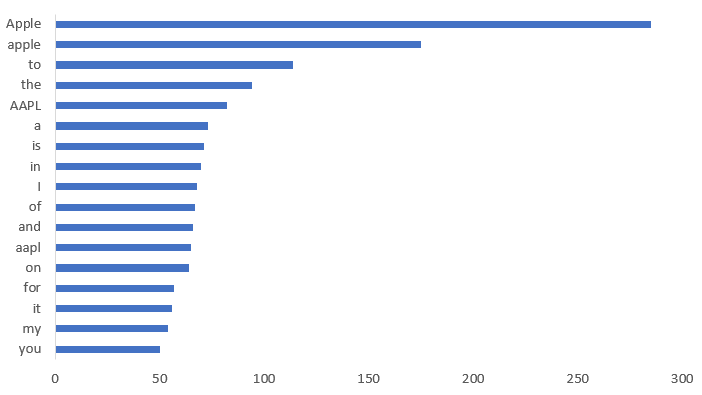
1. Method

Using *StringToWordVector* to split message content and J48 algorithm to build a model, as well as testing data by using filterclassifier algorithm.

1. Data Preview



Total 500 samples selected, 215 negative, 244 neutral and 41 positives. Positive twits is the target.

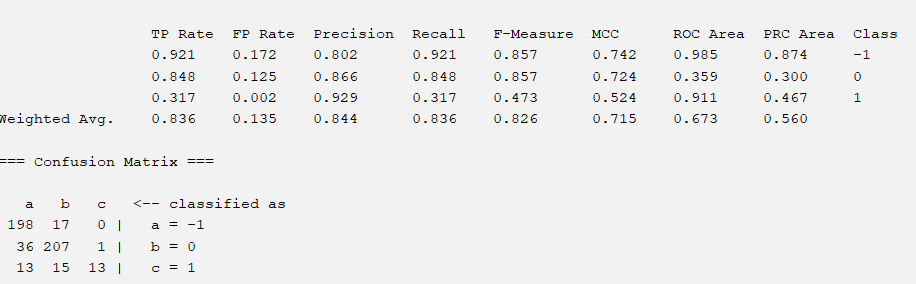


The most common words in twits by frequency.

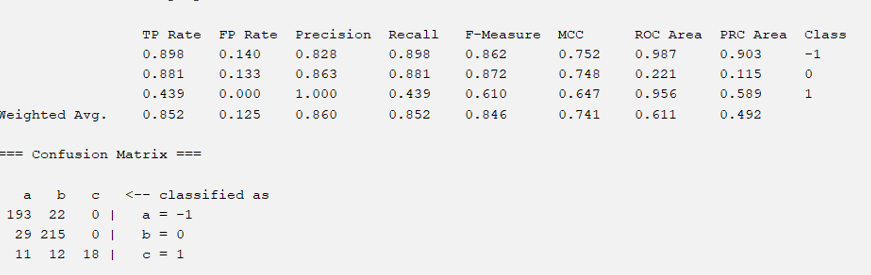
1. Performance Evaluation



The positive feedback is not much in the dataset.



Only 31.7% accuracy for the positive twits and the overall accuracy is 83.6% in the original preprocessing. After optimization in preprocessing, the accuracy% of positive feedback is improved by about 12%, while before or after optimization, the accuracy% of positive feedback in testing data is 0. It seems the training model needs more data and the data is slightly imbalanced.



**Time-series Forecasting**

**Wind Energy in Germany**

1. Data Introduction

The dataset is retrieved from Kaggle.com. The goal is to forecast the wind production. It is a time series dataset containing measurements of daily temperature, wind energy generation and wind capacity from 2017 to 2019.

1. Methods

Apply RandomForest, LinearRegression, MultiplayerPerception and SMOoreg to forecast wind production. MAE and RMSE will be used as comparison among those algorithms.

1. Data Preview

It seems wind energy generation is higher in winter than in summers.

The wind capacity is increasing over time.

The highest temperature can be 28.24 in summer and the lowest can be -9.363 in winter.

|  |  |  |  |
| --- | --- | --- | --- |
| **Correlation** |  |  |  |
| Wind Gen | 1.00 |  |  |
| Wind Cap | 0.13 | 1.00 |  |
| Temp | -0.33 | 0.11 | 1.00 |
|  | Wind Gen | Wind Cap | Temp |

The correlation between Wind Generation and Temperature is negative effect -0.33, while Wind Capacity and Wind Generation is positive effect. Therefore, wind capacity is slightly offset by the temperature.

1. Performance Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| **Wind Energy Generation** | | |  |
|  | **Mag.** | **MAE** | **RMSE** |
| Liner Regression | 637607.40 | 110773.58 | 144064.67 |
| Multilayer Perception | 725871.38 | 109633.82 | 144654.36 |
| SMOreg | 645825.28 | 107898.38 | 146922.39 |
| Random Forest | 552054.97 | 44555.67 | 57726.18 |
|  |  |  |  |
|  |  |  |  |
| **Wind Capacity** | |  |  |
|  | **Mag.** | **MAE** | **RMSE** |
| Liner Regression | 50436.23 | 67.72 | 92.54 |
| Multilayer Perception | 50418.28 | 30.80 | 39.18 |
| SMOreg | 50457.61 | 15.64 | 22.06 |
| Random Forest | 50246.04 | 39.93 | 52.47 |
|  |  |  |  |
| **Temperature** |  |  |  |
|  | **Mag.** | **MAE** | **RMSE** |
| Liner Regression | 1.52 | 1.19 | 1.50 |
| Multilayer Perception | 2.14 | 1.23 | 1.58 |
| SMOreg | 1.25 | 1.18 | 1.53 |
| Random Forest | 1.50 | 0.57 | 0.71 |

The lowest MAE and RMSE of Wind Energy Generation and Temperature is applying Random Forest algorithm, while for Wind Capacity is SMOreg. Therefore, Random Forest should be the most suitable algorithm.

Random Forest

