ICS-490-01: Special Topics I (T171)

Dr. Lahouari Ghouti

Sender’s Dataset Report

*Group Members*

|  |  |
| --- | --- |
| **Ibrahim BinAlshikh** | 201139750 |
| **Ibrahim Beladi** | 201224780 |
| **Mohammed Alfarraj** | 201229820 |
| **Omar Al-Dakhil** | 201381890 |

**Abstract**

Hand-curated Twitter sentiment dataset published by Sander’s Lab. It contains tweets from 2007 to 2011 that mention one of four major Tech companies. Sander’s Lab manually assigned labels for each tweet as either “Positive”, “Negative”, “Neutral”, or “Irrelevant” . “Positive” and “Negative” indicated whether or not the tweet showed a generally favorable or unfavorable opinion toward the mentioned company. We had to get some important words and then finally used them on classifiers.

**Table of Contents**

[**Toolbox**](#_sws21nhv82yg) **2**

[**Machine Learning Solution**](#_rk1fwuts7ar6) **2**

[Algorithm Description](#_72h0tvrln01c) 2

[Performance Measures Used](#_iy7cngvgim3b) 2

[**Exploratory Data Analysis**](#_cve23fyjoahi) **2**

[**Procedure**](#_egh42fiiaagr) **3**

[Data Cleaning](#_u9sbhvntkclx) 3

[Features Extraction](#_bxzhd023fw6g) 3

[Classifiers](#_mdldmyj3nz8c) 3

[**Performance Results**](#_gjdgxs) **4**

[Support Vector Machines Classifier](#_787dbhj5p9w8) 4

[Multinomial Naïve Bayes Classifier](#_kewudx1cvhem) 5

[Random Forest Classifier](#_w00qfzbx8tal) 6

[Logistic Regression Classifier](#_b3qqk0wdnidd) 7

[Voting Classifier](#_yziuqjhj1krc) 8

[**Conclusion**](#_6536oncmayky) **9**

# 

# 

# 

# Toolbox

* Sander’s Lab
* Scikit
* NumPy
* Jupyter

# 

# Machine Learning Solution

## Algorithm Description

* Support Vector Machines.
* Random Forests.
* Multinomial Naïve Bayes.
* Logistic Regression Classifier

## Performance Measures Used

* Accuracy
* Classification report
* Confusion Matrix

# Exploratory Data Analysis

There are more than 53000 reviews, and these are divided into 2 classes. They are balanced, the number of each class is too close to the other.

# Procedure

## Data Gathering

The data is available by Sander’s Lab. And it contains the sentiment classes for the reviews.

## Data Cleaning

* Removing Not/Applicable records (N/A).
* Removing the punctuation.

## Features Extraction

* Inserting 1 (bias) in the features matrix.
* Finding whether the important word in the tweet and then counting its occurance.
  + Both the existence and the count are features.
* Saving the features matrix as a file.

## Classifiers

* Running the classifiers with the exported features matrix.
* Feeding the classifiers model to an ensemble voting classifier.

# 

# 

# Performance Results

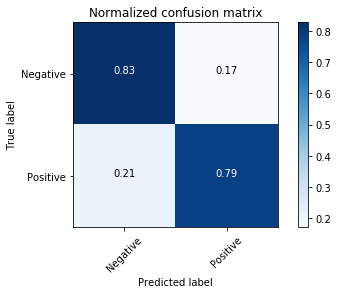
## Support Vector Machines Classifier

Accuracy: 80.46%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.79 | 0.83 | 0.81 | 2693 |
| Positive | 0.82 | 0.78 | 0.80 | 2615 |
| Avg / Total | 0.81 | 0.80 | 0.80 | 5308 |

Confusion matrix:



## 

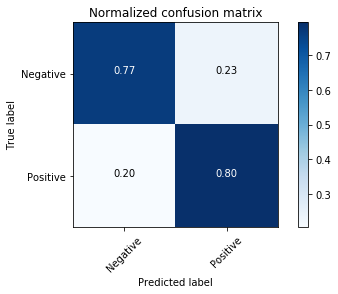
## Multinomial Naïve Bayes Classifier

Accuracy: 78.5%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.80 | 0.77 | 0.78 | 2693 |
| Positive | 0.77 | 0.80 | 0.79 | 2615 |
| Avg / Total | 0.79 | 0.79 | 0.79 | 5308 |

Confusion matrix:

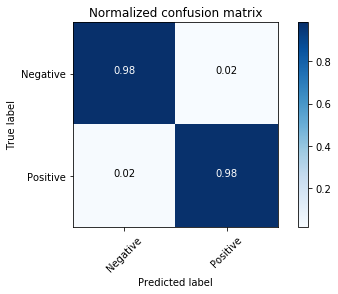


## Random Forest Classifier

Accuracy: 98.34%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.98 | 0.98 | 0.98 | 2693 |
| Positive | 0.98 | 0.98 | 0.98 | 2615 |
| Avg / Total | 0.98 | 0.98 | 0.98 | 5308 |

Confusion matrix:  


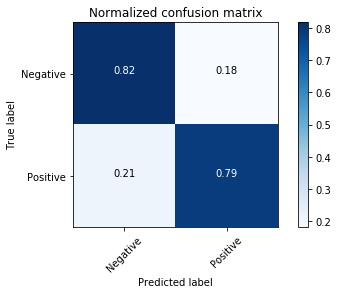
## Logistic Regression Classifier

Accuracy: 79.9%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.80 | 0.82 | 0.81 | 2693 |
| Positive | 0.81 | 0.79 | 0.80 | 2615 |
| Avg / Total | 0.80 | 0.80 | 0.80 | 5308 |

Confusion matrix:



## Voting Classifier

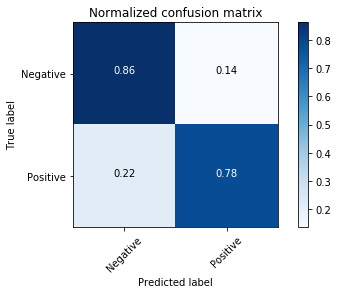
Ensembling the earlier classifiers

Accuracy: 82.23%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.80 | 0.86 | 0.83 | 2693 |
| Positive | 0.85 | 0.78 | 0.81 | 2615 |
| Avg / Total | 0.82 | 0.82 | 0.82 | 5308 |

Confusion matrix:



# Conclusion

Since the data is too big, we ran the feature extraction and the classifiers independently. And hence, we have a csv file containing the features. Furthermore, there is no best algorithm for all datasets, but for this dataset the best classifier algorithm is Random Forest Classifier algorithm.