**MTech EBAC New Media and Sentiment Mining**

# Sentiment Classification Worksheet #2

**Group Number: 15**

**Student Names:** Ethiraj Srinivasan

Naitik Shukla

Raghavan Kalyanasundaram

Shobhit Sandeep Jaipurkar

Sindhu Rumesh Kumar

Suya Hao

1. Build a model from “train.csv” to classify text into two classes, “Positive” and “Negative”. Test your model on “test.csv”. You can try different approaches to get your best model. ***Report your best model’s performance: overall accuracy, confusion matrix, precision/recall/F1 for each class as well as macro-average for overall.***

**Solution:**

The accuracy results of the models tried have been mentioned below through a Confusion Matrix for each of them. Out of the 5 models applied, SVM gave the best accuracy of 89.61%.

**Multinomial Naïve Bayes Classifier**

|  |  |  |  |
| --- | --- | --- | --- |
|  | PREDCITED | | |
|  |  | Negative | Positive |
| ACTUAL | Negative | 2861 | 578 |
| Positive | 205 | 1422 |

**Accuracy: 0.8454**

**Maximum Entropy**

|  |  |  |  |
| --- | --- | --- | --- |
|  | PREDCITED | | |
|  |  | Negative | Positive |
| ACTUAL | Negative | 185 | 16 |
| Positive | 2881 | 1984 |

**Accuracy: 0.4281**

**KNN**

|  |  |  |  |
| --- | --- | --- | --- |
|  | PREDCITED | | |
|  |  | Negative | Positive |
| ACTUAL | Negative | 2536 | 395 |
| Positive | 530 | 1605 |

**Accuracy: 0.817**

**Decision Tree**

|  |  |  |  |
| --- | --- | --- | --- |
|  | PREDCITED | | |
|  |  | Negative | Positive |
| ACTUAL | Negative | 2248 | 336 |
| Positive | 818 | 1664 |

**Accuracy: 0.7722**

**Support Vector Machine**

|  |  |  |  |
| --- | --- | --- | --- |
|  | PREDCITED | | |
|  |  | Negative | Positive |
| ACTUAL | Negative | 2738 | 198 |
| Positive | 328 | 1802 |

**Accuracy: 0.8961**

|  |  |  |
| --- | --- | --- |
|  | **Positive Sentiment** | **Negative Sentiment** |
| **Precision** | 0.90 | 0.89 |
| **Recall** | 0.84 | 0.93 |
| **F1** | 0.87 | 0.91 |
| **Macro Average (Precision)** | 0.895 | |
| **Macro Average (Recall)** | 0.89 | |

1. What are the words linked more to negative reviews? And those for positive ones? Find the most informative words found in Step 1.

***List top 5 words for positive and negative reviews respectively.***

**Solution:**

Most informative Negative Words:

|  |  |
| --- | --- |
| Negative Words | -1:+1 ratio |
| Worst | 32.7:1 |
| Awful | 28.3:1 |
| Disgusting | 24.1:1 |
| Stone | 21.5:1 |
| Horrible | 19.6:1 |

Most informative Positive Words:

|  |  |
| --- | --- |
| Positive Words | +1:-1 ratio |
| Eastway | 13.8:1 |
| pupusas | 13.8:1 |
| sketchy | 12.8:1 |
| Asu | 11.8:1 |
| Exterior | 11.8:1 |

The positive words are basically nouns and do not contribute to sentiment. Hence the frequency of individual words in the documents were analyzed.

Words like “food”, “good”, “like”, “place” are found more frequently in both negative and positive documents. So we use **bigrams** for positive and **trigrams** for negative documents.

* **Most Informative Bigrams for Positive Documents:**
* great food
* really good
* happy hour
* great place
* great service
* **Most Informative Trigrams for Negative Documents:**
* not go back
* would not recommend
* heart attack grill
* go somewhere else

1. Please refer to the attached files