Opinion Mining

SUBMITTED BY – ARKYA BAGCHI

Opinion Mining is the task of extracting opinions from a sentence on an instance. In these project, I have used the data from twitter and had proposed a methodology to judge the opinion of various tweets into three different classes "Offensive language", "Hate speech" and "Neutral speech" , using machine learning algorithm such as Naive Bayes and Random Forest classifier and the precision, recall and F-measure are computed.

I) Introduction

The data is stored as a CSV file and each data file contains 5 columns:

Count

number of CrowdFlower users who coded each tweet (min is 3, sometimes more users coded a tweet when judgments were determined to be unreliable by CF).

hate-speech
number of CF users who judged the tweet to be hate speech.

Offensive-language
number of CF users who judged the tweet to be offensive.

neither
number of CF users who judged the tweet to be neither offensive nor non-offensive.

Class
    0- Hate speech
    1- Offensive Language

2- Neither

3-

II) Methodology

In the proposed work, Opinion mining is mainly done in three stages.

a) Extracting data - Extracted the Tweets column
b) Processing data - Quotations extracted are processed before classifying by removing the stop-words and then stemming/lemmatizing the quotations.

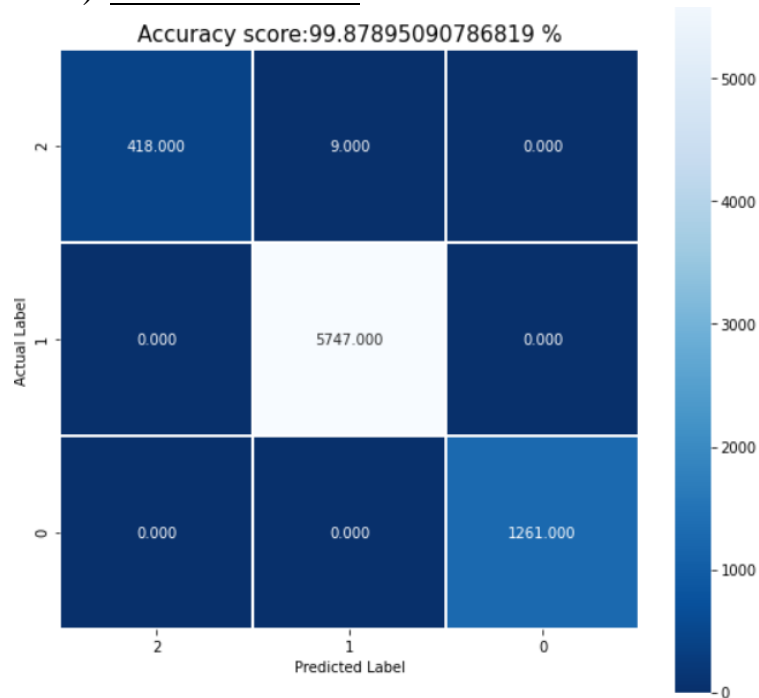c) Analysis of data – classification performed using Naive-bayes, Random Forest and SVM classifier
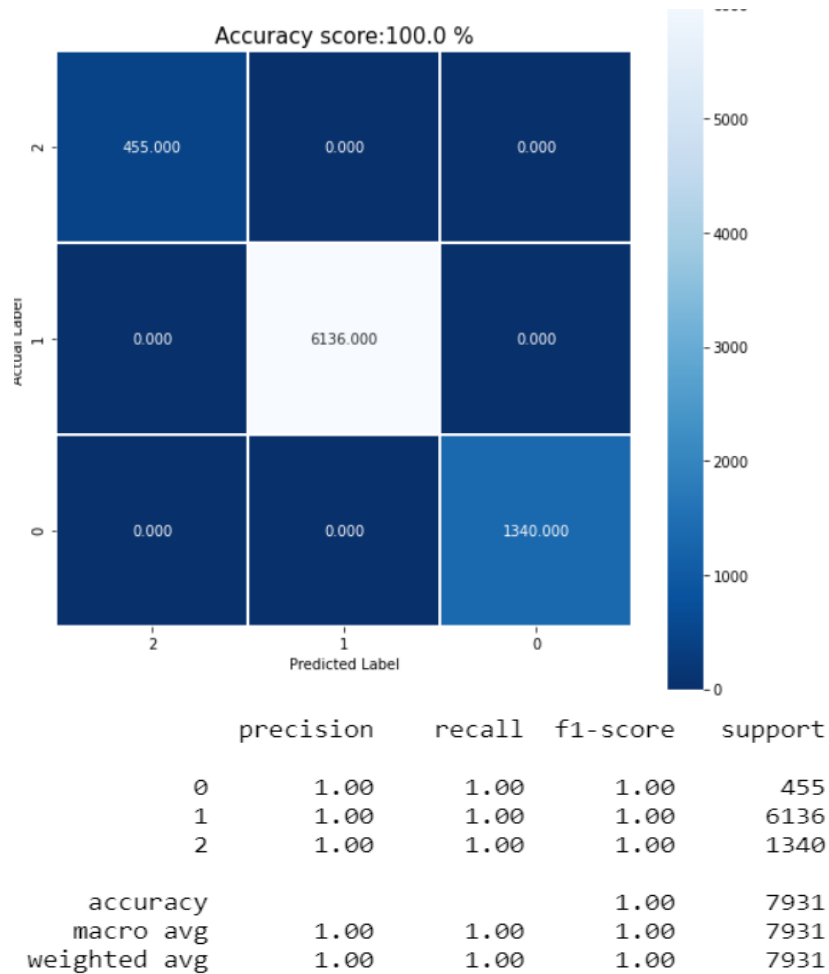
III) Scores

1) Multinomial Naïve-Bayes

	precision	recall	f1-score	support
0	0.80	0.73	0.76	427
1	0.98	0.98	0.98	5747
2	0.96	0.99	0.97	1261
accuracy			0.97	7435
macro avg	0.91	0.90	0.90	7435
weighted avg	0.96	0.97	0.97	7435

2) Random-Forest

Accuracy score:99.87895090786819 %

	precision	recall	f1-score	support
0	1.00	0.98	0.99	427
1	1.00	1.00	1.00	5747
2	1.00	1.00	1.00	1261
accuracy			1.00	7435
macro avg	1.00	0.99	1.00	7435
weighted avg	1.00	1.00	1.00	7435

3) Support Vector Machine(SVM)

Accuracy score:100.0 %

```
                precision    recall   f1-score   support

            0        1.00      1.00       1.00        455
            1        1.00      1.00       1.00       6136
            2        1.00      1.00       1.00       1340

     accuracy                             1.00       7931
    macro avg        1.00      1.00       1.00       7931
 weighted avg        1.00      1.00       1.00       7931
```

IV) Conclusion

We have achieved the highest accuracy value from the support vector machine model. In the present work we have only considered the tweet column as our feature data further this could be extended to some more attributes.