# Report

|  |  |
| --- | --- |
| Figure 1 Topic Frequency Chart | Figure 2 Sentiment Frequency Chart |

1. (1 mark) Give simple descriptive statistics showing the frequency distributions for the sentiment and topic classes across the full dataset. What do you notice about the distribution?

Topic statistics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | N\* | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
| Frequency | 20 | 0 | 100.0 | 20.8 | 93.0 | 7.0 | 26.0 | 57.5 | 157.3 | 358.0 |

Sentiment Statistics

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | N\* | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
| Count | 3 | 0 | 667 | 334 | 579 | 153 | 153 | 553 | 1294 | 1294 |

From the figures and tables above, many tweets were talking about topic 10003, economic management and were negative tweets

1. (2 marks) Vary the number of words from the vocabulary used as training features for the standard methods (e.g. the top *N* words for *N* = 100, 200, etc.). Show metrics calculated on both the training set and the test set. Explain any difference in performance of the models between training and test set, and comment on metrics and runtimes in relation to the number of features.

N – Number of feature words  
A – Accuracy from *sklearn.metrics.accuracy\_score*  
D – Data used as test sets [train, test]  
P\_MI – Precision score micro average  
P\_MA – Precision score macro average  
P\_W – Precision score weighted average  
R\_MI – Precision score micro average  
R\_MA – Precision score micro average  
R\_W – Precision score micro average  
F1\_MI – Precision score micro average  
F1\_MA – Precision score micro average  
F1\_W – Precision score micro average  
R – Runtime in seconds

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | N | A | D | P\_MI | P\_MA | P\_W | R\_MI | R\_MA | R\_W | F1\_MI | F1\_MA | F1\_W | Run  time |
| DT\_topics | 100 | 0.46 | Train | 0.46 | 0.41 | 0.47 | 0.46 | 0.30 | 0.46 | 0.46 | 0.31 | 0.44 | 0.0557 |
| DT\_topics | 100 | 0.25 | Test | 0.25 | 0.14 | 0.23 | 0.25 | 0.13 | 0.25 | 0.25 | 0.13 | 0.23 | 0.0520 |
| DT\_topics | 200 | 0.51 | Train | 0.51 | 0.43 | 0.51 | 0.51 | 0.32 | 0.51 | 0.51 | 0.33 | 0.48 | 0.0999 |
| DT\_topics | 200 | 0.31 | Test | 0.31 | 0.18 | 0.28 | 0.31 | 0.19 | 0.31 | 0.31 | 0.18 | 0.28 | 0.0800 |
| DT\_sentiment | 100 | 0.83 | Train | 0.83 | 0.76 | 0.82 | 0.83 | 0.71 | 0.83 | 0.83 | 0.73 | 0.82 | 0.0520 |
| DT\_sentiment | 100 | 0.67 | Test | 0.67 | 0.51 | 0.65 | 0.67 | 0.47 | 0.67 | 0.67 | 0.47 | 0.65 | 0.0420 |
| DT\_sentiment | 200 | 0.86 | Train | 0.86 | 0.82 | 0.85 | 0.86 | 0.72 | 0.86 | 0.86 | 0.76 | 0.85 | 0.0759 |
| DT\_sentiment | 200 | 0.65 | Test | 0.65 | 0.53 | 0.64 | 0.65 | 0.51 | 0.65 | 0.65 | 0.52 | 0.65 | 0.0560 |
| BNB\_topics | 100 | 0.41 | Train | 0.41 | 0.37 | 0.42 | 0.41 | 0.26 | 0.41 | 0.41 | 0.29 | 0.40 | 0.0550 |
| BNB\_topics | 100 | 0.27 | Test | 0.27 | 0.17 | 0.27 | 0.27 | 0.15 | 0.27 | 0.27 | 0.15 | 0.26 | 0.0450 |
| BNB\_topics | 200 | 0.50 | Train | 0.51 | 0.47 | 0.52 | 0.51 | 0.33 | 0.51 | 0.51 | 0.35 | 0.49 | 0.0700 |
| BNB\_topics | 200 | 0.33 | Test | 0.33 | 0.21 | 0.31 | 0.33 | 0.18 | 0.33 | 0.33 | 0.18 | 0.30 | 0.0620 |
| BNB\_sentiment | 100 | 0.73 | Train | 0.73 | 0.67 | 0.72 | 0.73 | 0.54 | 0.73 | 0.73 | 0.57 | 0.71 | 0.0790 |
| BNB\_sentiment | 100 | 0.70 | Test | 0.70 | 0.57 | 0.67 | 0.70 | 0.50 | 0.70 | 0.70 | 0.52 | 0.68 | 0.0400 |
| BNB\_sentiment | 200 | 0.76 | Train | 0.76 | 0.70 | 0.75 | 0.76 | 0.63 | 0.76 | 0.76 | 0.65 | 0.75 | 0.0840 |
| BNB\_sentiment | 200 | 0.69 | Test | 0.69 | 0.59 | 0.669 | 0.69 | 0.55 | 0.69 | 0.69 | 0.56 | 0.69 | 0.0529 |

1. (2 marks) Evaluate the standard models with respect to baseline predictors (**VADER** for sentiment analysis, majority class for both classifiers). Comment on the performance of the baselines and of the methods relative to the baselines.
2. (2 marks) Evaluate the effect that preprocessing the input features, in particular stop word removal plus Porter stemming as implemented in **NLTK**, has on classifier performance, for the three standard methods for both sentiment and topic classification. Compare results with and without preprocessing on training and test sets and comment on any similarities and differences.
3. (2 marks) Sentiment classification of neutral tweets is notoriously difficult. Repeat the experiments of items 2 (with N = 200), 3 and 4 for sentiment analysis with the standard models using only the positive and negative tweets (i.e. removing neutral tweets from both training and test sets). Compare these results to the previous results. Is there any difference in the metrics for either of the classes (i.e. consider positive and negative classes individually)?
4. (6 marks) Describe your best method for sentiment analysis and your best method for topic classification. Give some experimental results showing how you arrived at your methods. Now provide a brief comparison of your methods in relation to the standard methods and the baselines.

# Appendix A

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 10000 | 0.48 | 0.66 | 0.56 | 190 |
| 10001 | 0.41 | 0.53 | 0.46 | 104 |
| 10002 | 0.43 | 0.54 | 0.48 | 91 |
| 10003 | 0.41 | 0.61 | 0.49 | 288 |
| 10004 | 0 | 0 | 0 | 10 |
| 10005 | 0.61 | 0.6 | 0.6 | 141 |
| 10006 | 0.4 | 0.29 | 0.34 | 146 |
| 10007 | 1 | 0.2 | 0.33 | 5 |
| 10008 | 0.31 | 0.28 | 0.3 | 120 |
| 10009 | 0 | 0 | 0 | 11 |
| 10010 | 0.41 | 0.34 | 0.37 | 41 |
| 10011 | 0 | 0 | 0 | 8 |
| 10012 | 0.67 | 0.1 | 0.17 | 21 |
| 10013 | 0.44 | 0.29 | 0.35 | 72 |
| 10014 | 0.18 | 0.14 | 0.15 | 22 |
| 10015 | 0.89 | 0.71 | 0.79 | 82 |
| 10016 | 0.26 | 0.21 | 0.24 | 42 |
| 10017 | 1 | 0.09 | 0.16 | 35 |
| 10018 | 0 | 0 | 0 | 31 |
| 10019 | 1 | 0.03 | 0.05 | 40 |
|  |  |  |  |  |
| micro avg | 0.45 | 0.45 | 0.45 | 1500 |
| macro avg | 0.45 | 0.28 | 0.29 | 1500 |
| weighted avg | 0.47 | 0.45 | 0.43 | 1500 |

Table 1

Accuracy score: 0.45266666666666666

--- 0.0839850902557373 seconds ---