Precision, Recall and F1 for each class:

|  |  |  |  |
| --- | --- | --- | --- |
| Topic name | Precision | page3image32535488page3image32524736Recall | F1 |
| ARTS CULTURE ENTERTAINMENT | 0.33 | 0.67 | 0.44 |
| BIOGRAPHIES PERSONALITIES PEOPLE | 0.70 | 0.47 | 0.56 |
| DEFENCE | 0.83 | 0.77 | 0.80 |
| DOMESTIC MARKETS | 0.33 | 1.00 | 0.50 |
| FOREX MARKETS | 0.58 | 0.75 | 0.65 |
| HEALTH | 0.69 | 0.64 | 0.67 |
| MONEY MARKETS | 0.60 | 0.67 | 0.63 |
| SCIENCE AND TECHNOLOGY | 0.00 | 0.00 | 0.00 |
| SHARE LISTINGS | 0.60 | 0.86 | 0.71 |
| SPORTS | 0.95 | 0.98 | 0.97 |

Precision, Recall and F1 for recommendations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Topic name | Suggested articles | page3image32646144Precision  (weighted average) | page3image32535488page3image32524736Recall  (weighted average) | F1  (weighted average) |
| ARTS CULTURE ENTERTAINMENT | 9830, 9789, 9703, 9952, 9526, 9604 | 0.11 | 0.33 | 0.17 |
| BIOGRAPHIES PERSONALITIES PEOPLE | 9645, 9940, 9854, 9988, 9933, 9878, 9582, 9797, 9896, 9758 | 0.49 | 0.7 | 0.58 |
| DEFENCE | 9559, 9616, 9576, 9773, 9770, 9987, 9842, 9904, 9607, 9670 | 0.64 | 0.8 | 0.71 |
| DOMESTIC MARKETS | 9994, 9796, 9989, 9640, 9692, 9910 | 0.11 | 0.33 | 0.17 |
| FOREX MARKETS | 9743, 9565, 9944, 9875, 9584, 9718, 9977, 9530, 9894, 9506 | 0.25 | 0.5 | 0.33 |
| HEALTH | 9609, 9661, 9807, 9873, 9937, 9929, 9810, 9947, 9621, 9978 | 0.49 | 0.7 | 0.58 |
| MONEY MARKETS | 9516, 9769, 9618, 9602, 9755, 9765, 9871, 9995, 9707, 9863 | 0.49 | 0.7 | 0.58 |
| SCIENCE AND TECHNOLOGY | 9617, 9982 | 0 | 0 | 0 |
| SHARE LISTINGS | 9518, 9601, 9666, 9667, 9972, 9562, 9867, 9715, 9876, 9999 | 0.36 | 0.6 | 0.45 |
| SPORTS | 9857, 9774, 9787, 9997, 9760, 9620, 9568, 9596, 9922, 9848 | 1 | 1 | 1 |

10-fold Cross Validation is performed on a combination of parameters, the combination that produced the highest mean test score is chosen to be the final model.

The parameters used for cross validation are as follows:

* Alpha (the constant that multiplies the regularization term): 0.1, 0.01, 0.001
* Penalty/regularization term: l2, l1
* Features: all words (every distinct word represents one feature), stopping removed (stopping words such as ‘and’ ‘the’ are removed)
* Feature representation: CountVectorizer (convert article words into vectors of word count features), TfidfVectorizer (convert article words into TF-IDF features)

The above parameters are chosen for the following reasons:

* Alpha: To control the size of the weights, and prevent overfitting, a non-zero alpha is used. As the size of alpha grows, the impact of large weight increases too. However, alpha cannot be too big such that it underfits and under-estimates the weight. Selecting a good alpha is crucial hence it is chosen by cross validation.
* Penalty/regularization term: Compared to l2, l1 encourages the weights to be 0 hence reducing the number of features, (i.e. a sparse solution) such method is preferred when there are many redundant features. Whereas l2 has a non-sparse solution, it also penalises the loss function more with higher weights. Thus, while l1 reduce overfitting it is also relatively weaker in its prediction power. The effectiveness of the methods can be compared through cross validation.
* Features: there is a list of stop\_words in sklearn containing words that may not add value to classification, with the given dataset we want to compare whether removing such words can improve accuracy.
* Feature representation: CountVectorizer simply records the occurences of words whereas TfidfVectorizer balances out the frequency. This mean that we should use TfidfVectorizer when there are words with high occurrences but carry little meaningful information. CountVectorizer, if we want to be more biased in words that are more frequent. Depending on the dataset, we can only determine the better vectorizer by comparing its scores.

The chosen parameters are: alpha=0.1, penalty=l2, features=all words, feature representation=CountVectorizer.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| alpha | class\_weight | penalty | features | Feature representation | Mean test score |
| 0.001 | balanced | l2 | All words | CountVectorizer | 0.754 |
| 0.001 | balanced | l1 | All words | CountVectorizer | 0.66136842 |
| 0.01 | balanced | l2 | All words | CountVectorizer | 0.77073684 |
| 0.01 | balanced | l1 | All words | CountVectorizer | 0.52021053 |
| 0.1 | balanced | l2 | All words | CountVectorizer | 0.75231579 |
| 0.1 | balanced | l1 | All words | CountVectorizer | 0.43157895 |
| 0.001 | balanced | l2 | stopping removed | CountVectorizer | 0.75684211 |
| 0.001 | balanced | l1 | stopping removed | CountVectorizer | 0.66621053 |
| 0.01 | balanced | l2 | stopping removed | CountVectorizer | 0.76926316 |
| 0.01 | balanced | l1 | stopping removed | CountVectorizer | 0.52063158 |
| 0.1 | balanced | l2 | stopping removed | CountVectorizer | 0.75084211 |
| 0.1 | balanced | l1 | stopping removed | CountVectorizer | 0.43273684 |
| 0.001 | balanced | l2 | All words | TfidfVectorizer | 0.76105263 |
| 0.001 | balanced | l1 | All words | TfidfVectorizer | 0.68863158 |
| 0.01 | balanced | l2 | All words | TfidfVectorizer | 0.57410526 |
| 0.01 | balanced | l1 | All words | TfidfVectorizer | 0.49831579 |
| 0.1 | balanced | l2 | All words | TfidfVectorizer | 0.49831579 |
| 0.1 | balanced | l1 | All words | TfidfVectorizer | 0.49831579 |
| 0.001 | balanced | l2 | stopping removed | TfidfVectorizer | 0.76063158 |
| 0.001 | balanced | l1 | stopping removed | TfidfVectorizer | 0.68494737 |
| 0.01 | balanced | l2 | stopping removed | TfidfVectorizer | 0.57221053 |
| 0.01 | balanced | l1 | stopping removed | TfidfVectorizer | 0.49831579 |
| 0.1 | balanced | l2 | stopping removed | TfidfVectorizer | 0.49831579 |
| 0.1 | balanced | l1 | stopping removed | TfidfVectorizer | 0.49831579 |