COMP9417 Project Report

Group Name: BugMaker

Group Members:

Shuonan Wang Z5158229

Xin Cai Z5108236

Yukang Yan Z5158298

Yunfan Wang Z5171928

Zhihan Qin Z5290141

1. **Introduction**
   1. **Aim**

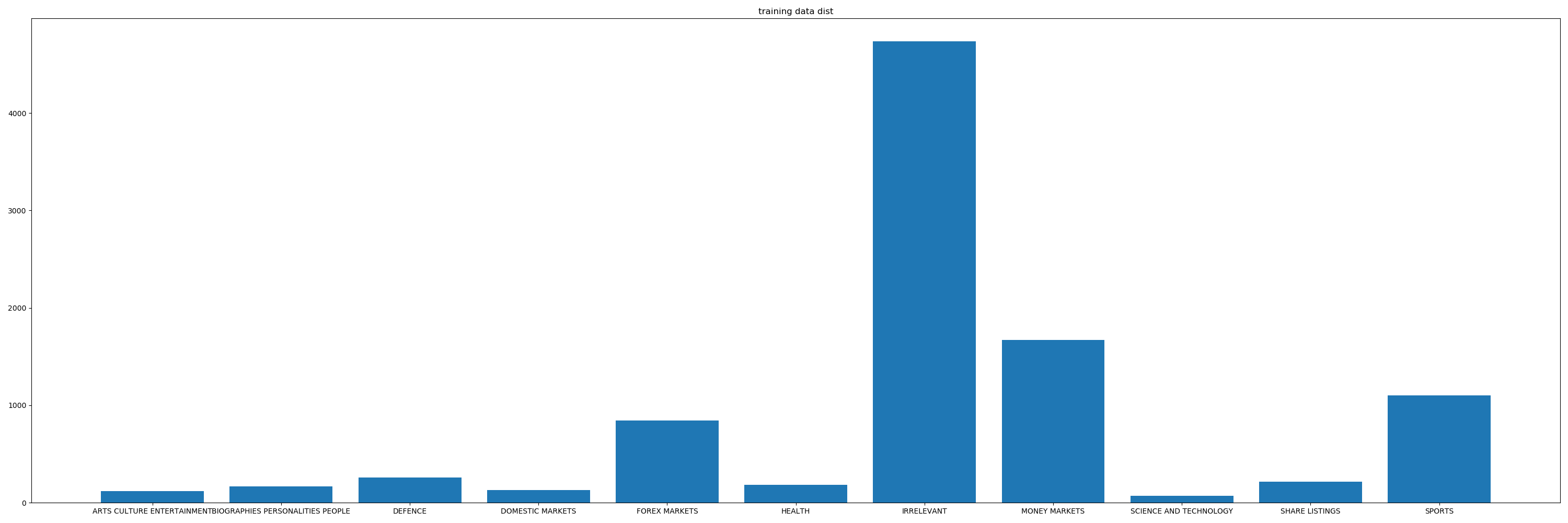
This project aims to apply machine learning algorithms to predict specific outputs in a dataset which is a set of pre-processed articles form 10 different topics. The final output will be the recommendation list of the top 10 articles for each topic from the test dataset at most. The project focuses on 2 main parts -- feature extraction models and classification models. Details about the model selection and hyper-parameter tuning will be presented in the next chapter.

* 1. **Dataset**

**Figure 1.1** demonstrates the distribution of each topic in the training set. There are 9500 articles in total, 49.93% of which are irrelevant articles. Besides the distribution of 10 targets is also imbalanced. **Figure 1.2** shows the word numbers in each article which is categorized by topic and most articles are around 1000 words.

Based on the above observations, we mainly pay attention to the scale and weight of features, which are the words in the training set and also anticipate that those methods which give more weights on rare topics will achieve better performance. To validate our training result, we split the training set into 2 parts -- new training set with 9000 articles and validation set with 500 articles on the basis of the portion.

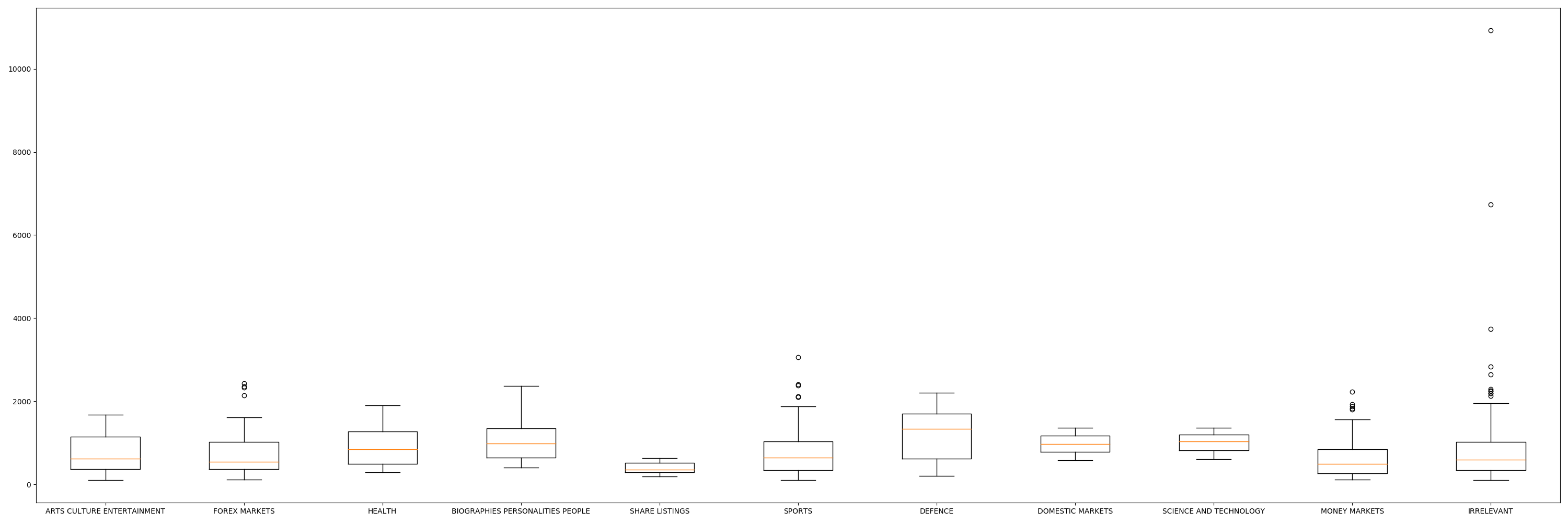
Additionally, based on the fact that the test set only has 500 articles probably with an extremely nonuniform distribution，we decide to use the predict probability matrix with a settled threshold to determine the recommended articles.



**Figure 1.1**

**Table 1.1**

|  |  |
| --- | --- |
| **Topic name** | **Articles numbers** |
| ARTS CULTURE ENTERTAINMENT | 117 |
| BIOGRAPHIES PERSONALITIES PEOPLE | 167 |
| DEFENCE | 258 |
| DOMESTIC MARKETS | 133 |
| FOREX MARKETS | 845 |
| HEALTH | 183 |
| MONEY MARKETS | 1673 |
| SCIENCE AND TECHNOLOGY | 70 |
| SHARE LISTINGS | 218 |
| SPORTS | 1102 |
| Irrelevant | 4734 |



**Figure 1.2**

1. **Methods**
   1. **Method choice**

To determine the extraction and classification model, the assignment splits the process into 2 main steps. First, all models are selected and compared under the policy of controlling variable. We compare extraction models with only using logistic regression classifier and compare different classifiers with document frequency (DF) extraction model. All hyper-parameters are set with a relatively good performance during the initial test phase. The reason to choose those 2 models is that their computation speed is fast, and they can achieve a relatively satisfying result at this stage. At the second step, 2 best extraction models and 2 best classifiers will be chosen for further adjustment.

Brief introduction of all methods tried and the process to determine the final model will be illustrated in this chapter.

* + 1. **Feature extraction choice**

This part will introduce all feature extraction methods we compared and detailed comparing result will show in **Chapter 3.4**

**Df**: Df is a basic method which counts the frequency of words in each article and form a word-frequency matrix. This method is usually a basis of many advanced feature extraction choice. The method usually tends to overweight the words with more frequency, however, considering the fact that some topic in the training set only have a small portion, this method performed well among all these choices.

**PCA**: Principal component analysis is helped to decrease dimensions of features. This method has subtle increase in some cases, however the process is very time-consuming, so we did not deploy this method eventually.

**TF-IDF**: TF-IDF is a commonly used weighting technique for data mining, TF means term frequency and IDF means inverse document frequency. IDF represents the overall relevance of the word to all documents. As a result, we found if a word appears more in a certain topic and less in others, IDF value would be relatively large. The performance is not that good as we expected, f1 score is lower than other extraction methods. This result might because if a feature only appears in a large number of individual texts in a certain topic and rarely occurs in others, then these individual texts are not excluded as special cases in this topic. Therefore, such feature items are not representative.

TF-IDF method based on variance: As we mentioned before, TF-IDF is weak at selecting some feature which takes possession of the most category discernibility, therefore, we designed a TF-IDF method based on variance, the weight value of the feature word in document is weight(, ) = TF-IDF \* var().

**Chi-square**: it calculates the actual deviation of the actual value and the theoretical value to verify the correctness of the theory. The disadvantage might be that it only counts whether a word appears in a document, regardless of how many times it appears, this would exaggerate the role of low-frequency words. However, in our comparison, it has a higher f1 score than other methods.

After comparing four of these feature extraction methods, we finally apply chi-square to train the model.

* + 1. **Classifier choice**

The main classifiers tested are from the lecture and result comparison is in **Chapter 3.3**.

**KNN algorithm**: KNN algorithm finds the k nearest neighbors of x among training instances and scores the category candidates based on the class of k neighbors. The similarity of x and each neighbor’s document could be the score of the category of the neighbor documents. **[1]**

**Ada-boost**: Ada-boost is an ensemble learning algorithm to solve the classification problem. Adjusting the sample weight and the weight of the weak classifier, the weak classifier with the smallest weight coefficient is selected from the trained weak classifier into a strong classifier with higher prediction accuracy.

**SVM :** Kernel SVM is traditionally used for the binary classification. In this project, we generate a multi-class SVM with a "one-vs-rest" approach for multi-class problems. And the probability SVM calculates the probability that the sample belongs to its category by using Sigmoid function.**[2]**

**Random forest**: Random forest is an algorithm that integrates multiple trees using the idea of Ensemble Learning. Its basic unit is the decision tree. We select Gini measure \ G(Xm)\ =\sumk\ p\_{mk}(1-\ p\_{mk}) in random forest algorithm, where p\_{mk} is the proportion of class k samples in node m **[3]**.

**Logistic regression**: This algorithm is to map the result of a linear function to sigmoid function y=1/(1+e−x) **[4]**. This algorithm is a kind of classification measure. It will establish a cost function, and then iteratively replace the optimal model parameters through optimization methods, and then test to verify the quality of our transformed model.

**Soft Voting Classifier**: It takes all model prediction samples’ the average probability of a certain class as the standard, and the corresponding type with the highest probability was the final prediction result

* + 1. **Justification of the method choice**

The performance of KNN is dependent on finding a meaningful distance function and text classification needs more feature values. Similar articles are easily misclassified in the same topic by distance function because the distance of the same topic article is too far away from each other.

Ada-boost is based on a simple decision tree with one node to improve weak classifiers through continuous training. Because there are 11 topics, it is difficult to classify successfully. And if the tree is too complicated, it would lose the simplicity of Ada-boost which lead to bad effect as well.

Compared with other methods above, SVM has a good performance in terms of data and results. It is available for data set with multi-features by using a kernel function. The kernel function is calculated on the low dimension in advance, and the virtual classification effect is shown on the high dimension, which avoids the complex calculation directly in the high dimensional space. It can also avoid overfitting, learning a form of decision boundary called the maximum margin hyperplane. Detailed experimental results are shown in the discussion section below.

After the experiment, we got the performance of random forest is unsatisfactory. Through analysis, we found that irrelevant articles are scattered, which will interfere with the Random forest classification. Additionally, the random forest will produce overfitting. Finally, some articles have a very low weight. After random separation, even if we can predict them correctly, their probabilities are very low. So, we give up this classifier.

We select the OVR method as multi-class, which is more suitable for a linear model to conduct binary classification. Furthermore, Logistic regression can easily update the model to absorb new data. At last, we can get the weight of articles directly.

Compared with other classifiers, we found the performance of Soft Voting Classifier is better. We vote between Logistic Regression and SVC. The soft voting classifier uses the sum of the probabilities of each category to make predictions. The accuracy of soft voting is slightly higher than that of hard voting because it considers the difference of each model. We will talk about more details in results and discussion part.

* 1. **Hyper-parameter tuning**

To tuning the hyper-parameters of chosen model, the project follows the rule of control variable and grid-search. For hyper-parameters which is hard to charge the effect and importance between them, we use grid-search to find the best value. For those which might consume too much time and relatively independent over others, we choose control variable method to find the hyper parameter. The process and result will be presented in **Chapter 3.3**.

**2.3 Evaluation metrics**

During the process of building a machine learning model, it is based on some evaluation indicators to evaluate the performance of both the learning and prediction, thus it is an important part of machine learning.

For classification, the common evaluation indicators include Precision, Recall and F1-measure. In most cases, the calculation of Precision and Recall should build the confusion matrix at first which is shown below.

|  |  |  |
| --- | --- | --- |
|  | the number of articles that practically belong to this topic | the number of articles that practically belong to this topic |
| the number of articles which are classified as this topic | TP | FP |
| the number of articles which are not classified as this topic | FN | TN |

Precision is the proportion of the articles that truly belong to a topic of all the articles which are classified as this topic:

P = TP / (TP + FP)

Recall is the proportion of the articles that are classified to a topic of all the articles that truly belong to this topic:

R = TP / (TP + FN)

F1 is a trade-off between Precision and Recall:

F1 = 2PR / (P+R ) = 2TP / (2TP + FP +FN)

1. **Results**
   1. **Presenting**

Finally, we choose chi-square and voting method to get the result. **Table 3.3.1** is the score over all test instances and **Table 3.3.2** is the score based on suggested articles.

**Table 3.3.1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic name** | **Precision** | **Recall** | **F1** |
| ARTS CULTURE ENTERTAINMENT | **0.50** | **0.67** | **0.57** |
| BIOGRAPHIES PERSONALITIES PEOPLE | **0.71** | **0.33** | **0.45** |
| DEFENCE | **0.90** | **0.69** | **0.78** |
| DOMESTIC MARKETS | **0.33** | **0.50** | **0.40** |
| FOREX MARKETS | **0.62** | **0.44** | **0.51** |
| HEALTH | **0.64** | **0.50** | **0.56** |
| MONEY MARKETS | **0.57** | **0.72** | **0.64** |
| SCIENCE AND TECHNOLOGY | **0.25** | **0.33** | **0.29** |
| SHARE LISTINGS | **0.67** | **0.57** | **0.62** |
| SPORTS | **0.95** | **0.98** | **0.97** |

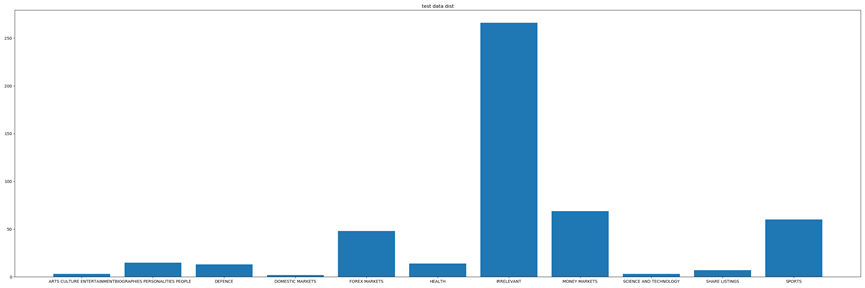
**Table 3.3.2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Topic name** | **Suggested articles** | **Precision** | **Recall** | **F1** |
| ARTS CULTURE ENTERTAINMENT | **9703,9789,9952,9830,9933** | **0.40** | **1.00** | **0.57** |
| BIOGRAPHIES PERSONALITIES PEOPLE | **9940,9988,9758,9933,9878, 9896** | **0.83** | **0.62** | **0.71** |
| DEFENCE | **9559,9770,9773,9616,9670, 9576,9987,9904,9706,9713** | **0.90** | **1.00** | **0.95** |
| DOMESTIC MARKETS | **9994,9989** | **0.50** | **1.00** | **0.67** |
| FOREX MARKETS | **9986,9977,9551,9875,9718, 9671,9588,9682,9786,9548** | **0.50** | **1.00** | **0.67** |
| HEALTH | **9661,9937,9807,9810,9929, 9873,9926,9609,9887,9833** | **0.60** | **0.67** | **0.63** |
| MONEY MARKETS | **9765,9618,9755,9602,9871, 9998, 9723,9737,9516,9967** | **0.80** | **0.62** | **0.70** |
| SCIENCE AND TECHNOLOGY | **9617,9722,9982,9621** | **0.25** | **0.50** | **0.33** |
| SHARE LISTINGS | **9518,9601,9666,9667,9972, 9715** | **0.67** | **1.00** | **0.80** |
| SPORTS | **9848,9596,9813,9942,9858,**  **9752,9620,9541,9656,9574** | **0.90** | **1.00** | **0.95** |

* 1. **Result evaluation using appropriate metrics**

Precision reflects the accuracy of each topic whereas recall means the ratio of the true positive number of each topic to the total number of the topic. F1 is the general evaluation combining the two. As the result shows, this method can make good suggestions on ‘BIOGRAPHIES PERSONALITIES PEOPLE’, ‘DEFENCE’,’ MONEY MARKETS’ and ‘SPORTS’ and the precession rate is very low on the topic of ‘SCIENCE AND TECHNOLOGY’ and there are only 2 recommendations of ‘DOMESTIC MARKETS’ . Figure 3.2.1 shows the distribution of test set, which indicates that the actual articles of these 2 topics are only 3 and 2.

Generally, the precision is higher than recall and has a good score in most topics, which means the final predict can distinguish the true articles well. And the relatively low recall means there are some irrelevant articles misclassified. To summarize, we think the final result is passable as all topics can suggest at least one right article and most topics performed well, considering the actual distribution.



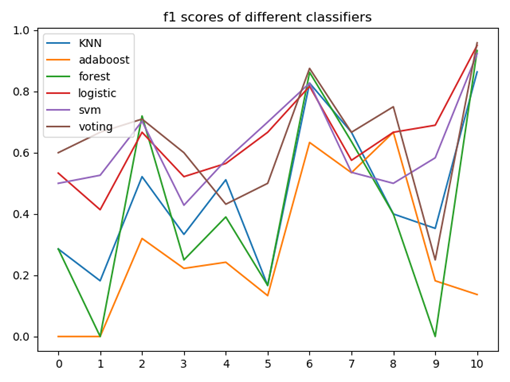
**Figure 3.2.1**

**Table 3.2.1**

|  |  |
| --- | --- |
| **Topic name** | **Articles numbers** |
| ARTS CULTURE ENTERTAINMENT | 3 |
| BIOGRAPHIES PERSONALITIES PEOPLE | 15 |
| DEFENCE | 13 |
| DOMESTIC MARKETS | 2 |
| FOREX MARKETS | 48 |
| HEALTH | 14 |
| MONEY MARKETS | 69 |
| SCIENCE AND TECHNOLOGY | 3 |
| SHARE LISTINGS | 7 |
| SPORTS | 60 |
| Irrelevant | 266 |

* 1. **Evaluation of various hyper-parameters and design choices**

After the laboratory of different kinds of classifiers shown in **Figure 3.3.1**, it is easy to notice that logistic regression and SVM have a better performance over others and they are complementary at some degree, so the final decision is to combine them into a voting classifier. As the figure is shown, the default voting classifier has a similar performance comparing to the two classic classifiers. The x-axis in the figure is the encoded topic with alphabetical order and 6 is irrelevant articles. The hyper-parameters are set with a relatively good result during the laboratory phase.

****

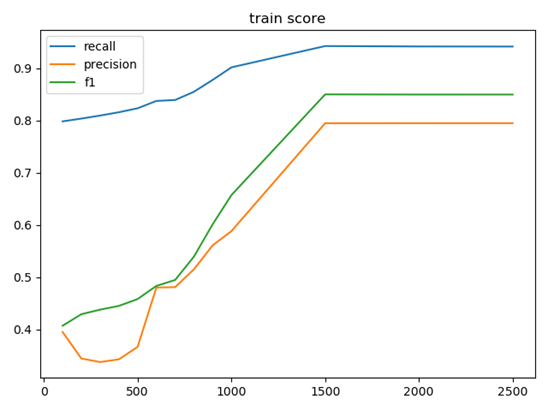
Encoded topics

**Figure 3.3.1**

Once the model is chosen, the next step is to tune its hyper-parameters. In this model, there are 3 classifiers need to be adjusted -- logistic regression, SVM and voting. We deploy a grid search and choose GridSearchCV from Scikit-learn with 5-fold cross-validation to complete this phrase. For logistic regression, we adjust the maxiter, multi-class and solver to find the best combination. Then, considering the fitting time of SVM, we first observe the effect of iterations on the f1-score with 10-fold cross-validation as shown in **Figure 3.3.2** and conclude that the iteration time is very subtle after 1500. Then we use grid-search to tune the C and gamma.

At last, we compare the overall score to decide the best weights for voting. All the tuning log are presented in **AP3.3.1** in Appendices.

After confirming the final model and hyper-parameters, we observe the output result based on the probability matrix on validation set and set the probability threshold to 0.4 so that only the article whose probability is bigger than 0.4 and which ranked in top 10 will be suggested.

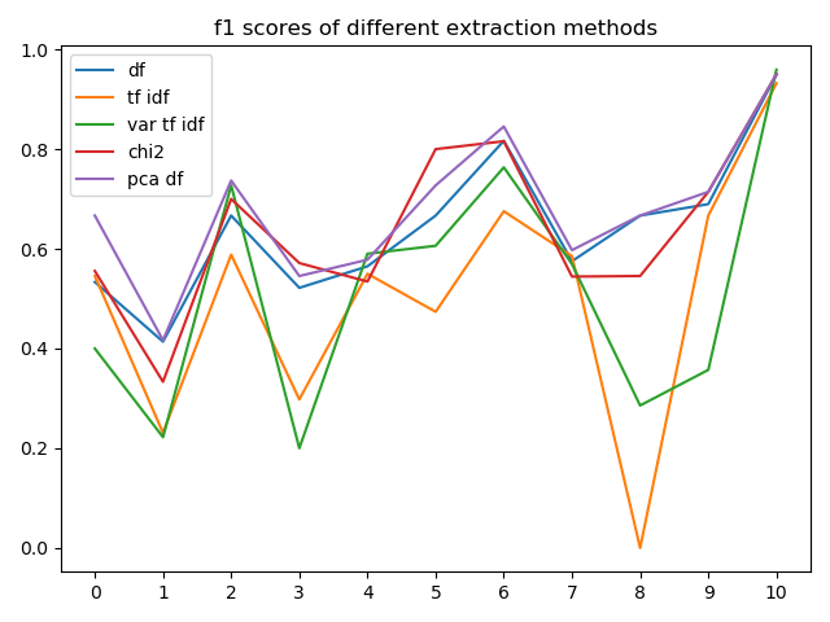


Encoded topics

**Figure 3.3.2**

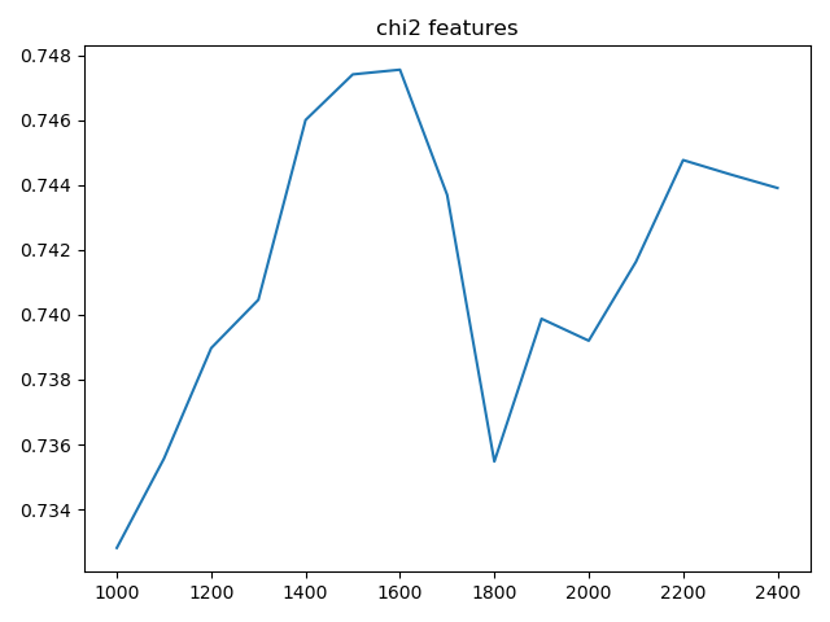
* 1. **Feature importance analysis**

Firstly, we used control variables to test the performance of each extraction method with the same number of features. As we can see from the figure below, chi2 generally performs well based on the performance of all 10 article topics. Thus, we chose chi2 as our best extraction method and then focused on the analysis of the optimal number of chi2 features used.

****

Encoded topics

**Figure 3.4.1**

****

Iteration time

**Figure 3.4.2**

The feature number of chi2 determines the number of category variables selected. By testing the iterative num between 1000 and 2400, it can be observed that the highest accuracy is 0.747 when iteration num is equal to 1600 from the above figure. Therefore, we choose 1600 as the best feature num of chi2 and combined with vote classifier to get a result of classification.

1. **Discussion**
   1. **Performance analysis**

As mentioned in the method choice section, figure 3.3.2 and figure 3.4.1, we select the best extraction model and classifier, which is CHI2 + vote. Compared with other methods, chi2 and voting classifier give us better performance. CHI2 is more suitable for the dataset of this project. Because we have 10 different topic and 1 irrelevant topic. The soft voting classifier uses the sum of the probabilities of each category to make predictions. This method can find the classifier suitable for our data set to the greatest extent. After the experiment, the effect of this integration method is indeed excellent.

* 1. **Evaluation metric analysis**

As discussed before, precision score reflects the accuracy of each topic whereas recall means the ratio of the true positive number of each topic to the total number of the topic. For this project, the reader wants to have as many as true positive articles presented on its suggested list, so precision score in the second result table can represent more accurately.

* 1. **Further improvement**

In the project, we mainly explore feature engineering, classifier modelling and evaluation methods of news classification. Feature engineering is based on Bag of Words and modeling is based on classifiers within machine learning. We have been devoted to developing different methods to deal with this problem and finally, the output performance has been effectively improved compared with the beginning. However, the performance still needs further promotion and will start from the following aspects.

1. For feature engineering, by reading the given articles of the dataset, we cannot get abundant semantic meaning from our perspectives. Without taking into account the semantic meaning, we have not tried word embeddings which are a common family of NLP techniques on feature extraction.
2. For building classifiers, we haven’t tried deep learning classifiers such as CNN and RNN in that they are mainly based on context-awareness and need features indicating the position of each word. Deep learning is powerful at learning details and as a result, the promotion of performance can be achieved when it comes to complicated situations.
3. For data augmentation, basic techniques of upsampling and subsampling had been adopted due to the imbalanced distribution of the dataset. However, we haven’t witnessed the improvement so we gave up data augmentation at last. Further research and experiment should be done in this step to promote the robustness of the model.
4. **Conclusion**

To finish the project, we first have a general trial of different kinds of extraction methods and classification methods. And after comparing with a relatively good hyper-parameter of all those methods, we choose 2 best extraction methods -- DF and Chi2 and also combine 2 best classifiers into the voting classifier. Then we deploy grid search and k-fold cross-validation on training set to tune the best hyper-parameters. To decide the threshold of recommendation articles, we observe the probability matrix in the validation set. Eventually, we compare 2 sets of results based on 2 different extraction methods and choose the final result.

Generally speaking, the influence of extraction methods and their parameters is remarkable, meanwhile, with all classic classifiers presented in chapter 2, linear model -- logistic regression and SVM tend to have a better scope. However, given the fact that the distribution of articles is very irregular and the way to process features is very primitive, the overall result is just passable and it is hard to have further improvement. To achieve more progress in text topic classification problem, more advanced algorithms such as attention-based LSTM and other neural networks may be compulsory.

**Reference List:**

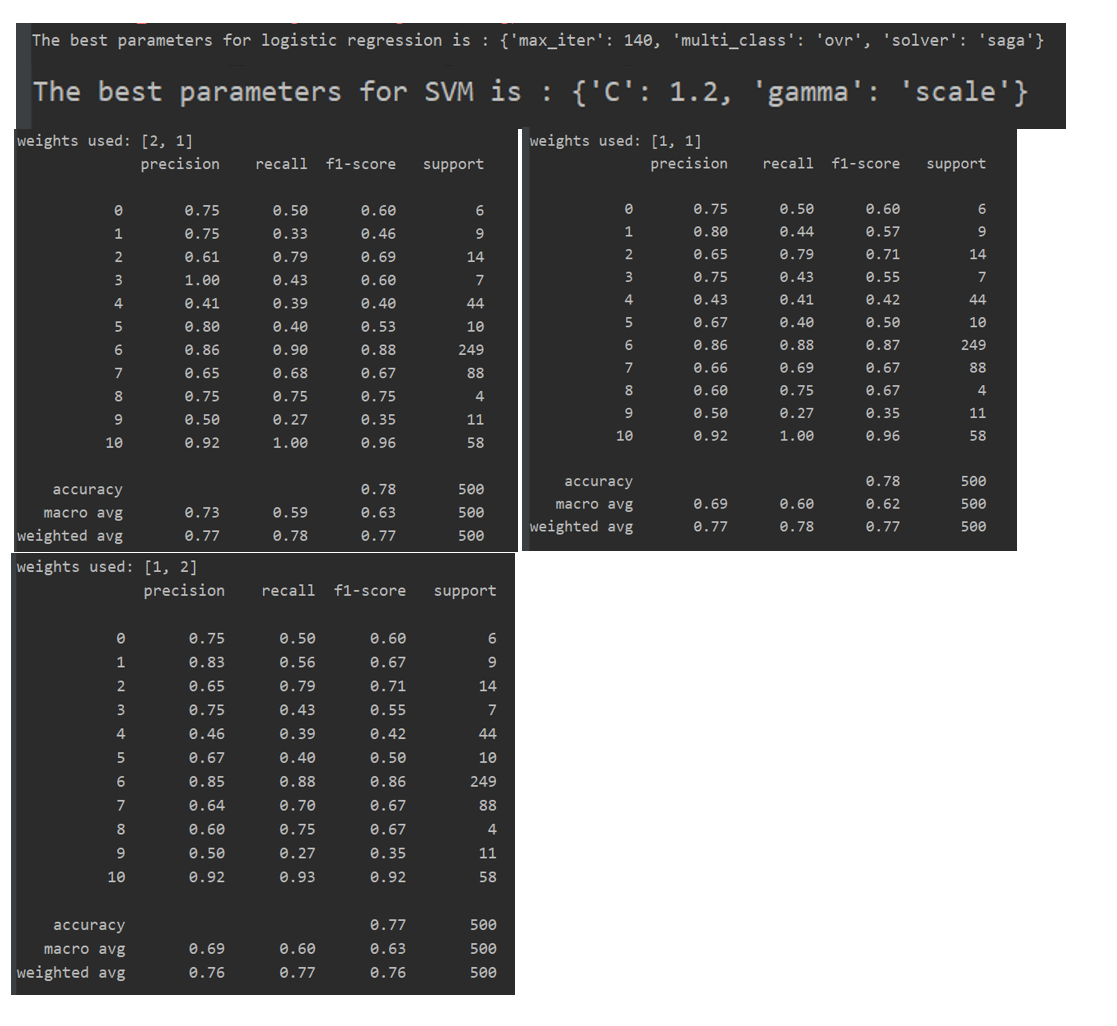
[1].Text Classification Algorithms: A Survey Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, Donald E. Brown (23 April 2019) DOI:10.3390/info10040150

[2]. Classification of News Dataset Olga Fuks Stanford University [2018]

[3]. Scikit-learn.org. n.d. 1.10. Decision Trees — Scikit-Learn 0.22.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/tree.html> [Accessed 19 April 2020].

[4]. Scikit-learn.org. 2020. 1.1. Linear Models — Scikit-Learn 0.22.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression> [Accessed 19 April 2020].

**Appendices:**



**AP 3.3.1**