UNSW

THE UNIVERSITY OF NEW SOUTH WALES

Group Project

COMP9417 Machine Learning and Data Mining T1, 2020

Group: Take Care Yourself

Group Member

Anrui Wang	z5183159
Chenfan Zhu	z5199811
Haoran Luo	z5194840
Jiling Yang	z5197332
Rongtao Shen	z5178114

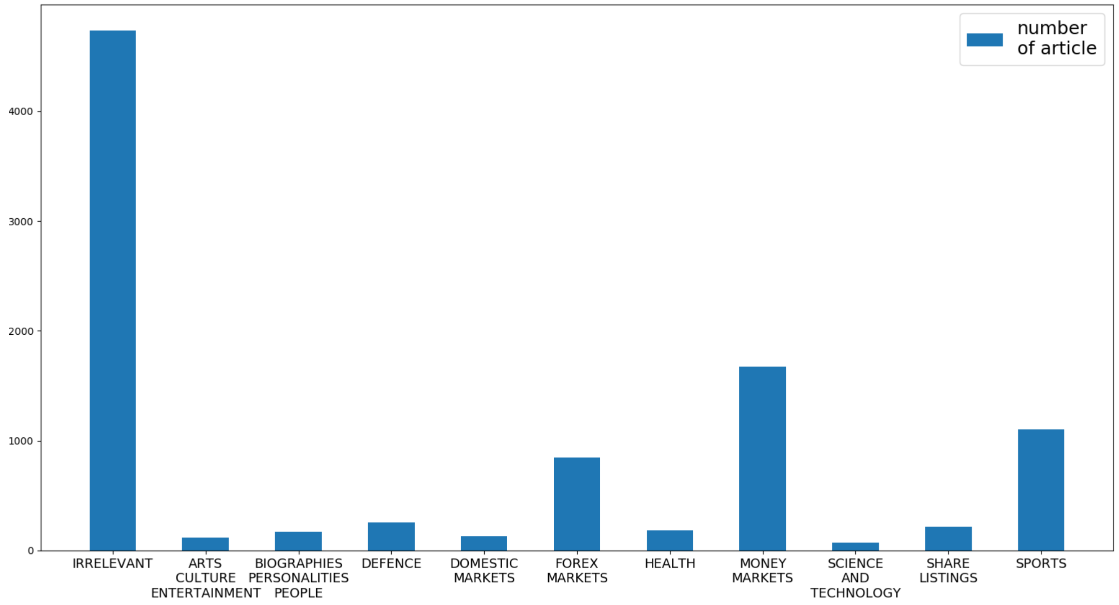
1. Introduction

With the rapid development of information technology, people no longer suffer from the lack of information and gradually enter the era of information overload. Nowadays, more than 500 new articles are published every day. As a result, information sifting as well as personalized recommendation is of great importance for news industry. In order to recommend each reader with articles that they are most likely to be interested in among a wide variety of fresh daily articles, this project aims to predict the most relevant news articles for each of the 10 topics (art culture entertainment, biographies personalities people, defense, domestic markets, forex markets, health, money markets, science and technology, share listings, sports) through the method of text classification.

The given datasets are already pre-processed, where unnecessary words are removed, and each word is separated by a comma. However, there is a class called 'IRRELEVNANT' in the datasets, which means no one is interested in such kind of articles. Meanwhile, there are in total 9,500 samples in the training set and the irrelevant data accounts for nearly 50% of it while some topics like domestic markets has less than 150 samples. So, in this project, the major challenge is to deal with the imbalance between different classes. Another challenge is to ensure the recommendation quality, since we do not want to post too much articles that are not in the right class to the readers, even if there are no more than 10 articles in that class.

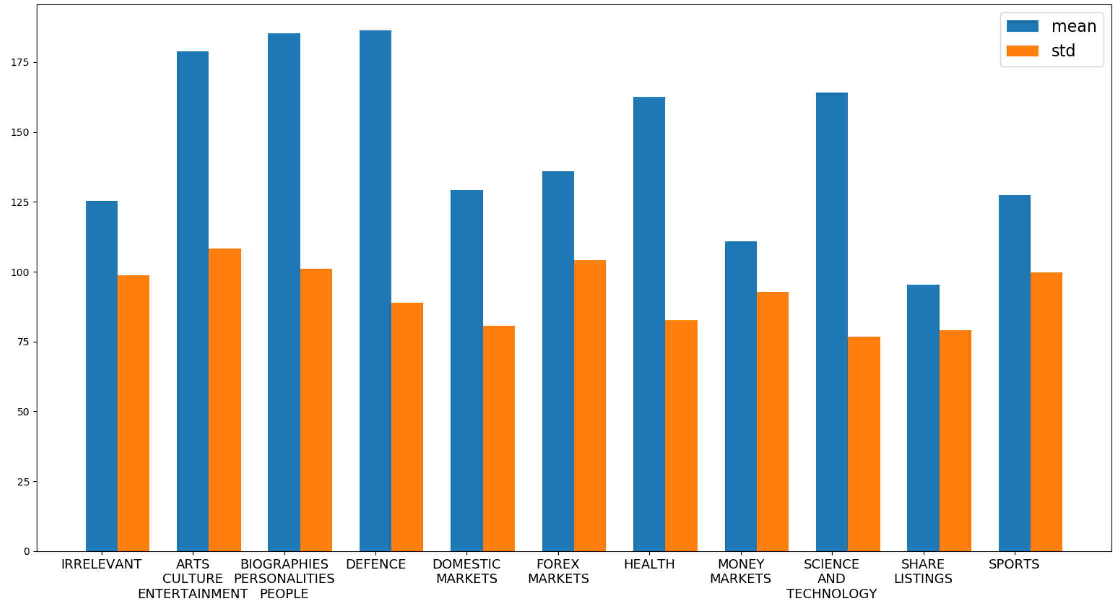
2. Exploratory data analysis

In order to get a better performance, it is a common practice to carry out an exploratory data analysis to gain some insights from the data. For the training dataset, there are only three columns. Two of them are really important, one is article words and the other is the topic of news articles. For article words column, each row in the training dataset contains all articles words of that article, some conjunction words and unimportant words have been deleted. As for label column, one of our main concerns when developing a classification model is whether the different classes are balanced which means whether the portion of each class is equal. After analyzing and counting the number of different classes in the training dataset, we can easily get a bar chart:



We can see that the number of samples for each topic varies significantly, 48% of the articles are in IRRELEVANT topic while 8 topics have less than 260 articles. Therefore, the imbalance sample in training dataset is a big problem which we need to deal with in preprocessing and we will mention the method in methodology part.

We also analyze the length of each article in the dataset and get the following result:



It can be noticed that the article length does not varies much between different topic, so the length will not be an important feature to classify those articles.

3. Methodology

Feature selection:

In this part, there are two things we need to explain. One is how to deal with the article words in the training dataset. As for article words column in the training dataset, all unimportant and conjunction words have been deleted, we can just focus on the frequent and important articles words in different news articles. In order to process and analyze this important feature in the training dataset, Bag-of-Words Model is an important concept to explain.

Since we cannot work directly with text in training text classifier, we need to convert the articles into vectors. Bag-of-Word Model focus on the occurrence of words in document and will convert an article into a matrix of token counts rather than the order information. Hence, there are two schemes that can be used in preprocessing. One method is CountVectorizer, and the other is TfidfVectorizer. There are some differences between these two methods. As for CountVectorizer, this method aims to convert text to the word count vectors. After creating an instance of the CountVectorizer class and using fit and transform functions, an encoded vector will be returned with a length of the entire vocabulary and the number of occurrences of each word in document. And for TfidfVectorizer, this method aims to convert text to word frequency vectors. TF means how often a given word appears in the document and IDF means inverse document frequency, we can use this numerical statistic to determine how important a word in a corpus. In preprocessing, we compare these two different methods and there are some differences on performance of classifier.

Data re-sampling:

Two methods are considered to deal with imbalanced datasets. One is to under-sample the majority class and the other one is to oversample every classes except the majority class. Both approaches intend to use imblearn package to solve the class imbalance problem. When doing under-sampling, we used RandomUnderSampler which can be imported from imblearn.under_sampling, deleting some samples which have large proportions in order to make all classes have a similar proportion. The advantage of this method is that the training speed is fast because the processed training set will be much smaller than the original training dataset. But it will randomly delete some valuable samples and cause the loss of some important information which will make the classifier has poor performance. Oversampling can be imported from imblearn.over_sampling, and it can add some samples which make the dataset larger than the original dataset. As a result, we will employ over-sampling to mitigate data imbalance.

Model:

We employ 2 models to do the classification task.

a. Logistic Regression classifier

Although the name of logistic regression is about regression, it is actually a classification model. It is a discriminant model that can be applied to solve multi-class classification tasks. By using the One vs One strategy, all the classes in a multi-class classification task are taken out in pairs and trained into several binary classification when we use logistic regression to train our model. For example, if we totally have N classes, it will generate N(N-1)/2 binary classification tasks. When we in the test process, the new sample will be submitted to all binary classifiers and generate N(N-1)/2 results. The final result will be produced by vote which means the most predicted class will be the final classification result. In addition, this model can directly model the possibility of classification without the need to assume the data distribution in advance. It calculates the probability that, given the input x, the probability that the data belongs to class y (Indra, S.T., Wikarsa, L. and Turang, R., 2016). As a result, it can obtain approximate probability predictions instead of which classes. After comparative analysis, it can be indicated that logistic regression classifier can get the best classification accuracy and precision.

The steps of generalize logistic regression to multi-class classification are as follow:

1. Class1 is regarded as a positive sample, and all other classes as negative samples, and then we can get the probability p_1 that the sample is belongs to class1.

2. Class2 is regarded as a positive sample, and all other classes as negative samples. Similarly, p_2 that the sample is belongs to class2 is obtained;

3. We can enumerate all the classes to obtain the probability p_i and finally we take the class that corresponding to the largest probability of p_i as final prediction.

b. Multinomial Naive Bayes

Naive Bayes is a generative model which aims to minimize conditional risk. It is based on attribute conditional independence assumption. For a given training dataset, it will first compute the joint probability of samples and hypothesis. For a new sample, it will compute the posterior probability according to the data from training set and find the class with maximal posterior probability. Multinomial Naive Bayes implements the naive Bayes algorithm for multinomially distributed data, which is an effective algorithm when we carry out the text classification. Since the relationship between two words in the article words is independent, we can change the probability of belonging to a category under certain conditions to the probability of having a certain characteristic under the condition of belonging to a certain category by using Multinomial Naive Bayes model to train the classifier. In addition, the naive Bayes model originates from classical mathematical theory with a solid mathematical foundation and stable mathematical efficiency. Although the training dataset is large, naïve Bayes model can also get a high training speed and accuracy. Therefore, it is a suitable method to implement.

4. Results

Final results for each class calculate

Topic name	Precision	Recall	F1
ARTS CULTURE ENTERTAINMENT	0.38	1.00	0.55
BIOGRAPHIES PERSONALITIES PEOPLE	0.78	0.47	0.58
DEFENCE	0.73	0.85	0.79
DOMESTIC MARKETS	0.40	1.00	0.57
FOREX MARKETS	0.54	0.77	0.64
HEALTH	0.69	0.79	0.73

MONEY MARKETS	0.94	0.84	0.89
SCIENCE AND TECHNOLOGY	0.33	0.33	0.33
SHARE LISTINGS	0.60	0.86	0.71
SPORTS	0.95	0.98	0.97

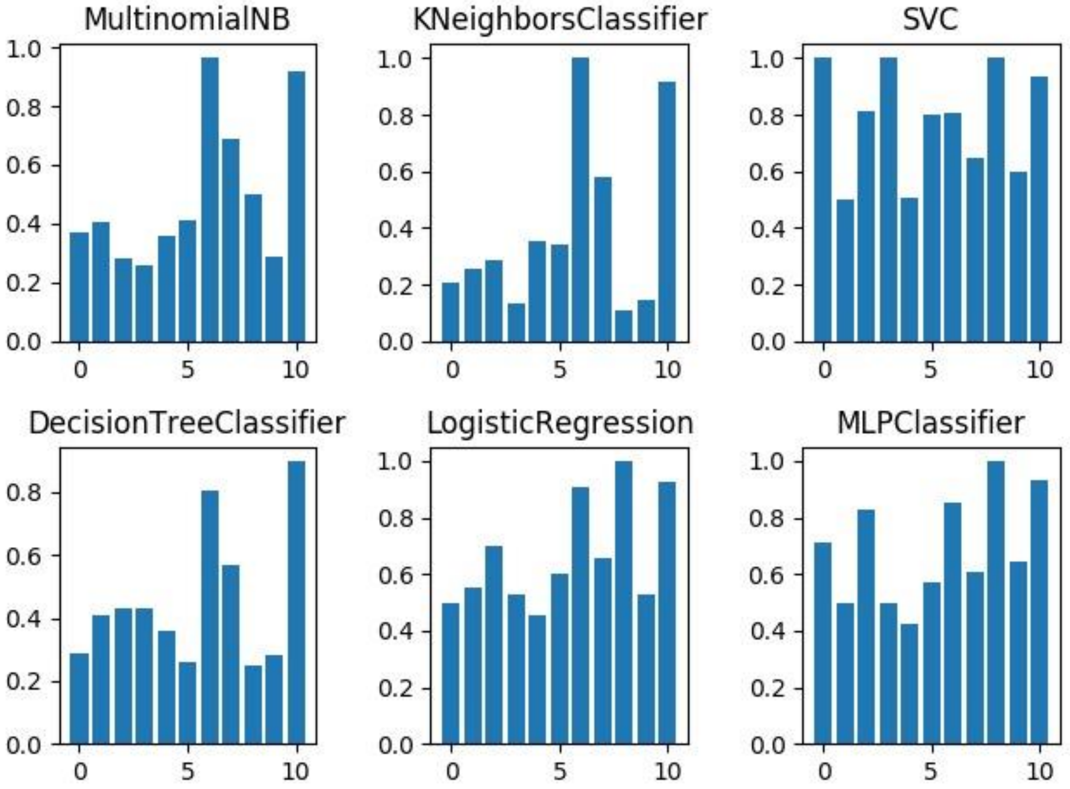
Topic name	**Suggested articles**	**Precision**	**Recall**	**F1**
ARTS CULTURE ENTERTAINMENT	9952, 9789, 9703, 9830, 9933, 9604	0.33	0.67	0.44
BIOGRAPHIES PERSONALITIES PEOPLE	9878, 9896, 9695, 9940, 9758, 9988, 9854, 9645	0.75	0.40	0.52
DEFENCE	9559, 9576, 9616, 9770, 9773, 9759, 9607, 9670, 9842, 9987	0.80	0.62	0.70
DOMESTIC MARKETS	9994, 9796, 9989, 9640, 9923, 9954, 9750	0.29	1.00	0.44
FOREX MARKETS	9986, 9772, 9632, 9748, 9718, 9588, 9682, 9693, 9894, 9529	0.40	0.10	0.14

HEALTH	9873, 9937, 9661, 9887, 9807, 9810, 9609, 9735, 9833	0.78	0.5	0.61
MONEY MARKETS	9998, 9871, 9618, 9602, 9761, 9516, 9863, 9755, 9766, 9898	0.80	0.12	0.20
SCIENCE AND TECHNOLOGY	9617, 9982, 9722	0.33	0.33	0.33
SHARE LISTINGS	9518, 9666, 9601, 9715, 9972, 9667, 9562, 9655, 9999, 9876	0.60	0.86	0.71
SPORTS	9886, 9596, 9574, 9752, 9857, 9569, 9568, 9760, 9573, 9942	1.00	0.17	0.29

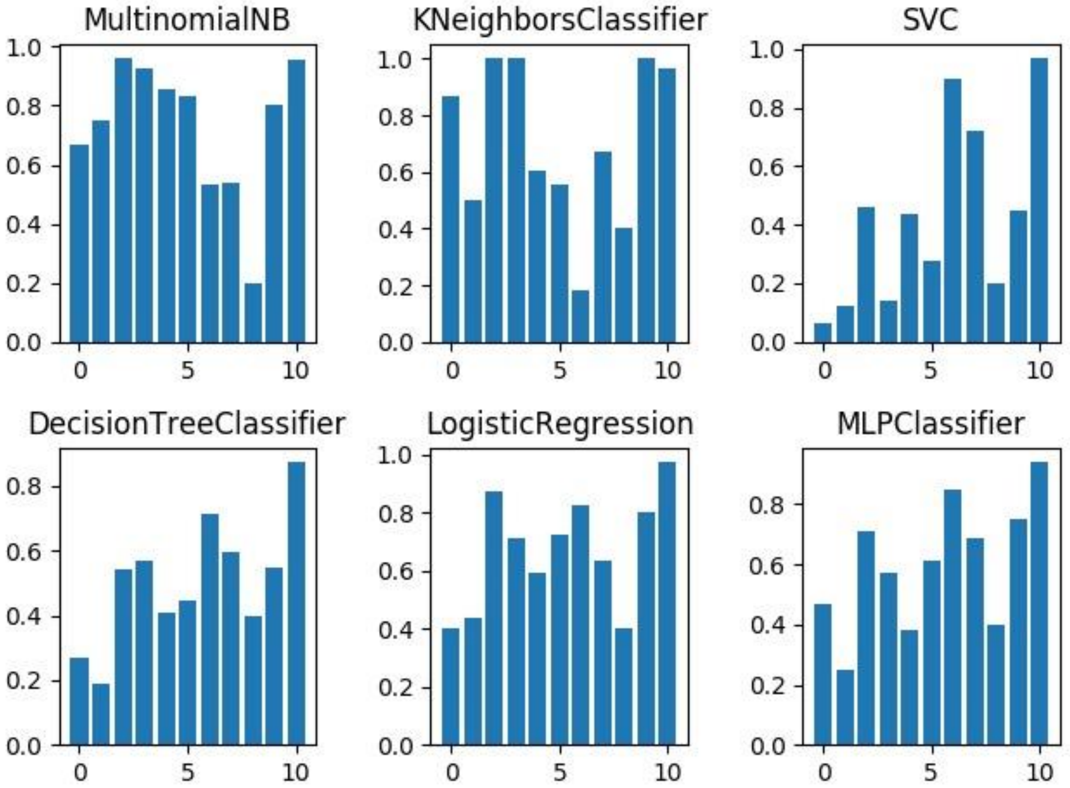
5. Discussion

Methods Comparison

We employ a 5-fold cross validation on 6 models and evaluate based on their precision and recall.



precision



recall

We can see from the charts, Logistic Regression model has achieved high score in both precision and recall. MNB without prior has the highest recall and an average precision. So, we select those 2 models as candidates. Due to the imbalance distribution of topics, models like k Neighbors Classifier cannot perform very well since there is no enough samples for some topics.

Choosing Metrics

For this project, 3 metrics are used to evaluate final output of the model including precision, recall and f1-score. Since we do not want to recommend much articles that are out of their preference to the readers, we consider that precision is the most important metric. For these three metrics, precision is used to show the correctness of categorize of all recommend articles; recall demonstrate that if all article belong to this certain categorize are picked from dataset; and f1-score is harmonic mean of precision and recall which is a more comprehensive metric to give consideration to both of these two metrics. Thus, as we emphasized earlier, the critical aim of this project is to provide right article to users, the model with higher performance on precision is the fittest metric, and after ensuring that, providing all articles relevant to a certain topic can be focused.

Possibility of improvement

According to the accuracy of our current model, there are still large space to improve. In our consideration, BiLSTM might result in better performance on this project. Because, if someone want to predict the topic of an article, it is necessary to know the context of the word, rather than just figure all the words. In BiLSTM, both long-term and short-term context can be saved in the model. Unfortunately, due to hardware restriction we cannot train a BiLSTM model in a reasonable time.

Except from utilizing neural network to extract features, we also can select features manually. We only consider the frequency of words appear in the article and have not considered the impact or word ordering. So some information is lost after convert article words into vectors. If we take word ordering into account and make it as an additional feature, a better performance may be achieved. Besides, in this project we treat all words equally so employ methods like information gain, which is a metric that can evaluate the amount of information a word has, could also lead to a better result.

6. Conclusion

In this project, the Logistic Regression model with TfidfVectorizer preformed best and we choose precision, recall and f1-score as final metrics, where the precision is the most important one.

First of all, in different news articles, we only focus on those words that appear frequently and are relatively important in the specific topic. As we mentioned in exploratory data analysis part, the distribution of articles over various topics is not uniform. And among three method we have tried (binary classification, undersample, oversample), oversample gets a better result by increasing the size of the original datasets to achieve a relatively balanced total number of each topic. Undersample also do well sometimes, however, it is unstable because more likely to delete the important information which leads to a low accuracy in the test data. In addition, the parameter of smote is vital point to improve the final result.

Due to the emphasis on the frequency of words, Bag-of-Word is chosen, and it computes the occurrence of different words in articles, where we use CountVectorizer and TfidfVectorizer in the sklearn lib.

As for the training model, there are two models finally chosen, which can get better accuracy than others we have tried. The first and best one is logistic regression, which can avoid the problems on topics distribution and get probability of each topics. The second one is Multinomial Naive Bayes, originates from classical mathematical theory with high efficiency and accuracy.

In a news article, the semantics of a word often depend largely on its context. Therefore, for the improvement of this project, we believe that the use of Bi LSTM can improve the existing accuracy.

In conclusion, although the method in this project has not gain a really high accuracy, for larger datasets, it is still efficient with short runtime.

7. Reference

1). Rennie, J.D., 2001. Improving multi-class text classification with naive Bayes.

2). Xu, S., Li, Y. and Wang, Z., 2017. Bayesian multinomial Naïve Bayes classifier to text classification. In Advanced multimedia and ubiquitous engineering (pp. 347-352). Springer, Singapore.

3). Pranckevičius, T. and Marcinkevičius, V., 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), p.221.

4). Indra, S.T., Wikarsa, L. and Turang, R., 2016, October. Using logistic regression method to classify tweets into the selected topics. In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 385-390). IEEE.

5). Moreo, A., Esuli, A. and Sebastiani, F., 2016, July. Distributional random oversampling for imbalanced text classification. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 805-808).

6). Zheng, Z., Wu, X. and Srihari, R., 2004. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter, 6(1), pp.80-89.

7). Mladenic, D. and Grobelnik, M., 1999, June. Feature selection for unbalanced class distribution and naive bayes. In ICML (Vol. 99, pp. 258-267).