# BIA 660 Final Report

Advisor:

Emily Liu

Group 3

Xiao Zhou

Tianhao Ruan

Qinglin Yang

Xiaoyu Zhou

1. Overview

### 1.1. Motivation and Research Questions

In this project, we will study the effects of the product titles to the customers on a crowdfunding website - Kickstarter. A clear name and a brief description can help the customers quickly understand the product. The customers will click onto the products, only if the name and the subtitle are creative and interesting enough. A good and unique product name can attract customers and build a connection between the designer and the customer.

There are 15 categories of projects on Kickstarter based on the industry and the form of the project. The film & video, music, and games are the top 3 categories with highest total funding. However, these projects are visually sensitive, which are not ideal for text mining. The wording in these projects is not well controllable, as it could be particularly abstract. Thus, we will study the titles and subtitles of the products of design.

### 1.2. Background

At the initial stage of the project, we were trying to scrape the data directly from Kickstarter using Python. However, the data was limited, because the website does not provide information about the failed projects. We could only scrape the live projects and the previous successful projects from Kickstarter. In order to do effective text mining and prediction, we would need the data of the failed projects. That's why we imported pre-crawled datasets from reliable resources – Web Robots (Kickstarter Datasets) and Harvard Dataverse (Li, 2009). After comparing, we have decided to use the datasets from Web Robots, which have less complex structure and more controllable data size. Besides, Web Robots provides more historical data. Unlike the datasets from Harvard Dataverse, which consist of multiple SQL tables with a large number of variables, the essential variables from Web Robots' datasets are collected within one csv file.

### 1.3. Literature Review

While most text mining research of marketing focuses on customer reviews, we want to know what contributes to a successful product by analyzing the product titles. Some researchers study from sentimental aspects and prove that the emotion of the designer can stimulate consumers' willingness to consume (Wang et al., 2017). The predictive accuracy reaches 71.7% after applying the sentimental factors with the Support Vector Machines (SVM) classification method. Support Vector Machines Recursive Feature (SVM-RFE) is used in another study to select and evaluate important features and to find the best classification result (Chen & Shen, 2019). This study predicts the success of the project funding with the SVM prediction model, which received an accuracy of 78.9%.

Various researchers study the factors influencing crowdfunding success on Kickstarter using text mining. R. S. Kamath et al. (2016) designs a classification model for the analysis of Kickstarter campaigns to identify the possibility of success of a campaign. From the analysis of this research, project properties play a vital role in predicting success and neural network is the suitable classifier for Kickstarter campaigns. It also proves the claim that projects with a video report a higher success rate than those without. Wei Wang et al. (2016) studies the influence on the investment willingness among different language persuasion styles by conducting empirical analysis on 128,345 projects on Kickstarter. The research concludes that different project categories correspond to different persuasion styles due to the nature of the projects. For the gaming industry, Yang Song et al. (2019) uses principal components analysis (PCA), logistic regression, and the OneRule method to analyze 9962 game projects on Kickstarter between 2009 and 2018, which highlights the importance of choosing appropriate words in a project's title to increase the probability of success.

**Key words**: Clustering, Random Forest, Classification, Naïve Bayes, SVM

2. Dataset

We are using the pre-crawled datasets from Web Robots (Citation). The researchers from Web Robots crawl the projects of Kickstarter every month. The data we use in this project was crawled from January 16, 2020 to April 15, 2021 by Web Robots.

After organizing the csv files together, there are 38 columns and 224,179 samples of the overall data. To compare the performance of the models, we will also apply the models to the products from all categories. Then, to minimize the gap and the possible errors between different categories, we have decided to mainly focus on the subcategories of design, which includes Interactive Design, Graphic Design, and Product Design. Some useful variables in our datasets are – slug, blurb, category, state, usd_pledged.

- Slug

  Slug is the title of the product. The product name is at the top of the product page. The product name highly summarizes the features of the product in one sentence.

- Blurb

  Blurb is the subtitle of the product, which is a short paragraph with 1 to 2 sentences. Blurb is a brief description of the product. Both the subtitle and the name of the project are the most fundamental points of a project. When the customers are browsing projects on the main page, tittle and subtitle are the ones they look at first.

- Category

Category is the product category. We mainly focus on the category of design, while we are subtracting other categories, such as typography, civic, and toys.

- State

  State is the status of the project, there are 4 kinds of status – canceled, failed, live, and successful. We need to select and analyze the succussed and failed projects. For better calculation, we have replaced the "succussed" and "failed" with "1" and "0". The state will be the label of the data.

The following is a sample of the data.

"*slug*": "r rugged-black-best-damn-athleisure-wear"

"*blurb*": "Performance joggers, hoodies, and henleys made with premium 4-way stretch fabric and tailored fit. Made for action. Good for lounging."

"*category*": "Product Design"

"*state*": "1" or "successful"

## 3. Methodology

### 3.1. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). What we want to do is to cluster the Kickstarter projects which belong to the design category. After data preprocessing, we have 7043 rows to cluster. The text data we use to train our model is the title and blurb of each Kickstarter projects (the 'name' and 'blurb' in web robots dataset). The Kickstarter website has labeled the projects of design category to 7 different subcategories, they are architecture, civic design, graphic design, interactive design, product design, toys, and typography. We could use these labels to evaluate our cluster model to test the performance of our model.

The algorithm we use is K-means clustering algorithm, and we use cosine similarity as our distance measure. We load our data, split our data into train and test parts with test size of 0.2, and generate the tf-idf matrix. As the projects on Kickstarter have been labeled into 7 different categories, we set the number of clusters as 7. We use nltk K-mean algorithm to train our clustering model by repeating 20 times with different initial centroids. The results of clustering would be assigned to cluster labels which starts from 0 to 6. After clustering, we want to

interpret each cluster by centroid, and we found a problem. We get the top 20 tf-idf weight words in each cluster and most of clusters have the word ‘design’. We realize that we need to add ‘design’ to our stop words list, because all of our text data is related to design. We regenerate the tf-idf matrix and rerun our clustering model. The cross tabulation of our model is shown below.

| label<br>cluster | architecture | civic design | graphic design | interactive design | product design | toys | typography |
|---|---|---|---|---|---|---|---|
| **0** | 11 | 3 | 202 | 6 | 49 | 13 | 7 |
| **1** | 5 | 1 | 30 | 8 | 15 | 0 | 14 |
| **2** | 81 | 42 | 159 | 29 | 10 | 5 | 6 |
| 3 | 69 | 13 | 19 | 16 | 88 | 5 | 0 |
| **4** | 20 | 12 | 57 | 31 | 114 | 8 | 3 |
| **5** | 5 | 1 | 5 | 4 | 138 | 1 | 0 |
| **6** | 10 | 10 | 17 | 12 | 53 | 2 | 0 |

From the cross tabulation we can see that the projects which belongs to “typography” or “toys” have very small numbers and could not be divided by the clustering model. And the “architecture” and “civic design” group also been split to each cluster. Therefore, it is essential to change our parameters to fit our model better. We consider the label that Kickstarter provides may cause some confusion. ‘Toys’ is a certain kind of ‘product design’ and ‘interactive design’ and ‘typography’ are parts of ‘graphic design’. What’s more, we could merge ‘civic design’ and ‘architecture’ to a new group: ‘environment design’.

Now we have 3 kinds of labels, and we could rerun our clustering model with number of clusters 3 and 4 to find a well-performanced one. The performance is shown in the figures. From the cross tabulation with 3 clusters, we found that most product design projects could be clustered, but about half of the graphic design projects and almost all the environment design projects mix together. The performance is better with 4 clusters, but we also could not explain the meaning of the first cluster well and there are still some graphic design projects are in the same cluster with environment design projects.

| label | environment design | graphic design | product design |
|---|---|---|---|
| cluster | | | |
| 0 | 17 | 291 | 80 |
| 1 | 242 | 268 | 47 |
| 2 | 24 | 66 | 374 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| environment design | 0.00 | 0.00 | 0.00 | 283 |
| graphic design | 0.59 | 0.89 | 0.71 | 625 |
| product design | 0.81 | 0.75 | 0.78 | 501 |
| accuracy | | | 0.66 | 1409 |
| macro avg | 0.47 | 0.55 | 0.50 | 1409 |
| weighted avg | 0.55 | 0.66 | 0.59 | 1409 |

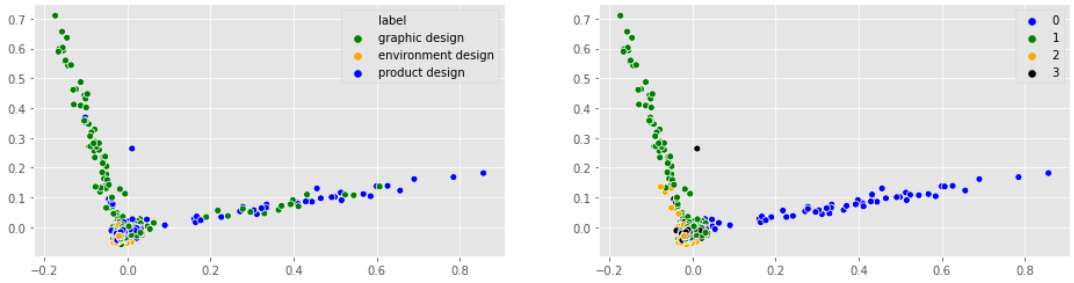
Performance of model with 3 clusters

| label | environment design | graphic design | product design |
|---|---|---|---|
| cluster | | | |
| 0 | 0 | 26 | 46 |
| 1 | 38 | 381 | 48 |
| 2 | 226 | 173 | 46 |
| 3 | 19 | 45 | 361 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| environment design | 0.51 | 0.80 | 0.62 | 283 |
| graphic design | 0.82 | 0.61 | 0.70 | 625 |
| product design | 0.82 | 0.81 | 0.82 | 501 |
| accuracy | | | 0.72 | 1409 |
| macro avg | 0.71 | 0.74 | 0.71 | 1409 |
| weighted avg | 0.76 | 0.72 | 0.72 | 1409 |

Performance of model with 4 clusters

We could use PCA to visualize our clustering data. We can see that our clustering model has done a relatively good job in clustering the Kickstarter design projects. As we only use the project name and blurb as our text which is about 10 to 50 words, it could be of much difficult to cluster the data more accurately. If we want to increase the performance, we could only use more text data in the story part of each project to build clustering model. And it is what we need to do in the future steps.




### 3.2. Classification

We firstly import the data and select our target category - design. Because the dataset of 2021 has only 4,902 projects with duplicates, we will combine the dataset of 2020. The data have been marked as failed or successful with proper labels. We could convert the words “successful” and “failed” to 1 and 0. Combining the datasets of 2020 and 2021 and removing the duplicates, we have 7,014 samples of successful and failed projects in the design category, where 1,636 samples are failed projects and 5,378 samples are successful projects. The inputs are the blurb and the state of the projects.

To prepare the data for model training, we split the data into 70% for training and 30% for testing. We will import and initialize the TFIDF Vectorizer. We manually select the parameters, where the stop word is "english" and default threshold, and transform the test documents by the fitted TFIDF Vectorizer.

### i. NB & SVM

We're comparing 2 models in this section- Naive Bayes model (NB) and Support Vector Machine model (SVM). NB model and SVM model are widely used in text mining and text analysis. We train the multinomial Naïve Bayes model using the testing data and train the linear SVM model with the testing data. We will compare the model with higher accuracy and evaluate the performance of the models.

In order to get the optimal results, we can then use Grid Search to find the best parameters. Here, we create a pipeline with integrates TF-IDF Vectorizer and SVM classifier. We use the metric to select the best parameters. Then, we use Grid Search CV function with 4-fold cross validation to find the best parameter values based on the training dataset. If we input the training data of the blurb and the state, we can get the best parameter for the model. The outputs here are the optimal parameters of the model. According to the results, we need to set the stop words to "english" and min_df to 5. Applying the function with proper parameters, we can finally run the function with testing data. Comparing with the previous models, we achieved a slightly higher accuracy rate of 81% and 77%.

We've achieved good accuracy rates with the products of design category. However, if we expand the dataset to all kinds of categories, we receive different results. There will be 7,006 samples in the dataset, where 3,072 of the samples are failed projects and 3,934 samples are successful projects. In SVM model, by selecting the best parameters with "None" as the stop word and "1" as the threshold, we get an accuracy of 59%. In the NB model, we receive an accuracy of 63%. The overall performance of the models is not as good as the ones we use in the design category, because the sample size is not big enough and the features of the products in all categories are dispersive.

### ii. Random Forest

In this part, we will train a random forest classification model to forecast whether a crowdfunding project is successful according to analyzing the key words from the title of the project. Since random forest builds multiple decision trees and merges them together to get a

more accurate and stable prediction, it can help us to get the accurate decision factors from the random forest model.

We selected the design module information of Kickstarter from the Web Robots website as the dataset, represented by “blurb” in the code. First, we used blurb dataset as input, and then we defined the label that if a project is successful and it will show as “1”. Next, we transferred the dataset to the TF-IDF weight. Otherwise, we optimized some parameters to apply the random search. We set the number of parameter settings as 100 and the cross- validation as the integer 3 here.

Finally, we got the accuracy that the random forest score is approximate 77.5%. We also illustrated virtualized tree as Figure Tree showed below. The figure only shows a snippet of the full tree because the full tree figure is too small to show the detail.

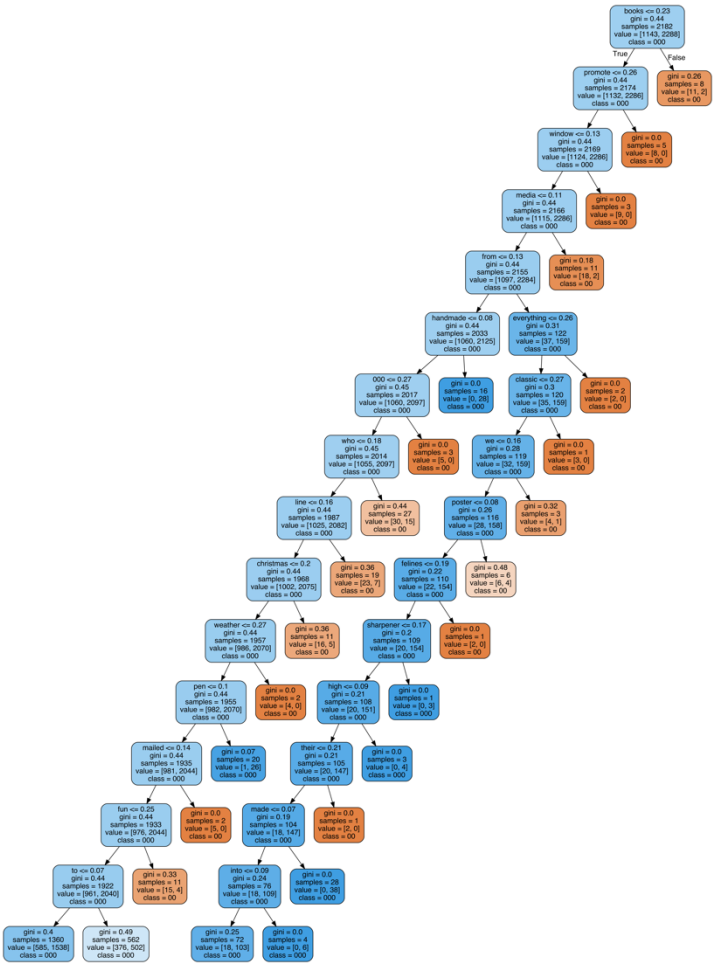
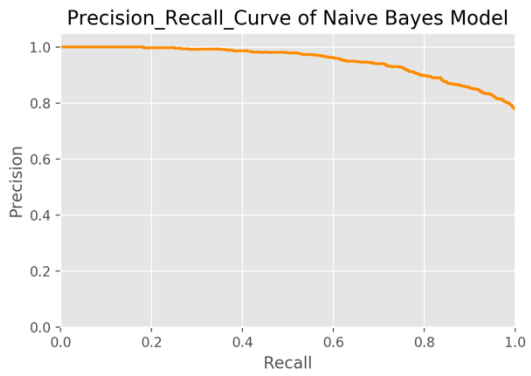
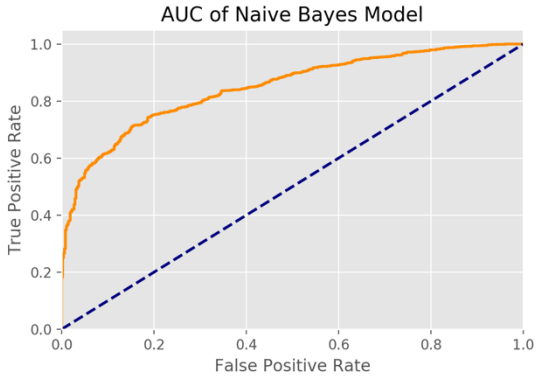


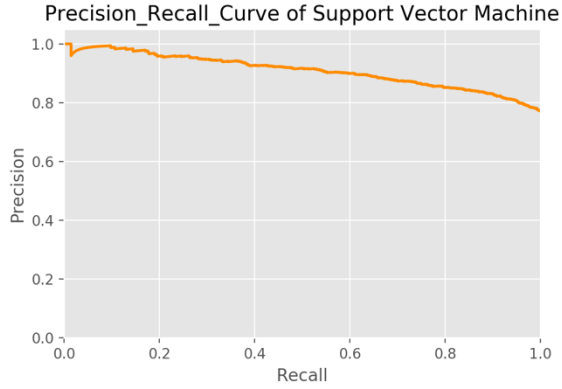

Figure Tree

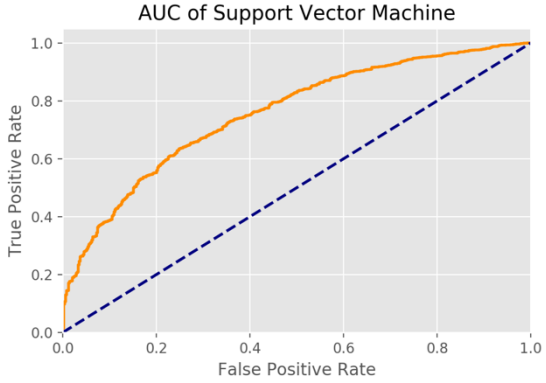
### 3.3.Performance metrics AUC and PRC

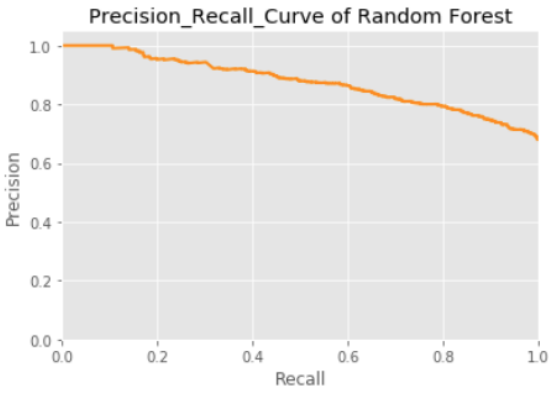
Then we can calculate the score of Area Under the Curve (AUC) and generate the plots. The AUC score of Naïve Bayes model is 84.75%, SVM is 75.6% and Random Forest is 76.62%. After calculation, the Precision Recall Curve (PRC) score of Naïve Bayes model is 95.10%, SVM is 90.89%, and Random Forest is 87.52%. The plots are shown below.

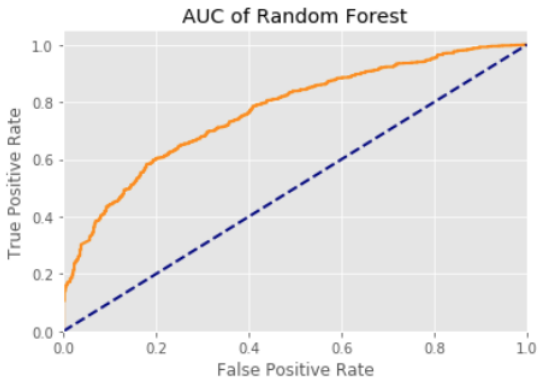





The Naïve Bayes model has moderately better performance in our project. We believe that because we have short texts for each sample, Naïve Bayes performed better.

4. Analysis of Experiment results

4.1.What part of your methodology worked or didn't work and why?

We have achieved the highest accuracy of 81% with Naïve Bayes model. It worked well because the documents, which are the brief subtitles of the products, are short snippets. Having better performance on short documents is the feature of NB model. Though the result of SVM was not too low, the results from these two models were very close.

4.2. How to improve?

Selecting an appropriate model in text classification is essential, as models have different features and disadvantages. Our project mainly uses short text as the target. In this case, a Naïve Bayes model is an optimal model to use. The results also show the superior advantages of Naïve Bayes model in analyzing short snippets. If we are analyzing a longer paragraph, SVM model and Random Forest model might have better performance.

4.3. How to utilize your results? What business insights can be derived from your analysis?

The best business insight is the application of the analysis. The result that we analyzed may be a good example that help us to know if we raise a project, it will give us some idea what kind of key words and description that we should add into the project. Otherwise, according to the analysis result, we can estimate the rough success rate of the project and decide whether the crowdfunding website deserves that we spend time and energy on it.

5. Conclusion and future work

### 5.1. Analysis of full product descriptions

The datasets online do not include the information of the video and the script of the video, but we can scrape by ourselves in the future. Our teammate Xiao Zhou has successfully scraped the video script from a couple of videos, however, due to the limited capability of our laptops, we couldn't scrape all the information in a short time. It would be very helpful to analyze the script along with the product description.

### 5.2.Analysis of video scripts

We couldn't find a dataset with the full description of the product at this time, but the description is very important to learn about before the customers make the purchase. Since Kickstarter is not releasing a lot of failed projects, we will need to figure out the way to prompt the failed projects. The next possible step is to scrape the full description of the product manually, including the rewards to the backers. Besides the text, the information we need includes the count of the videos and images in the description.

### 5.3.Conclusion

In our project, we have tried to use NB and SVM as machine learning models for classification. We also introduced Random Forest in this project as a comparison. The results show that the NB model has a higher accuracy rate than the other two models in this experiment, the scores of accuracy is from 0.70~0.73. In terms of running speed, the NB model is also the fastest, followed by Random Forest.

***References***


Chen, L. and Shen, E., "Finding the Keywords Affecting the Success of Crowdfunding Projects," 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA), Tokyo, Japan, 2019, pp. 567-571, doi: 10.1109/IEA.2019.8714815.

Chen, W., Zhu, K., Wang, W., & Wang, H. (2016). The success rate of crowdfunding financing and the persuasiveness of language style: Empirical research based on Kickstarter. *Management World, 5*. doi:10.19744/j.cnki.11-1235/f.2016.05.008

Kamath, R. S. and Kamat, R. K., Supervised Learning Model for Kickstarter Campaigns With R Mining (2016). *International Journal of Information Technology, Modeling and Computing (IJITMC)* Vol. 4, No.1, February 2016, Available at SSRN: https://ssrn.com/abstract=3513341 or http://dx.doi.org/10.2139/ssrn.3513341

"Kickstarter Datasets." Web Robots Kickstarter Datasets, April 22, 2021. https://webrobots.io/kickstarter-datasets/.

Li, Guan-Cheng. "Kickstarter Structured Relational Database." Harvard Dataverse. Harvard Dataverse, January 31, 2019. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FEOYBXM.

Sawhney, Kartik, Caelin Tran, and Ramon Tuason. "Using Language to Predict Kickstarter Success," n.d.

Song Y, Berger R, Yosipof A, et al. Mining and investigating the factors influencing crowdfunding success[J]. *Technological Forecasting and Social Change*, 2019, 148:119723-.

Wang, W., Zhu, K., Wang, H., & Wu, Y. J. (2017). The impact of Sentiment orientations on successful crowdfunding campaigns through text analytics. *IET Software*, 11(5), 229-238.

Project task table

| | Tasks | Members assigned | Signatures |
|---|---|---|---|
| 1 | Proposal | All | XZ, XZ, QY, TR |
| 2 | Data collection / processing (midterm) | Xiao Zhou, Xiaoyu Zhou | Xiao Zhou |
| 3 | Midterm report | Xiao Zhou, Xiaoyu Zhou, Qinglin Yang | XZ, XZ, QY |
| 4 | Analyze dataset by clustering | Tianhao Ruan | Tianhao Ruan |
| 5 | Implement NB\SVM\RF models | Xiao Zhou | Xiao Zhou |
| 6 | Testing different datasets | Xiaoyu Zhou | Xiaoyu Zhou |
| 7 | Results Analysis | Xiaoyu Zhou | Xiaoyu Zhou |
| 8 | Final report writing | All | XZ, XZ, QY, TR |
| 9 | Final presentation | All | XZ, XZ, QY, TR |