

Application of a Text Analysis Approach in O2O Customers Service Records Categorization

Zhang Yichi

University of Tampere
School of Natural Sciences
Software Development M.Sc. thesis
Supervisor: Jyrki Nummenmaa
July 2018

University of Tampere

School of Natural Sciences

Software Development

Yichi Zhang: Application of a Text Analysis Approach in O2O Customers Service

Records Categorization

M.Sc. thesis, 71 pages, 4 index and appendix pages

July 2018

With the O2O industry entering oligarchic era, many O2O companies in China like DiDi face the challenges of acquiring the helpful information to improve the service quality and obtain the new requirements. As the big data processing and machine learning is getting popular, one of the direct to fulfill the needs of DiDi is to acquire the information from the customer service records. With that purpose, automatic classification of the customer service records data is the initial step.

This thesis aims to find a solution for DiDi to categorize their customer service records data into the pre-defined categories. The solution is based on the traditional text categorization flow and introduces a way to build enhancement feature collection instead of using the original feature collection. Classic supervised learning algorithms in the traditional text categorization flow are also demonstrated in the thesis.

According to the regulation in the expression of customer service records data. The pre-defined syntax pattern is introduced to build the enhancement feature collection. The text analysis approach Dependency Parser is first introduced to obtain the syntax pattern from the document.

Tests are conducted to compare the performance of most of the algorithms mentioned in the thesis and the best ones are chosen to be applied in the final solution. The performance tests are also made to prove the better performance in using the enhancement feature collection than the original feature collection.

Key words and terms: O2O, text categorization, classifier, feature selection, dependency parser, syntax pattern

Contents

1. Introduction	1
2. O2O Business Model and Introduction to DiDi Case	3
2.1. O2O Business Model.....	3
2.1.1. O2O Operation Mode	5
2.1.2. Features of O2O	7
2.2. The Development of O2O in China.....	9
2.2.1. Chaos stage	9
2.2.2. Germination stage.....	10
2.2.3. Primary development stage	10
2.2.4. The status of Chinese O2O.....	12
2.3. Introduction to the DiDi case	14
2.3.1. The challenges of DiDi service	15
2.3.2. The brief introduction to DiDi customers service records data.....	16
2.3.3. Features of the DiDi customers service records data	17
3. Traditional supervised text categorization.....	19
3.1. Traditional supervised text categorization flow	20
3.2. Document Representation	22
3.2.1. Bag-of-Words Model(BOW).....	22
3.2.2. Vector Space Model	23
3.2.3. Term Frequency-Inverse Document Frequency(TF-IDF).....	24
3.3. Feature Selection	24
3.3.1. Document Frequency (DF).....	25
3.3.2. Information Gain (IG)	25
3.3.3. Pointwise Mutual Information(PMI).....	26
3.3.4. χ^2 Statistic (CHI)	27
3.3.5. Summary of The Feature Selection Approaches	27
3.4. Classifiers	28
3.4.1. Naïve Bayes(NB).....	28
3.4.2. K-Nearest Neighbor(KNN)	29
3.4.3. Decision Tree(DT).....	29
3.4.4. Support Vector Machine(SVM)	31
4. Syntax pattern application in short text categorization	33
4.1. Approaches of short text categorization	34
4.1.1. N-gram.....	34
4.1.2. Feature co-occurrence set	35
4.1.3. Syntax Pattern.....	36
4.2. Dependency Parser	38
5. A solution for DiDi customer service records classification.....	42

5.1. Implementation.....	44
5.1.1. Feature Enhancement	44
5.1.2. Document Representation.....	47
5.1.3. Feature Selection and Model Training	49
6. Testing and evaluation.....	51
6.1. Full features data collection testing.....	51
6.2. Traditional feature selection methods testing.....	54
6.3. Enhancement features data collection testing.....	60
6.4. Discussion	62
6.4.1. Limitation and future work.....	62
7. Conclusions	64
References	66

Acknowledgements

I would like to sincerely express my thanks to my supervisor Prof. Jyrki Nummenmaa for his patiently guidance and selfless help through the process of my master thesis. I also wish to thank my second reviewer Prof. Jaakko Peltonen who gave me specific corrections on my thesis to highly improve its quality. Without them, the accomplishment would be hardly achieved.

Finally, I would like to thank to my parents, girlfriend and all the friends for their significant support and encouragement which accompany with me in every day and night throughout my study period.

1. Introduction

With the wide application of the internet in life, e-commerce is no longer satisfied with staying online and making simple virtual business and convenient shopping experiences. Based on the explosive increase of smartphone use and mobile applications, a new type of business model known as Online to Offline combines online transaction with offline service. Compared to the traditional online business model O2O is more focused on consumption of services including catering, film, beauty, SPA, tourism, health, car rental, house rent, etc. [Stan, 2011] which means traditional e-commerce model concentrates more on merchandise quality and requirements while O2O model is customer-centric, to meet the different needs of customers, and to create a good online and offline experience for them as well [Shen, C., & Wang, Y. ,2014].

Considering the customer-oriented model build, analyzing the customer feedback, advice and even complaints becomes a significant approach to seek out the potential problems with both offline service and the online business platform. One of the most immediate methodologies for obtaining the customer needs is to review the customer service records. However, with the huge amount of structured or unstructured reports stored in the customer service database [Hui, S.C., & Jha, G., 2000], it is impossible to manually review and extract valuable information from the customer service records.

The traditional method to obtain helpful information from customer service records is similar to the brainstorming requirements elicitation method in requirement engineering [Pohl, K., 2010]. The experienced customer service team members are selected as domain experts to participate in the brainstorming workshop to gives feedback and advice to the service management team and the platform development team. The service management team and development team will try to acquire customer requirements and strategic decision through the analysis of the advice and feedbacks. Nevertheless, with the integration of a few O2O platforms, the offline service district coverage of every single O2O platform increases extremely fast which brings a great challenge to this method in precision and effectiveness.

Together with the popularization of distributed file systems, a variety of big data platforms like Hadoop and Apache Spark have appeared to enable the O2O platform-based enterprise to build big data process platforms with existing server clusters. Advanced machine learning methodologies become available in big data processing of the O2O platform enterprise. To automatically extract valuable information from customer service records, text mining is considered to be a proper technique. According to Tan, 80% of a company information is contained in text documents [Tan, A. H., 1999].

So, compared to data mining through customer behavior data statistics, text mining would extract more useful and specific information on the common problems in the customer experience in both service and platform use.

Text mining is a multidisciplinary field, involving information retrieval, text information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining [Feldman, R., & Sanger, J. 2007]. Within text mining, text categorization plays a significant part in minor information filter and main information elicitation. By applying the machine learning algorithms in the text categorization, the customer service records data could be automatically categorized more accurately and effectively than manual categorization.

In this thesis, the traditional flow of the text categorization will first be introduced. The approaches and algorithms in every step of the text categorization including document representation, feature selection, and categorization function inference will be demonstrated and further discussed. The differences between short text categorization and long text categorization will be discussed. Finally, an application of syntax pattern in the pre-processing step of text categorization to construct a new text feature set will be presented. A text analysis-based approach named dependency parser will be applied to acquire the syntax pattern of the text sample.

2. O2O Business Model and Introduction to DiDi Case

As a new type of e-commerce business model, O2O is more customer oriented and more focused on the consumption of service than the traditional e-commerce business model. With recently stepping into the oligarchy time, the Chinese O2O platforms are even slowly integrating the service part into themselves instead of being co-operators of offline service companies. This developing trend of O2O platforms makes customer service records play an even more important role in improving the online platform quality and the offline service quality. This chapter describes the specific definition and current situation of O2O business and shows the significance of the customers' service to O2O platforms. It will also introduce DiDi and present their demands for mining useful information from the customer service.

2.1. O2O Business Model

The concept of the O2O business model was first raised by an American scholar Alex Rampell in August 2010. When analysing the similarity of Groupon, OpenTable, Restaurant.com and SpaFinder companies, Rampell found that all their business models combine the online platform business with the offline service business. Therefore, in August 2010, Rampell [2010] summarized this new kind of business model as promoting and attracting customers online, then leading the customers to the offline shop, store, restaurant or bar. He named this business model “Online to Offline” i.e. “O2O”, which keeps the consistency with the business terminology words e.g. “B2C”, “B2B”.

With growing popularity, the concept of O2O is becoming increasingly generalized. From the perspective of Wang [2013], it is difficult to explain the phenomenon of O2O because there is no clear definition of O2O industry. As the name implies, the entire industry chain that relates both online and offline can be called O2O. For example, the O2O can include not only the group purchase, coupons and the other local daily services but also the traditional B2C industry services, for instance, furniture, cabinet and some other retailers. In the opinion of Yang, Liang, and Wu [2014], O2O is an integration of an online channel and an offline channel. Specifically, it means that online promoting, online purchase and payment lead to the offline sale and consumption. O2O makes the online platform a uniform front desk for the offline service suppliers. The customers can directly enjoy the pure service without experiencing the annoying recommendation from the salesperson or wasting time at checkout. Li and Dou [2013] explain O2O as a chance to unite the Internet and the local service. The Internet plays a role as the front desk in transactions and the advertising board for promotion. This mode frees both the shopkeeper and the customers. The shopkeepers can derive their marketing strategy from

the customers' choices and the customers can filter the shopkeepers who receive negative feedback. Li and Dou also pointed out that O2O will be a trend in the future to realize the "Internet plus" (a concept raised by Chinese IT industry to develop an extension of the internet to connect everything in life). Overall, with the upgrading of the internet devices, the increasing appearance of new forms of O2O industries makes the definition of O2O even more varied than any other e-commerce business model.

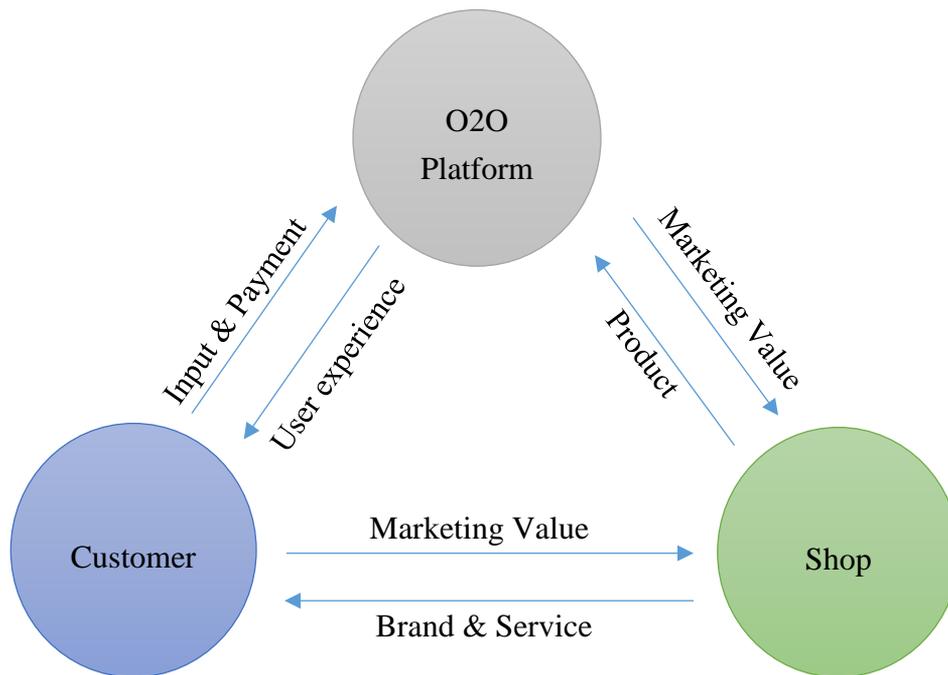


Figure 1 basic O2O business model

Figure 1 shows the basic business model of O2O. Different from the traditional business model, the numbers of essential elements in the transaction progress vary from customer and shop to customer, shop and O2O platform. Inside this triangle relationship, O2O platform acts more like a medium or an agency. It keeps a bidirectional relationship with both the customer and the shop and covers the whole transaction and promotion functions which look almost the same as the offline agency and medium. However, unlike an offline medium and agency, the O2O platform makes the transparency of the transaction progress possible, enhances the shop marketing effect and simplifies the whole "front desk" service for the customers, due to its online product attribute. Compared to the refunds and the advertising investment from the shop, O2O cares more about the number and satisfaction of the customers. This feature results in the independence of O2O and builds a relatively objective word-to-mouth system of the shops. With the slow introduction of the customer credit system into the O2O platform, the bidirectional pre-choice of the customer and the shop is also gradually realized.

2.1.1. O2O Operation Mode

Following the basic model of O2O, many companies and even small shops have started to join the O2O business and create the operation modes with their own characteristics in specific fields. According to research, most of the operating modes can be divided into four categories [Guo, 2015]:

1. **Online to Offline mode.** This mode refers to the business that starts from online transaction and moves to offline consumption experience. This mode is the most classical and common mode of O2O business. Most of the O2O platforms for insurance Groupon, Uber, Foodora, Parkman and Booking now follow this mode to finish the transaction online and let the customer enjoy the service offline. This business mode is suitable for the platform based O2O companies to attract both customers and shops (co-operators) and build the uniform O2O entry to the strict field.
2. **Offline to Online mode.** This mode refers to the business that starts from the offline promotion or consumption experience (service) and moves to the online transaction. This mode has become more popular with the spread of the mobile internet and the application of QR codes in online payments [Lu & Li, 2013]. The most famous example of this mode may be the Amazon Go (<https://www.youtube.com/watch?v=NrmMk1Myrxc>). Amazon tries to build a supermarket offline which puts all the order and purchase progress online. The customers only need to pick the merchandise they want and go out of the supermarket. The order list will be automatically generated in the smartphone of the customers and they can fully finish the transaction with their Amazon account. However, Amazon Go is still at the conceptual stage while in China this kind of mode has already been implemented with the help of Alipay and Wechat.



Figure 2 Scanning QR code for the online transaction

As shown in Figure 2, Some local supermarkets try to use the QR code as the price tag for the customers to scan to finish the online transaction. If this mode is simply defined as moving offline cash or card payment to online payment, this mode is already quite popular among many shopkeepers in China.

3. **Offline to Online to Offline mode.** This mode refers to a business that starts with offline promotion and moves to the online transaction then back to offline consumption experience (service). Obviously, this mode is an extension of the two modes above. Specifically, it refers to the companies which have done a period of promoting offline to promote their online transaction then offer the service to customers offline. Since most of the O2O platform-based companies are used to doing promoting online, this mode can usually be seen in the offline service based companies Such as restaurants which cooperate with a famous O2O platform to encourage their customers to use the discount and Groupon they serve and sell in the platform
4. **Online to Offline to Online mode.** This mode refers to the business that starts from online transaction or promotion and moves to offline consumption experience and then to the online transaction. This mode is rarely seen in a service based O2O platform. Nevertheless, this mode is now popular in some B2C platforms to make consumption progress visible for the elder customers. For instance, the famous Chinese B2C company JD.com has started to build offline shops in many big cities in China. The customers can finish the order in their online store and then directly pick up the merchandise they order from the offline shop. These shops are not meant to attract customers to pick merchandise offline. On the contrary, these shops are meant to help the elderly customers to experience the convenience and security of online shopping, and finally, transfer themselves to fully online customers.

2.1.2. Features of O2O

Compared to the traditional e-commerce business models, the O2O business model holds a longer consumption lifecycle. The offline part takes the same or even a higher share than the online part. This kind of difference makes O2O a customer-oriented business, instead of the merchandise-oriented, traditional business model. Moreover, in the beginning, most of the O2O platforms are suppliers of online transaction and comprehensive information. They usually choose to cooperate with the local service suppliers instead of offering the online to offline service all by themselves. The independence of the two parts makes O2O benefit more related companies and shops while bringing the O2O its own following features.

1. Mobilization

As introduced by Guan Xu and Wang [2015], connecting to the internet is no longer a limitation with the increasing popularity of smartphones and other mobile devices. This trend frees the online transaction from being an in-door behaviour.

Combined with the service leading consumption experience of O2O, the consumption habit of the customers also changes from reviewing online and choosing the offline service to review the service supplier offline when visiting it. Among many famous O2O platforms, many of them are more likely to focus on the mobile clients or they are even based on them. Although some B2C platforms are also slowly trying to become more mobile client-oriented, most of the customers are still used to finishing the consumption on their desktop.

2. Localization:

Also raised by Guan Xu and Wang [2015], as the service plays the role as main merchandise in the O2O, the development of O2O platform relies more on their cooperation with the local service suppliers or even their own service position.

This feature brings sometimes challenges to the O2O platforms. Compared with the B2C the offline operation is no longer just building and maintaining the delivery system, the O2O Offline operating requires the platform to maintain the offline service (i.e. physical shops), supervise its service quality and find more reliable offline co-operators to build their own business network. However, this feature also makes O2O businesses more agiler and easy to start, especially for those offline shops who want to find an entrance to do online promotion or even build their own online part.

3. Experience-orientation

As mentioned above, instead of the merchandise, service becomes the main part of the O2O business. Therefore, in comparison with the B2C or other traditional models, O2O is an experience-oriented consumption progress. The main reasons for O2O business being attractive to customers are not just the so-called bridge between online and offline or its convenience, but the whole experience of consumption should be included [Guo, 2015].

Indeed, in the traditional e-commerce business model of consumption, the customers will usually not concern too much about the attitude of the online salesman or the efficiency of the delivery if the quality of the merchandise is good enough. The service of the platform will be paid attention to only if the customer runs into aftersales problems. The key point of maintaining the loyalty of the customer is just supervising the quality of the merchandise itself. However, in the O2O business, the whole consumption progress is about the experience. Therefore, usually, the customers will not divide the online experience and offline experience clearly in their mind. This type of customer consumption habit makes them more concerned about every part of the service including the convenience of the transaction, the recommendation of platform, the convenience of transformation of online receipt to offline service and, of course, the quality of offline service.

4. Interaction:

Obviously, O2O businesses build a very important bridge between customers and offline shops. In the past, the customers have only been able to interact with shops after they reach the shops and end the interaction once they step out. It is hard for the shop to collect enough feedback from the customers and for the customers to know the quality of the shop. The O2O platform offers the best solution for both sides.

Through the O2O platform, the limitation of the time and place of interaction between customers and shops disappears. The interaction between the two parts is brought forward before the transaction happens. The shops can do promoting and answer the questions of the customers, while the customers can review the comments to the shops from other customers to form an opinion about the general service quality. In addition, the interaction is also extended after the end of the transaction. The shops can easily review the feedback from the customers, meanwhile, the customers can feel totally free to express their true feelings about the service quality.

2.2. The Development of O2O in China

China is one of the countries that started the O2O business model very early. Two years after when the concept of O2O was first raised by Rampell [2010], the first group of O2O businesses began to appear in China. Although with few kinds of service and simple applied of operation modes this newly imported e-business model still soon was accepted by local people and soon popular with the start-up circle.

According to Tian [2017], by now the development of Chinese O2O business can be separated into three stages which are the chaos stage, the germination stage and primary development stage.

2.2.1. Chaos stage

In this stage, most of the O2O businesses inherited the idea from the traditional B2C business model. Without high-speed and wide coverage mobile networking, online payment processes, and QR code identification could barely be used apart from the PC clients. This kind of technical barrier became also a limitation for the ability of online platforms to reach offline service suppliers. Therefore, during that period, the O2O platforms in China mainly played the role of an information integration platform, and most of the information is supplied by the customers. The classical representation of that period is DaZhongDianPing which is known as the Chinese version of Yelp.

The screenshot displays the DaZhongDianPing (大众点评) website interface. At the top, there is a navigation bar with options like '写点评', '你好, 请登录', '免费注册', '个人中心', '商户服务', '帮助中心', and '网站导航'. Below this is the main header with the '大众点评' logo, a search bar, and a QR code for mobile access. The main content area features a restaurant listing for '桐记小灶儿牛板筋火锅(悠唐购物中心店)'. The listing includes the restaurant's name, a '手机买单 积分抵现' badge, a '添加分店' button, and a star rating of 4.8 based on 1934 reviews. The average price is listed as 106元, with ratings for taste (8.8), environment (8.7), and service (8.8). The address is '3丰北里2号楼悠唐购物中心4层001号' and the phone number is '010-59775951'. There are also buttons for '写点评' and social sharing options. Below the listing, there are '优惠促销' (Promotions) and '推荐菜' (Recommended Dishes) sections. The '优惠促销' section shows two deals: '超值午间单人套餐' for ¥68.00 (original ¥112.00) and '精品午间单人套餐' for ¥78.00 (original ¥118.00). The '推荐菜' section lists various dishes with their prices, such as '香酥焗饼 (229)', '香辣牛板筋锅 (229)', '港式芝士丸 (148)', '泡菜牛板筋锅 (141)', '原味牛板筋锅 (120)', '干叶豆腐 (64)', '咖喱锅 (54)', '香辣牛杂锅 (52)', '韩式辛拉面 (50)', '韩式清汤牛尾锅 (49)', '毛肚 (31)', '蟹子丸 (28)', '雪花肥牛 (26)', '墨鱼丸 (22)', and '金桔冰糖茶 (22)'. A '手机抢红包, 还能赢免单!' (Mobile lucky draw, can win a free meal!) banner is visible at the bottom right.

Figure 3 Example of the detail information shown in the user interface of DaZhongDianPing

Figure3 presents the detail information of a random restaurant in Beijing, although the main business of DaZhongDianPing has changed the website still keeps its original function of providing restaurant information and customer reviews. Together with the increased effectiveness of the platform, allowing the restaurant hosts to be the sponsors and improve their review scopes once became a main profit method of the website. This kind of phenomenon indicates that the Customer-orientation feature of the O2O model had already appeared since the very beginning of the O2O business.

2.2.2. Germination stage

An example of this stage is the popularity of Groupon. Compared with the diversity of systematical O2O businesses today, the Groupon can no longer be strictly defined as a typical O2O service. However, Groupon has proved to be the strongest, easiest to duplicate and most widespread O2O business. From the end of 2010 to the middle of 2011 in the fastest growth period of Groupon business in China, the user number of Groupon business websites grew from 18.75 million to 42.2 million. The rate of increase reached 125.0% [Huang,2012]. By the end of 2012, the total amount of the Groupon websites in China was above 1000. The Chinese O2O business was entering a famous period of “Thousand Regiments”.

Unlike the O2O business in the first stage, the Groupon O2O business basically implements the business model of an O2O service. In the Groupon business, the service suppliers are more interactive than the first stage business. The online platform comes up with a reasonable and continuable profit model for the first time which is to get the profit of each group order from the service suppliers. Through the short message service, the customers got the identity code of their Groupon orders. In this way, the customer finished a whole consummation progress truly online for the first time. It can be said that Groupon business reached a serious milestone of O2O business. In the meantime, Groupon business explored a variety of Offline services which could be possibly extended in the future.

2.2.3. Primary development stage

This stage is presented by Tian as a stage of the diversity of the O2O business. With the development of the mobile network, the online transactions of O2O are no longer limited

to the non-mobile network clients. More and more O2O businesses are trying to become independent from the traditional integration into a Groupon website, especially the one on one service businesses including the transport service (for instance DiDi), housekeeping service, take-out service etc. No matter whether successfully running or failed, growing numbers of service industries are willing to step out to get closer to customers and make the offline more focus on the service itself and take into more serious consideration the customer feedback and consumption experience.

Meanwhile, with the appearance of the P2P (point-to-point) form of O2O business, more customers are transferring their role from customers to service suppliers. The typical example is DiDi, with increasing numbers of people, join as drivers to DiDi, the service attitude, and quality of the taxi industry have improved a lot. The problem of “hard to take a taxi” is also solved or at least alleviated in many cities. However, new O2O business attempts to bring new challenges. According to Li [2016], at this stage, a huge amount of the O2O enterprises in China will go out of business. Li presents that in regardless of the different specific reasons for the failure of the companies, some common problems of the O2O mode itself in this stage can be inferred:

- **Lack of innovation:**

Many of O2O industry practitioners who experienced the Groupon period still hold the inertial thinking that the O2O model is an easily copied model for gaining profit. Simply copying the model of an O2O business make them struggle in the fierce competition of hundreds of similar O2O companies and finally they become swallowed by the big enterprises.

- **Poor localization of the market:**

Introduced by Li many of the O2O companies localized the wrong market group from the very beginning of their business. This problem resulted in the direct failure of the later operation. The wrong localization of the market group will seriously reduce the consumption requirements of the consumers. This problem is caused mainly by the shortage of experienced managers and inaccurate analysis of the customer needs.

- **Poor standard of the service:**

Demonstrated by Li [2016], one customer will tell his dissatisfaction to eleven customers, the loyalty of the customers will reduce by 5%, and the profit of the enterprises will be reduced at least by 25%. Lacking specific analysis of customer feedback to formulate a suitable standard for the O2O service is the core issue of this problem.

From the descriptions of problems above, it can be indicated that apart from hiring more experienced managers and operators for creating marketing strategies, analysis of the customer-related data is a more accurate and direct way to solve the core issues of the problems.

2.2.4. The status of Chinese O2O

Now, the Chinese O2O has reached the end of the third stage. As predicted by Tian [2017], the O2O business will enter into a fourth stage which is called the high-speed development stage. In this stage, much of the traditional industry will step into the area of O2O and the service will become more personality oriented. The relationship between data, customers and multiple types of resource will be integrated to affect the development of O2O. O2O business will be the main trend of the future business model.

In the latest report of the development status of O2O business in China, iResearch [2017] mentions the same status of the Chinese O2O business with more evidence. According to the report, by the end of the year 2016, the O2O market scale in China has reached 665.94 billion Yuan (about 83.24 billion euros), increased 42.7% compared to the year of 2015. Based on this trend, and the existing O2O industry data, iResearch predicts that by the end of 2017 the scale of the O2O market will reach over 900 billion Yuan (about 112.5 billion euros) which will grow 28.3% compared to the scale of the year of 2016. Nevertheless, on the contrary, the penetration rate of the O2O business market in China is only 10.8% as predicted by iResearch at the end of 2017 which means only 10.8% of the traditional offline service business have been covered or partly covered by O2O business. The low penetration rate is a good signal to both O2O service suppliers and platforms which means there is still an enormous potential of the O2O market in China to be tapped. By now the O2O business has permeated into many industries including healthcare, traveling, housekeeping, local transportation, interior trim, cloth washing, food delivery & restaurant etc... Moreover, the report says that from the financing situation of the O2O business companies at least more than 50 traditional industries will join the family of O2O in the coming year.

In the opposite of the prosperous situation of the new coming industries of O2O, the early bird O2O industries are facing fierce competition in these years. With the high-speed growth of O2O in recent years, the markets of some early bird O2O industries are slowly becoming saturated. The customers are losing their enthusiasm for some unpractical requirements of these service industries. At the same time due to the problems I mentioned above most of the O2O companies failed to detect the new real requirements and the requirements changes in their fields. To fight for the profits of the few practical requirements in the industries, the companies choose the simplest method- "Money Burning". According to the report of Ebrun [2017], since the end of the third stage, most of the O2O companies had long stayed at a loss. O2O companies devoted most of their

money to the consummation allowance to the customers and the rewards to the cooperation service suppliers. The income of the companies is mainly from the rounds of investment. However, together with the saturation of the market and the decreased enthusiasm of the customers, the winter of the O2O investment also arrived. A huge amount of small companies fell in this period, the integration of the rest into large companies was continually going on in this stage.

Name	Established Time	Class of Business	Name (after integration)	Integration Time
DiDi	07.2012	Smart private taxi service	DiDi	14.02.2015
KuaiDi	05.2012	Smart taxi service		
WuBa	12.2005	Classified life info platform	WuBaGanJi	17.04.2015
GanJi	03.2005	Classified life info platform		
DaZhong DianPing	04.2003	Local Service info and transaction platform	MeiTuan + DaZhong DianPing	06.10.2015
MeiTuan	03.2010	Groupon service platform		
Ctrip	10.1999	Travel agency	Ctrip + Qunar	26.10.2015
Qunar	02.2005	Travel agency search platform		
MoGuJie	02.2011	Woman fashion e-commerce platform	MoGuJie + MeiLiShuo	11.01.2016
MeiLiShuo	11.2009	Woman fashion e-commerce platform		
Ele.Me	04.2009	Food delivery platform	Ele.Me + KouBei	14.04.2016
KouBei	12.2013	Food delivery platform		

Table 1 The integration cases of O2O magnates within 2015-2016. From Qi [2016]

From the Table 1 presented by Qi [2016], it can be inferred that from 2015 to 2016 within this short period six cases of integration between O2O companies happened. Among these cases only one belongs to the capital acquisition, the rest all belong to the group operation which clearly shows the hardness of maintenance of these companies in this period. After the round of integration in these O2O industries the early bird O2O industries entering the time of oligarchies. Although this reduced the fierce competition, the basic problems of most of the O2O companies mentioned above are still not solved. Moreover, from the perspective of Qi [2016], the integration will definitely bring

challenges to the new companies such as the merging of the employees, the potential chaos of the company structure and management. Among them, the most important challenge is the maintenance of the service quality and the acquisition of the potential requirements of the customers to extend the business.

2.3. Introduction to the DiDi case

DiDi is one of the early bird local transportation companies. It was founded in June of 2012. In the beginning, the business of DiDi was similar to Uber which mainly attracted the car owners to register as private taxi drivers and offer their service to customers. After two years of operation, DiDi successfully received investment from Tencent and started an “allowance battle” with KuaiDi in which they give allowance to the drivers and the customers at the same time for building their loyalty to the platform. By February 2015, DiDi integrated KuaiDi and became the largest O2O local transportation platform in China. By the same time, DiDi has expanded its business areas to more than 10 fields in local automobile transportation.

Business Name	Business Description
DiDi Taxi	DiDi is a leader in the digital transformation of the taxi industry. The integrated big data approach enables continuous improvement in both service and driver's income.
DiDi Express	DiDi Express is the world's largest affordable mobility network providing individual and pooled rides across all income groups.
DiDi Premier	DiDi Premier offers higher-end premier mobility experience with luxury vehicles and drivers trained to the highest service standards.
DiDi Hitch	DiDi Hitch is the world's largest social carpooling platform that helps commuters save money, make the right friends while helping conserve energy.
DiDi Designated Driving	DiDi provides designated driving service for commuters, partygoers and business travelers who can now sit back, relax and enjoy a ride free from the hassle of driving.
DiDi Bus	Besides providing real-time public bus schedule services, DiDi Bus helps plug-in DiDi's data analytics into the digitization of urban public bus systems.

DiDi Car Rental	DiDi Car Rental provides hassle-free, door-to-door online car rental services, which are now extended to over 175 countries around the world through global partnerships.
DiDi Enterprise Solutions	DiDi Enterprise Solutions provides corporate clients with flexible, efficient and reliable corporate mobility solutions free from reimbursement and processing pains.
DiDi Minibus	DiDi Minibus helps solve the "last mile" challenge for urban transportation by shuttling commuters in 5- and 7-seaters between transportation hotspots, including public transit terminals.

Table 2 The main businesses of DiDi since 2015. From DiDi [2017]

The Table 2 provided by DiDi [2017] demonstrates the main businesses of DiDi after the year 2015. From the figure, it can be seen that since 2015 DiDi has extended its business into almost all the local private and public transportations. Among all the businesses, apart from the original two businesses of Taxi and Express, most of the branch businesses are built to depend on the integration of the related companies or business models copied from other competitors. Standing on the top of the O2O field DiDi still faced the same problems of continuing business innovation. By the summer of 2016, DiDi received 1 billion dollars investment from Apple and integrated the last public transportation O2O tycoon-Uber China. Since then, DiDi became the biggest public transport O2O platform company which covers more than 90% market shares and have its businesses spread into more than 400 cities in China.

2.3.1. The challenges of DiDi service

Chances always come with challenges. Considering the marketing situation and the pressure from the investors, DiDi wanted to end the ‘money-burning’ business competition strategy. Therefore, most of the same time, DiDi started to reduce the allowance to both the customer side and the driver side. Without the allowance support to maintain the growth of the users, the problems behind the services of the platform were magnified. From the beginning of 2017, the number of high-frequency users started to decrease, meanwhile, the growth rate of new users also started to decrease. Moreover, negative comments and news were continually appearing on some popular social media and communities.

Unexpectedly, more than 70% of the negative information was about the replies of the customer complaints. More specifically, most of the customers’ dissatisfactions are about

the solutions to their complaints. For instance, some drivers reflected that the customer service is partial to the customers when dealing with the customer drive conflicts, whereas the customer reflected that some drivers required unreasonable tips during rainy days, but the customers replied that operating on those days the choice of the drivers.

Obviously, after many years of operation, DiDi still didn't find a good method to properly acquire the requirements from the users and improve their service standard and their business strategy. Luckily, with the popularity of the big data analysis and machine learning the collected customer data like customer service records or customer order records are no longer data that only stays on the server. With the application of text analysis or machine learning methods, not only the general trends but even the existing specific problems and user preference can be discovered.

2.3.2. The brief introduction to DiDi customers service records data

The data sample used in this thesis is real data supplied by the DiDi customers service backend. The original sample of the data contains over 22000 pieces of customers service records from over 10 cities in one week. The original data are from both the customer side and driver side and have been manually categorized into three-level categories which contain 32 first level categories but 500 third level categories. As described by my friend in DiDi, the main reasons for the wide range between the number of the first level and third level categories is the diversity of the categorization standard in different branch offices.

订单订单信息查询订单信息	司机来电，乘客下错订单地点了，没取消定单，已告知，此订单乘客已经取消了								
订单订单信息查询订单信息	此张订单关闭了，还会有车来，乘客称已经提示没有司机接单了，告知让乘客重新发单。								
订单订单信息查询订单信息	司机来电称接了个订单马上被乘客取消了，怀疑是刷单已告知该乘客是选择了其他交通方式								
订单订单信息查询订单信息	司机来电咨询联系不上乘客，已告知经查询此订单为乘客取消状态								
订单订单信息查询订单信息	乘客是否已经支付，告知已经支付了。								
订单订单信息查询订单信息	司机来电查询订单 已告知 已支付								
订单订单信息查询订单信息	司机来电表示公里数为零，已告知解决方法								
订单订单信息查询订单信息	司机来电查询订单状态 查询 告知司机订单已完成								
订单订单信息查询订单信息	司机来电查询订单信息已查询告知服务中订单改派完成								
订单订单信息查询订单信息	查询订单，告知已经完成。								
订单订单信息查询订单信息	师傅来电称订单公里数不对，已告知需要与乘客协商后致电更改								
订单订单信息查询订单信息	司机来电乘客要求返回起点，告知记录								
订单订单信息查询订单信息	司机来电咨询乘客未上车点击了开始计费结束计费并且付款，告知将钱还给乘客以后正常服务即								
订单订单信息查询订单信息	已告知司机师傅，乘客已经把订单取消了。								
订单订单信息查询订单信息	司机来电查询订单信息，已告知乘客没有支付，建议乘客重新刷新手机尝试								
订单订单信息查询订单信息	司机咨询订单状态，已告知乘客已取消。另咨询代驾电话，已告知电话。								
订单订单信息查询订单信息	乘客来电查看订单信息，告知订单司机已经代充值，不会有费用产生。								
订单订单信息查询订单信息	司机来电注册手机号无法显示乘客的联系方式，以核实告知了								
订单订单信息查询订单信息	司机来电，点击开始计费是突然出现元，已告知，没有点击结束计费，看不到价格的分类，可能								
订单订单信息查询订单信息	司机来电查询订单里程，里程不对，已告知结束计费，系统会刷新价格								
订单订单信息查询订单信息	司机询问乘客取消订单有什么影响，已告知乘客取消订单对司机是没有影响的								
订单订单信息查询订单信息	司机来电查询账户信息。已告知填写完信息后建议耐心等待审核。								

Figure 4 Part of the original sample data

Figure 4 shows some of the original customer service records data in one category, from the figure it can be clearly indicated that even in simply one category the records are in different record formats. Therefore, it could be inferred that the different customer support experts may have different habits of recording and standards of categorization. It means that the first step to acquire useful info from the customer service records should be the automatic categorizing of the original record data into some predefined uniform categories.

2.3.3. Features of the DiDi customers service records data

Due to the records are written by the professional customer service team of DiDi, some general features of all the records can be found in data collection.

司机 15903512022 来电取消订单，已告知拨打 4000000666 按 3.

Driver 15903512022 called to cancel the order, have told to call 4000000666 and press 3.

In the sample above, we can see that the record sentence is consist of two equal parts. The first part usually describes the purpose of the customer call while the next part records the solution which supplied by the customer service team. To precisely clarify the content of the customer service record, both parts use the simple subject-predicate structure for expression. To shorten the sentence the second part of the sentence is usually lack of the subject. However, both parts contain entire predicate structures which are easy for the reviewer to get the core information of the record.

3. Traditional supervised text categorization

The text categorization approaches can be divided into supervised and unsupervised. Supervised categorization approaches are applications of supervised machine learning approaches, meanwhile, the unsupervised categorization approaches are applications of unsupervised machine learning approach.

The supervised machine learning approaches refer to the machine learning task of inferring a function from labeled training data [Mohri, M., Rostamizadeh, A., & Talwalkar, A. 2012]. In supervised learning based categorization methods, the training data contains a list of training samples. Every sample is a pair of objects composed of an input sample and a desired output value. In practical use, the input sample is usually a feature vector that can represent graph information, text paragraphs or even voice pieces etc. The output usually marks the class membership of the input sample, which can be simply understood as categories. Through the analysis of the training data, the supervised learning algorithm generates an inferred function which can be used to label the new data. This inferred function is also called a model. The performance of the functions will be judged by multiple of performance indicators. Therefore, the supervised learning algorithms are also called classifiers. Some popular classifiers are KNN, SVM, DT, NB etc.

Unsupervised machine learning refers to the machine learning task of inferring a function to describe the hidden structure from “unlabeled” data [Wikipedia, 2017]. In contrast to supervised machine learning, the desired outputs are not necessary for unsupervised learning. The learning goal is described as an optimization fast given the input samples only, for example clustering the samples into a number of different clusters, where the number of different categories should be able to be the number of clusters is inputted as a parameter. Therefore, as presented by Jordan [2004], a central task of the unsupervised learning is the optimization of the density estimation in statistics. Most of the algorithms are created based on the two principles below:

1. **Similarity estimation.** To estimate the similarity of the data samples is one method to help detect borders of each cluster. Topic modeling approaches find latent directions in data which can be used to evaluate similarity, and which can be used for clustering.
2. **Clustering.** Unsupervised learning approaches usually categorize the data by calculating the similarity between different samples or can be seen as calculating

the fitness between the sample vectors and their clusters. The famous unsupervised learning method like K-means is based on this principle.

In the field of text classification, these two types of categorization approaches have their own advantages in different situations. In Chaovalit [2005], an experiment is done with both supervised learning approach and an unsupervised learning approach for opinion mining of movie reviews from different websites. From the result, the performance of supervised learning improved with the size of the training set and ultimately reached 85% accuracy in supervised learning outcomes while the unsupervised learning approach reached up to 77% accuracy but finished the task faster. Lee [2009] also indicates a similar result in his paper that the performance of unsupervised learning is slightly worse than with supervised learning, but the learning is more efficient. However, he also demonstrates that the performance of the unsupervised learning decreases seriously when the number of categories increases. Considering the large total amount of the data and marked categories in the DiDi data, the supervised learning approach will be used in this thesis.

3.1. Traditional supervised text categorization flow

The original text data, especially Chinese text cannot be used directly by the classifiers to train and infer their models. Thus, appropriate methods are needed to transfer the text sample into vectors and integer labels which can be directly accepted by the classifiers.

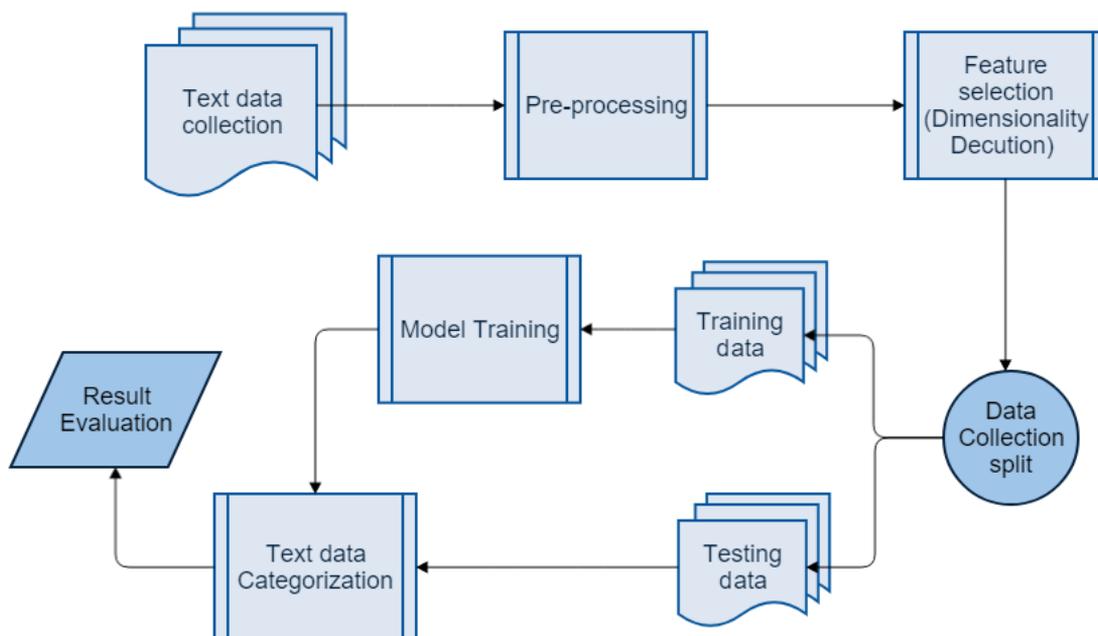


Figure 5 Traditional supervised text categorization flow

Figure 5 describes the steps of a typical supervised text categorization flow. From the figure, it can be indicated that the progress from the input data to inferring the function is separated into the following steps:

1. **Pre-processing.**

The text paragraphs are split into a set of words for further processing. Most languages can be easily separated with white space, whereas Chinese is a bit special and needs to be split with word split tools or programming packages. Then the words set is initially filtered with a list of stop-words. Stop-words refer to words that commonly appear in the text samples but are not helpful to recognize the categories of the text samples. Usually, stop-words contain punctuation, prepositions, conjunctions, articles etc. The set of filtered words set is the final set to be transferred into input objects, this step is also called document representation in the text categorization domain. According to machine learning categorization approaches the Vector Space Model (VSM) is often used in document representation.

2. **Feature selection.**

By the application of VSM in the document presentation, every text sample is represented as a vector. The specific process of transferring the text sample into the vector will be presented later in the thesis. The total set of words in the training data are used as dimensions of the vector, every word in the set is called a feature of the training data. Thus, if the number of sample in the training data is large or every single sample is a long text the dimension of the vector would be too enormous for the classifier to handle. Feature selection is the step to filter the useless words in the total words set to accomplish dimensionality deduction. In this step, features which are less related to the categorization of the text samples will be filtered out.

3. **Model training.**

After feature selection, the new training data which is constructed by the text sample vectors are ready to be inputted into the classifiers for model training, that is, inferring the function to categorize the new data. According to the common procedure in text categorization research the original data will be partitioned into a training set and a testing set, the partition ratio is usually 8 to 2. Partition approaches where the partition is done several times such as 10-fold cross-validation or 3-fold cross validation are also sometimes put in use. The

system performance is evaluated in multiple ways which will be introduced later in the thesis.

3.2. Document Representation

Document representation is the most important step of the pre-processing. Through document representation, the labelled text data will no longer be an exact representation of the text. Only the values will be used to present the importance of every feature and the characters of one sample within the whole data collection. The usual way to do document representation is to use the Vector Space Model. However, the feature improvement based method like N-grams is also used depends on the feature of the data set.

3.2.1. Bag-of-Words Model(BOW)

The Bag-of-Words model is one of the earliest and easiest models of document representation. The idea of Bag-of-Words is naïve and simple: to treat the text sample as a bag of words. The order of the words is not important. The words are used to build a dictionary with every word being expressed with an index, the dictionary has nothing to do with the order of words in every text sample. According to the dictionary, a first-order matrix will be built to represent every sample, each element of the matrix is the Term Frequency of a particular dictionary word in a particular document sample.

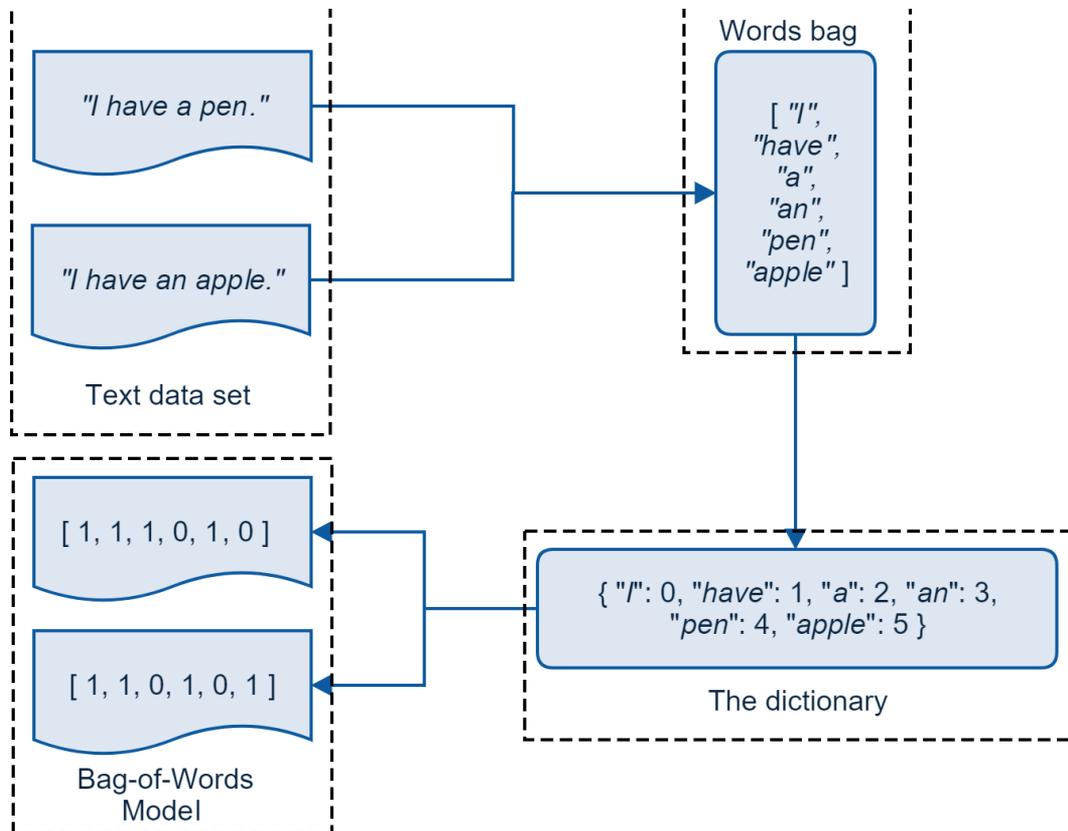


Figure 6 The instance of generating a Bag-of-Words Model

Figure 6 demonstrates a simple instance of BOW model extraction. The BOW model is often treated as the prototype of VSM while VSM as a concept also includes definitions of how BOW is applied in the machine learning approaches and includes other aspects like term weight based representations. However, the BOW model is popular in topic modeling, pattern recognition and some computer vision machine learning approaches due to its simple structure.

3.2.2. Vector Space Model

The Vector Space Model which is usually denoted as VSM was first introduced in the 1960s by Gerard Salton. It is the most popular document representation model in text categorization. The basic idea is to represent document D as an n -dimensional vector in vector space $(T_1, W_1; T_2, W_2; \dots T_n, W_n)$. The T in the vector denoted the dictionary indices of every feature word (term) in the document D . W is the weight of T , and the weights are usually chosen using the Term Frequency (TF) or TF-IDF value of the feature.

3.2.3. Term Frequency-Inverse Document Frequency(TF-IDF)

TF-IDF was introduced by Sparck Jones [Sparck Jones, K., 1972, 2004] and introduced IDF as a term in the term weight calculation instead of simple TF alone. The whole name of the algorithm is Term Frequency-Inverse Document frequency. According to Jin [Jin, X.B., 2006], the weight of a feature word in the document should be in a relationship with the following elements. First, Term frequency, the higher the appearance frequency of a feature word in a document the larger weight the term should gain. Second, Inverse Document frequency, the more documents the feature word is involved in, the more common the feature word will be. Therefore, the IDF value is defined as

$$IDF = \log(N/DF)$$

Where N is the total number of the documents, DF is the number of documents that involve the feature word. Lastly, to take into account the difference between the length of each document, the length of the documents should be normalized to compare their weight. The weight of the feature word can then be calculated with the following formula:

$$W(t, d) = tf(t, d) \times \frac{\log\left(\frac{N}{df(t)}\right)}{\sqrt{\sum_{k=1}^M \left[tf(t, d) \times \log\left(\frac{N}{df(t)}\right)\right]^2}}$$

Inside of it, the $W(t, d)$ represent the weight of the feature word t in the document d , $tf(t, d)$ refers to the term frequency of feature t in document d , M is the total number of document and N is the total number of documents.

The paper of Zhou introduces a method of pre-processing of Chinese text [Qin, 2005]. The progress of pre-processing includes using a proper word splitting system e.g. ICTCLAS for word splitting into sentences, an English words and letters filter, a number filter, a single split Chinese character filter and a stop-words filter.

3.3. Feature Selection

After document representation, the document is represented as high dimensional vectors. The vectors need further dimension reduction to reduce the calculation cost and improve the categorization precision. Thus, feature selection methods need to be applied for further processing in text categorization

3.3.1. Document Frequency (DF)

Document frequency is the number of times that a feature appears in the whole set of documents. Filtering by document frequency calculates for every feature its appearance frequency in the total document set and removes the features with the lower frequency than a pre-defined threshold value. This methodology assumes that low-frequency features have less impact on the categorization result. However, as indicated by Dai, Huang, and Chen, in their survey of Information Retrieval, some terms with lower DF can contain more information than the terms with higher DF. Therefore, the lower DF terms should not be totally filtered and removed [Dai, L., Huang, H., & Chen, Z., 2004]. As a result, this methodology will lose precision in categorization. However, this methodology still performs well in huge data set processing and reduces complexity.

3.3.2. Information Gain (IG)

Information gain is one of the popular approaches employed as a term importance criterion for text document data [Uğuz, H., 2011]. The idea is from the perspective of the information theory [Mitchell, T. M., 1997], and works so that every feature, in turn, is used to divide the learning samples into two groups based on whether the feature appears in the sample or not and select the features where their corresponding grouping provides the most information in the category. The IG of the term can be calculated with the formula below:

$$\begin{aligned} IG(w) = & \sum_{i=1}^k P(C_i) + P(w) \sum_{i=1}^k P(C_i|w) \log P(C_i|w) \\ & + P(\bar{w}) \sum_{i=1}^k P(C_i|\bar{w}) \log P(C_i|\bar{w}) \end{aligned}$$

where $P(C_i)$ represents the probability of a document belonging to the category C_i , $P(w)$ is the probability that feature w appears in a document, $P(\bar{w})$ is the probability that feature w doesn't appear in a document, $P(C_i | w)$ is the probability that the feature w appears in the documents belonging to category C_i , $P(C_i | \bar{w})$ is the probability that feature w appears in the documents not belonging to the category C_i , and k is the total numbers of the categories. The larger the $IG(w)$ is the larger the possibility is that the feature will be selected.

Information Gain is one of the best performing feature selection algorithms in terms of precision and efficiency. In the paper of Zheng [2007], the precision experiment resulted in 1000 features with a test accuracy of 96.974%, just a bit lower than the CHI methodology but with five seconds running time advantage. In the paper of Yang [1997], it was shown in an experiment that IG is able to remove 98% of the unique terms which indicates high effectiveness in dimensionality reduction. However, the most serious weakness of IG is that it can only check the contribution of features to categories of the whole dataset, but not to specific categories. Therefore, IG is suitable for overall common feature selection for the document set, but meanwhile can be weak for selecting features useful for specific categories.

3.3.3. Pointwise Mutual Information(PMI)

Pointwise Mutual Information is a criterion commonly used in statistical language modelling of word association and related applications [Church, 1990]. In the application of text categorization, PMI can measure co-occurrence between features and the categories, the larger PMI feature has to one category the more information the feature has to affect identification to this category. Therefore, PMI represents the correlation between features and categories. In text categorization, a feature can have only one IG and DF. Nevertheless, a feature can have PMI to every category that is included in the training document set. Generally, the features are selected within the top PMI ranking of the features to the category or a threshold will be set to filter the features with low PMI to the category. The PMI can be calculated with the following formula:

$$PMI(t, c) = \log \frac{A \times N}{(A + C) \times (A + B)}$$

Where A represents the frequency of documents which include term t and belongs to category c, B is the frequency of documents which include term t but do not belong to category c, C represents the frequency of documents which do not contain term t but belong to category c. If there is no relationship between term t and the category c the PMI(t, c) will be 0. To measure the relationship between t and the total category set the following formula can be applied [Fan, X.L. 2010]:

$$PMI(T, C) = \sum_{i=1}^m [PMI(t, c_i) + PMI(\neg t, c_i)]$$

The m represents the total number of the categories. Due to the low complexity in the calculation and lack of consideration for term frequency in calculation PMI is usually lower in precision but higher in efficiency. However, with increasing computing power the advantage of PMI is becoming less obvious.

3.3.4. χ^2 Statistic (CHI)

The CHI is another well performing and widely applied feature selection methodology. CHI measures a significance test statistic the correlation between term t and category c while assuming that the correlation follows the chi-squared distribution with one degree of freedom. The correlation between term t and the category c is in proportion to the value of CHI, the higher the value is, the less the term t seems independent of category c , the probability of selecting term t into the feature subset will increase. The CHI value of the feature can be calculated with the following formula:

$$\chi^2(t, c) = \frac{(AD - BC)^2 \times N}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Where N is the total number of documents within the training document set, c is a specific category, t represents a specific term, A represents the frequency of the documents that contain term t and belongs to the category c , B represents the frequency of the documents that contain term t and do not belong to the category c , C indicates the frequency of documents which do not include term t but belong to category c . D is the frequency of documents that do not contain c and do not belong to category c either.

CHI based on the assumption that the high term frequency features in the documents in a specific category and the high term frequency features in the documents in other categories are all helpful to decide whether a document belongs to the specific category. The CHI value reflects the contribution of a term to a specific category and the other categories at the same time in two perspectives. This idea highly increases the accuracy in feature selection and the precision in categorization result. However, according to Li, lack of consideration of term frequency results in its relying on low term frequency features.

3.3.5. Summary of The Feature Selection Approaches

Yang's experiment [Yang, 1997] measuring precision of the four above described algorithms indicates that with the proper setting of the thresholds, apart from the PMI,

the rest of the algorithms achieve almost the same results in classification accuracy, the simplest algorithm DF's testing performance was surprising, she concluded this result as some correlation within DF, IG and CHI. However, Dai [2004] point out in their paper that this kind of result may highly depend on the balance of DF, IG and CHI values for English text and on suitable scale selection with the training document data set.

In Chinese text categorization, the number of features is far larger than that in English which results in more low DF values for features that seriously affects the precision of IG and CHI. This situation is also mentioned in the paper of Shen [2006]. In his paper, Dai [2004] come up with a practical improvement method to combine the DF values with the IG and CHI to eliminate their dependence on low DF terms.

Additionally, in the research field of short text categorization, in the paper of Wang [2009] authors take into account the low TF values of the features, the Association Rule [Agrawal, 1993] is introduced to create a co-occurrence feature set to extend the features and eliminate the effect of synonyms.

3.4. Classifiers

The purpose of the classifiers is to find a rule function for processing the new data, based on the labels of the given training data. The algorithms of classifiers are various. They can be statistics based, logical based etc. The classifiers usually contain more than 1 parameter to be selected. To properly choose the classifiers and their parameters is directly related to maximizing the performance of the classifiers.

3.4.1. Naïve Bayes(NB)

Naïve Bayes [Lewis, D. D. 1998] is one of the most popular and easiest implemented algorithms in text categorization. The idea is based on Bayes' theorem, to decide a document category by calculating the posterior probabilities of the document belonging to every category and select the category with the largest posterior probability as the result. We assume document d can be represented with the feature vector $d = (w_1, w_2, \dots, w_n)$. Based on Bayes' theorem the posterior probability that d belongs to category c can be calculated as:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

In Naïve Bayes the probabilities $P(c)$ and $P(d)$ need not be calculated as the former probability $P(c)$ is assumed uniform and thus constant with respect to c , and the latter $P(d)$ is always constant with respect to the category c . Therefore, the probability $P(c|d)$ will be decided by $P(d|c)$. To calculate $P(d|c)$, the NB makes an assumption that all the feature dimensions in document d are independent. Finally, the probability will be:

$$P(d|c) = \prod_{i=1}^n P(w_i|c)$$

NB is easy to implement with low complexity and performs well with irrelevant feature and noise. However, the assumption does not consider relationships between feature which may affect its precision.

3.4.2. K-Nearest Neighbor(KNN)

The K-Nearest Neighbour is a simple categorization method based on VSM. The idea of the method is simple, given a test document, the system will find K nearest neighbours in the training set, and put the document into the category that most of the nearest neighbours belong to. The similarity between the neighbours and the test document will be calculated by the system as the weight to decide which neighbours are the nearest ones.

The KNN needs no modelling and training and performs more efficiently for new patterns. The K should be pre-defined through experiment experience, normally between 20 – 50.

3.4.3. Decision Tree(DT)

The Decision Tree is one of the most well-known logical supervised learning algorithms. The algorithm was introduced by Hunt [1966] in 1966 for concept learning. The categorization progress is the same as its name: decisions are made from the root node following a path down to a leaf node. The root node represents the total set of input data and each leaf node represents a subset classified to a particular value of the target variable. Through training a model, a tree of decision rules will be built to describe the internal relationship between the input data set to the target variable. Every branch node of the tree model is a single attribute test to the input data set. The data set is split based on the test result where each split subset can then be subject to further testing, finally, the surviving data set at a leaf node is predicted to output a particular value of the target variable.

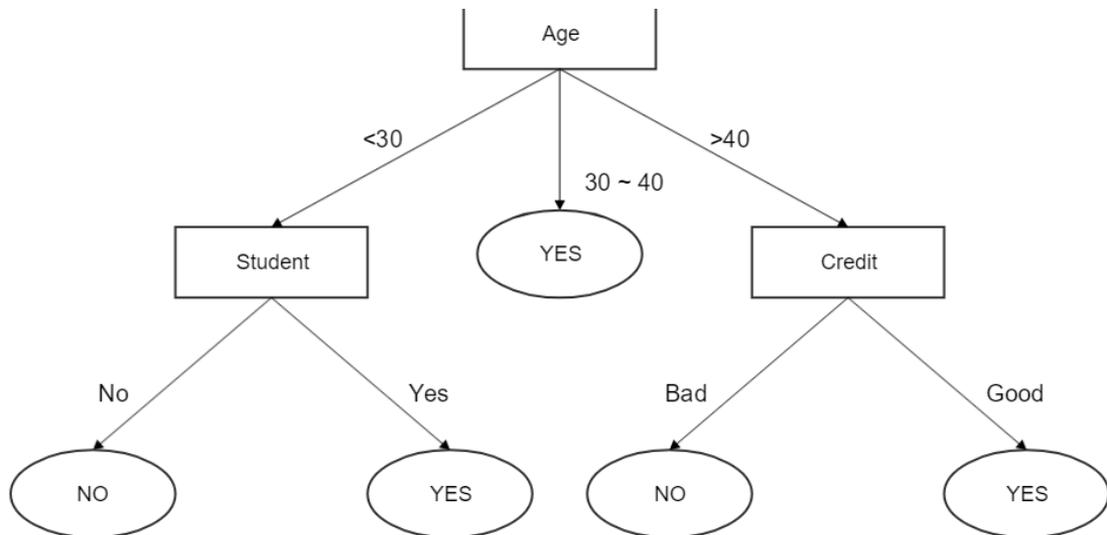


Figure 7 An instance of a decision tree [Wang, 2006]

Figure 7 is an instance that presented by Wang in his paper to explain the decision progress of the decision tree. This tree is used to predict whether a person would buy a computer. The input data set is a group of different people and the target variable takes two values, either “YES” or “NO”. Three attributes of the person will be tested which are age, student status, and credit. According to the test results, five decision rules are set to decide the group of people:

1. A person who is younger than 30 and is a student will buy a computer.
2. A person who is younger than 30 but not a student will not buy a computer.
3. A person who is between 30 and 40 years old will buy a computer
4. A person who is older than 40 and has good credit will buy a computer.
5. A person who is older than 40 but has bad credit will not buy a computer.

From this instance, it can be inferred that the key problem of modelling the decision tree is to choose the testing attributes and decide the number of the branches. The most popular two algorithms to decide the testing attributes are ID3 and C4.5.

1. ID3 Algorithm

ID3 was introduced by Quinlan [1986] as the extension of one of the branches of Concept Learning System [Hunt, 1966]. ID3 gives up the pre-planned attributes approach and constructs the decision tree based on the information maximization. ID3 builds the decision tree from the root node to the leaves and uses the information entropy as the standard for attribute selection which is similar to the IG algorithm in feature selection. This standard chooses the attributes which take more values from

the data set and makes the ID3 algorithm categorize the sample based on features appears more in a specific class.

2. C4.5 Algorithm

C4.5 is an improvement of ID3, it imports the gain ratio to overcome the weakness of ID3 which prefers to choose high-frequency attributes and can process the continuous attributes. The gain ratio can be calculated with the formula below:

$$\text{Gainratio}(a) = \frac{IG(a)}{\text{splitinf}(a)}$$

In the formula, a stand for the candidate attribute, $\text{splitinf}(a)$ is the split information which means the generated potential information when the data set T is split into h parts, the $\text{splitinf}(a)$ can be acquired through the formula below:

$$\text{splitinf}(a) = - \sum_{i=1}^h \frac{|T_i|}{|T|} \times \log_2 \left(\frac{|T_i|}{|T|} \right)$$

Where $|T_i|$ is the number of samples in the sub-dataset I , $|T|$ is the number of samples in the whole data set, h is the number of sub-datasets. Although accompany with the weakness of repeatedly scanning the data set and not be able to do the incremental learning, C4.5 is still one of the most popular algorithms for the DT classifier implementation in many machine learning libraries.

3.4.4. Support Vector Machine(SVM)

The Support Vector Machine (SVM) is proved by many experiments to be an efficient and effective method for text categorization. SVM tries to separate the points that belong to the category from the points that not belong to the category by finding a hyperplane in the feature vector space. The best hyperplane which separates the points belong to the category from the points that not belong to the category with the largest distance is selected during the process of the model training. This distance is measured from the hyperplane to the nearest point from category related points and category unrelated points. Below is an example of best hyperplane choosing in two-dimension presentations

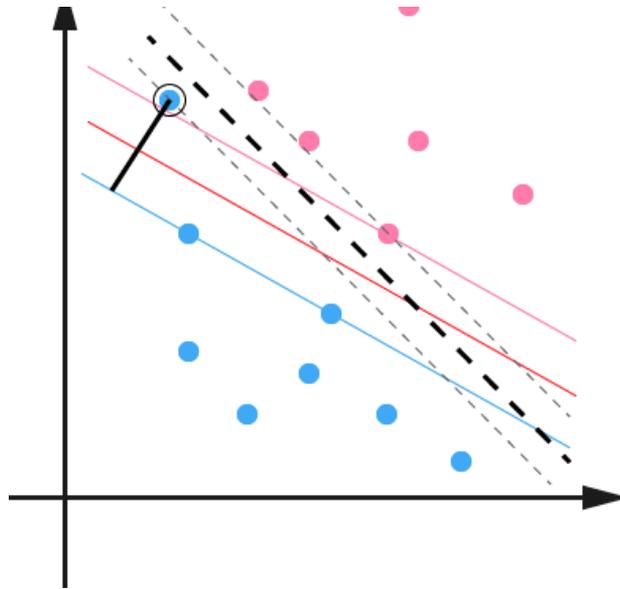


Figure 8 The hyperplane extraction of SVM

As shown in the Figure 8. Only the relative subset of the training set is used to find the best hyperplane. The points in the subset are called support vectors. The rest data all have no effect on determining the best hyperplane. This property of SVM is the significant difference between itself and other classifiers.

Generally, the SVM classifier has two advantages. First, the process of SVM model training remains the same in multidimensionality feature spaces which theoretically resolve the overfitting problem. Second, the parameter selection work is automatically done within the model training process which aims to find the best hyperplane.

Considering the unknown nature of features to be selected and their sparsity in well-structured short texts the specific classifier selection is still a topic for further study.

4. Syntax pattern application in short text categorization

The traditional text categorization flow works effectively and performs well with long text sample data collections, for instance, articles and blogs. In most of the cases, with the proper choice of every step method and optimizations in the variables of the core algorithms, the results are acceptable and being put to practical use. However, in most of the short text sample data collections, the traditional text categorization flow based systems usually do not work well.

Short text usually refers to a text sample that was less than 200 words [Cui, 2011]. In the paper of Li [2013] the typical short text contains the 3 features below:

1. Extremeness.

The structure of the short text sample is quite simple. Due to the length of every sample, most of the short text sample is formed with few sentences. Some of them are even formed with few words. The simple structure of the short text results in two extreme situations of the features in the sample.

- 1) Highly discrete. The features are nearly nonrelated to each other or have few inner sample connections which make the feature vector extremely high dimensional. It is hard to select proper features from it for text categorization.
- 2) Overlap. Few words are used in one big topic which makes it hard to find effective information in the sentence to make the further levels of text categorization. The feature vectors are in low dimensionality which is hard to be divided from each other.

2. Real-time

Most of the short text samples appearing on the internet are updated in real-time. From the traditional short reviews and comments to the modern Twitter twits, Facebook statuses, the updating speed of the short text samples is counted with seconds or even microseconds. This feature makes the size of the short text data set dramatically large and hard to acquire which brings more challenges and requirements for the text categorization.

3. Irregularity

The text formats and the words in use are freer than long text. Many of the short text expressions are incomplete and many catchwords and abbreviations often appear in the short text samples. These features will make some noise in feature selection.

According to the features of the short text, the traditional text categorization method is definitely not fit for the short text categorization. More processing towards the features should be added to the text categorization flow.

4.1. Approaches of short text categorization

Presented by many papers, the short text lacks enough features to demonstrate its category. Therefore, many approaches in short text categorization aim to form the words partly together in the text sample to construct new features which include the categorization information of the text sample. N-gram, feature co-occurrence set, syntax patterns are the approaches to fulfil these requirements from different aspects.

4.1.1. N-gram

N-gram is a type of model which is widely applied in the field of the Natural Language Processing. Originally, N-gram is used to separate the document into elements. This model defines a very basic operation to a document e.g. use a window whose length is N elements to scan through the document and separate the whole document into a new set of windowed elements [Tripathy, 2016]. Usually, the 1-gram (uni-gram), 2-gram (bi-gram), 3-gram (tri-gram) are most frequently applied. For instance, I have a document sample below:

text categorization

If the letter is used as the elements and bi-gram is applied with the document splitting, the document sample will be split into following windows flow of letter set:

te, ex, xt, t_, _c, ca, at, te, eg, go, or, ri, iz, za, at, ti, io, on

N-gram model is based on the assumption of Markov Chains that the appearance of the Nth element in the document is only related to the former N-1 elements while has no effect with the rest elements [Brown, 1992]. This assumption has a direct effect on how the mathematic model of the N-gram model is built.

For instance, in the application or words splitting, the whole word splitting strategies can be treated as the document set. In the bi-gram model the appearance rate of every flow set element is $P(w_i | w_{i-1})$, according to the Markov Chain assumption the document appearance rate can be calculated with the following formula:

$$P(w_1, w_2 \dots w_m) = \prod_{i=1}^m P(w_i | w_{i-1})$$

Therefore, the document with the largest document rate is the best word splitting strategy.

The elements of the N-gram model can be not only letters or characters but also words, POS tags or any elements which appear as a chain in the document [Sidorov, 2014]. Following this property, the feature set of the short text document is able to be re-organized using the N-gram model. For instance, with the sample document, "I have a pen and an apple.", the original feature set is:

I, have, a, pen, and, an, apple

The new feature set applied the bi-gram is:

I have, have a, a pen, pen and, and an, an apple

In the new feature set of the instance document the benefits of applying the N-gram model to form the new feature set is obvious, the semantic features, for example, *I have* as subject-predicate structure, *an apple* as quantity phrases appear in the new feature set which would be recognized as a strong feature in the document representation. However, some totally none-sense noisy features for example *pen and*, *and an* appear at the same time in the new feature set which is harder to detect and filter comparing with the single word feature set.

What's more, facing the long sentence or the sentence with some complicated structure for instance inversion or rear object it would also be hard for this method to obtain some core expression semantic elements.

4.1.2. Feature co-occurrence set

Feature co-occurrence set based on the feature co-occurrence model in the field of Natural Language Processing [Rak, 2005]. This model assumes that if in the large scale of the corpus, two words often occur in the same document, the two words are considered as associated with each other. Moreover, the higher the occurrence rate of the two words is, the closer the relationship between the two words is.

The paper of Wang [Wang, 2009] presents that feature co-occurrence set aims to figure out the occurrence relationships of the original feature set, filter out some feature pairs with high occurrence rate and combine these feature pairs into new features. The feature co-occurrence set is formed with the new features and the rest former word features. For instance, in the corpus of the sports field, the short phrases like “play basketball”, “play football” are the typical co-occurrence features.

In her paper, Wang also introduces that the high occurrence rate feature pairs can be obtained with association analysis. From her perspective, the documents set e.g. corpus may be treated as a database with association rules, while the feature set is treated as an item set. The occurrence rate feature pairs are considered as the strong association rules in the association analysis. Therefore, by setting up proper threshold minimum values of the support and confidence, the co-occurrence features can be obtained with the FP-growth algorithm.

The advantages of applying feature co-occurrence sets to form the new feature set are obvious. Compared to N-gram feature co-occurrence set contains more short phrases. According to Liu [Liu, 2007], the short phrases especially the key phrases are more semantically complete and make it easier to express the core meaning of the document. However, feature co-occurrence requires rapid statistical analysis to find the proper minimum threshold values of the support and confidence. The key phases of some specific categorization may also be filtered due to the affection of high occurrence phrases. To avoid this situation, the large-scale corpus especially the domain term corpus need to be imported into the analysis which decreases the efficiency of building the new feature set.

4.1.3. Syntax Pattern

The concept syntax pattern, based on feature reform methodology, is first found in the paper of Yan [Yan, 2007]. As presented by them the popular statistics based feature selection methods perform well with their own characteristics English feature selection, but face two big problems in Chinese feature selection: a huge amount of time cost in calculation and high dimensionality in feature representation as a feature vector. Bunch of irrelevant words and low TF words affect the precision and efficiency of feature selection.

Combine with the situation now in short text categorization, the short document length makes relevant features low in TF within the document as well. As presented by Lin [2014], the researchers attempt to use association rules to enhance or extend the feature.

By the opinion of Lin, this method performs well in high DF terms but weak in low DF terms.

Based on the expression of Chinese, Yan [2007] come up with brief steps of feature selection and introduced the regular expressions to recognize the single parts of speech. In the paper of Zheng [2010], the parts of speech are combined to select the emotional feature of the expression. Finally, in the paper of Lin, the specific progress and the basic subexpression pattern are given to formalize this method. The common feature selection will follow the steps below:

1. Split the expression based on the punctuations into some short subexpressions.
2. Use ICTCLAS to split words and mark the parts of speech.
3. Make the pattern combination rule with manufacturer analysis depending on the subexpression structure, use regular expressions to select features as candidate features.
4. Noisy term removal, feature set generating.

The basic expression syntax pattern and java regular expressions are defined in the table below:

ID	Syntax pattern	Regular expression
1	noun; gerund; noun phrase + noun phrase	(.{1,4}/v?n){1,2}
2	verb + noun	.{1,4}/v.? .{1,4}/n
3	noun phrase + auxiliary word + noun phrase	(((.{1,4}/v?n){1,2}) (.{1,4}/v.? .{1,4}/n)) 的 /ude1 (((.{1,4}/v?n){1,2}) (.{1,4}/v.? .{1,4}/n))
4	noun phrase + conjunction + noun phrase	(((.{1,4}/v?n){1,2}) (.{1,4}/v.? .{1,4}/n)) . {1,2}/cc (((.{1,4}/v?n){1,2}) (.{1,4}/v.? .{1,4}/n))

Figure 9 The java regular expressions of basic expression syntax patterns [Lin, 2014]

Compare with the N-gram and feature co-occurrence set approaches to reform the feature set. The syntax pattern approach emphasizes more the artificial analysis in defining the rules to regroup the features. With the additional artificial analysis work, the new features are more semantic and meaningful. In addition, the syntax pattern requires only the

simple match processing which deduces the additional processing of the program. Nevertheless, syntax pattern performs weakly in random short text data sets since a random short text data set usually contains vast types of expression, which is hard to find the general syntax pattern to process all the data. On the contrary, it performs well in processing the documents data with the similar sentence construction. Therefore, in the DiDi case, the syntax pattern is more suitable to be chosen as the feature reconstructing method. However, the introduced regular expression approach to obtain the syntax pattern has a strict limitation in the word format and the grammar expressions of the text. These properties may cause the low accuracy in matching syntax pattern with a high degree of freedom of language, for instance, Chinese.

4.2. Dependency Parser

Dependency grammar is first introduced by the French linguistic expert Tesniere [1953]. The dependency grammar argues that there exists a centre word in the sentence which depends on no other parts in the sentence. All the rest words of the sentence fulfil a binary relation. This relation contains two roles a head and a dependent with each role is mapped to one word in the sentence. This relation is called dependency relation. The dependency relation offers new association rules to word pairs which guarantee the direction and affliction of the words pair.

Robinson [1970] elected the four axioms of the dependency relations, which became the basis of dependency relations. According to the special property of Chinese language, Bai [2004] extend the four axioms to five. Below the “member” will be used instead of “word” to describe the entire five axioms:

1. Every sentence consists of only one independent member.
2. All the other members of the sentence are directly depending on some members.
3. Any member of the sentence should not depend on two or two more other members.
4. If member A depends on member B, while member C locates between A and B, accordingly member C depends on A, B or some other member between A and B.
5. The other members at the two sides of the centre member do not have a relationship with each other (Chinese specific).

The dependency relation can be syntax relation or linguistic relation. Using the syntax relation as dependency relation to analyse a sentence is called Dependency Parser. The progress of the dependency parser can be generalized into two steps [Tang, 2014]:

1. Build the dependency tree according to the syntax dependency relations within the sentence.
2. Find the dependency direction of every dependency member couple and mark its syntax dependency relation.

Different languages usually share the similar syntax characters but diverse in syntax relations. In the case of DiDi, the syntax dependency relations are defined by the Language Technology Platform which is one of the best open source Chinese Natural Language Processing platform. The following table lists all the definitions of the syntax dependency relations and their marking tags. The quick example of the relations and their direct English translations are also given in the table.

Tag	Description	Chinese Example	English Translation
SBV	subject-verb	我送她一束花 (我 ← 送)	I send her a bunch of flowers (I ← send)
VOB	verb-object	我送她一束花 (送 → 花)	I send her a bunch of flowers (send → flower)
IOB	indirect-object	我送她一束花 (送 → 她)	I send her a bunch of flowers (I → send)
FOB	fronting-object	他有很多书来读 (书 ← 读)	He has a lot of books to read (book ← read)
DBL	double	他请我吃饭 (请 ← 我)	He invites me for dinner (invites ← me)
ATT	attribute	红苹果 (红 ← 苹果)	Red apple (red ← apple)
ADV	adverbial	非常美丽 (非常 ← 美丽)	Very beautiful (very ← beautiful)
CMP	complement	做完了作业 (做 → 完)	Finish the homework (finish)
COO	coordinate	大山和大海 (大山 → 大海)	Mountain and sea (mountain → sea)
POB	preposition-object	在房间内 (在 → 内)	Is inside of the room (is → inside)
LAD	left adjunct	大山和大海 (和 ← 大海)	Mountain and sea (and ← sea)
RAD	right adjunct	士兵们 (士兵 → 们)	Soldiers (soldier → s)
IS	independent structure	两个单句在结构上彼此独立	Two sub-sentences independent from each other in structure.
WP	punctuation	!	!
HED	head	整个句子的核心	The centre word of the sentence.

Figure 10 The dependency relation marks from LTP [2018]

The progress of building the dependency tree starts with confirming the dependency relations of every member of the sentence. This step is usually dependent on the existing corpus. For instance, with the sample sentence from the DiDi database:

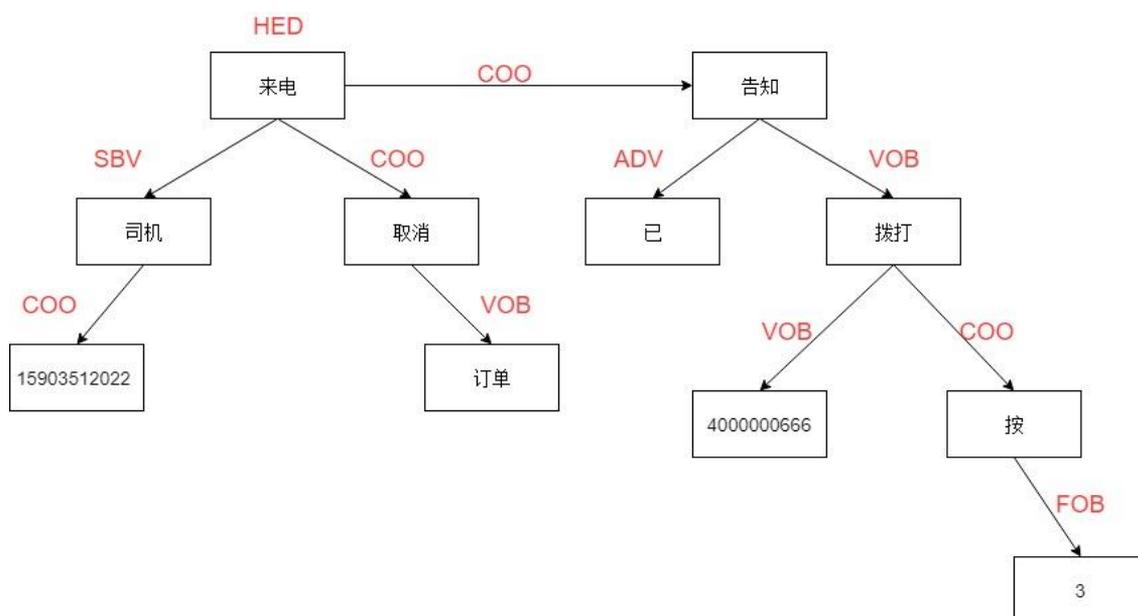
司机 15903512022 来电取消订单，已告知拨打 4000000666 按 3。

Driver 15903512022 called to cancel the order, have told to call 4000000666 and press 3.

By analysing from left to right through the sentence, it can be derived from the following dependency relations:

司机 ← 来电， 15903512022 ← 来电， 来电 → 取消
 取消 → 订单， 来电 → 告知， 已 ← 告知， 告知 → 拨打
 拨打 → 按， 拨打 → 4000000666， 按 → 3

Next, choose the member which does not depend on other members as head and build every branch according to the head to form the dependency tree. After proper marking of the dependency relations the entire dependency tree is:



The dependency tree contains all potential basic patterns defined in the syntax pattern feature reforming method and extends the patterns with more relations. Therefore, compared to the use of sub-expressions to obtain the syntax patterns, using the dependency parser to detect out the syntax pattern members for the documents is more accurate and specific. With the marking rules applied in the dependency parser, it would be easier to the obtain the customized syntax pattern.

5. A solution for DiDi customer service records classification

According to the description in 2.3.2, most of the records in the DiDi data set are presented in a similar format and a general sub-expression can be extracted from them to describe the potential categorization information of the records. Therefore, the syntax pattern method will be used to reform the new feature set. A proper tool will be used to fulfil the dependency parser to extract customized syntax pattern. The solution will base on the traditional text classification flow. To apply the syntax pattern to reform the simple words feature set into the new words pair feature set, the pre-processing step of the traditional text classification will be separated into two new steps. The first step contains the word segmentation, dependency parser, obtaining the pre-defined syntax pattern to reform feature set and stop words filter. This step is defined as Feature Enhancement. The next step will focus on the document representation. The general flow of the solution is shown as below:

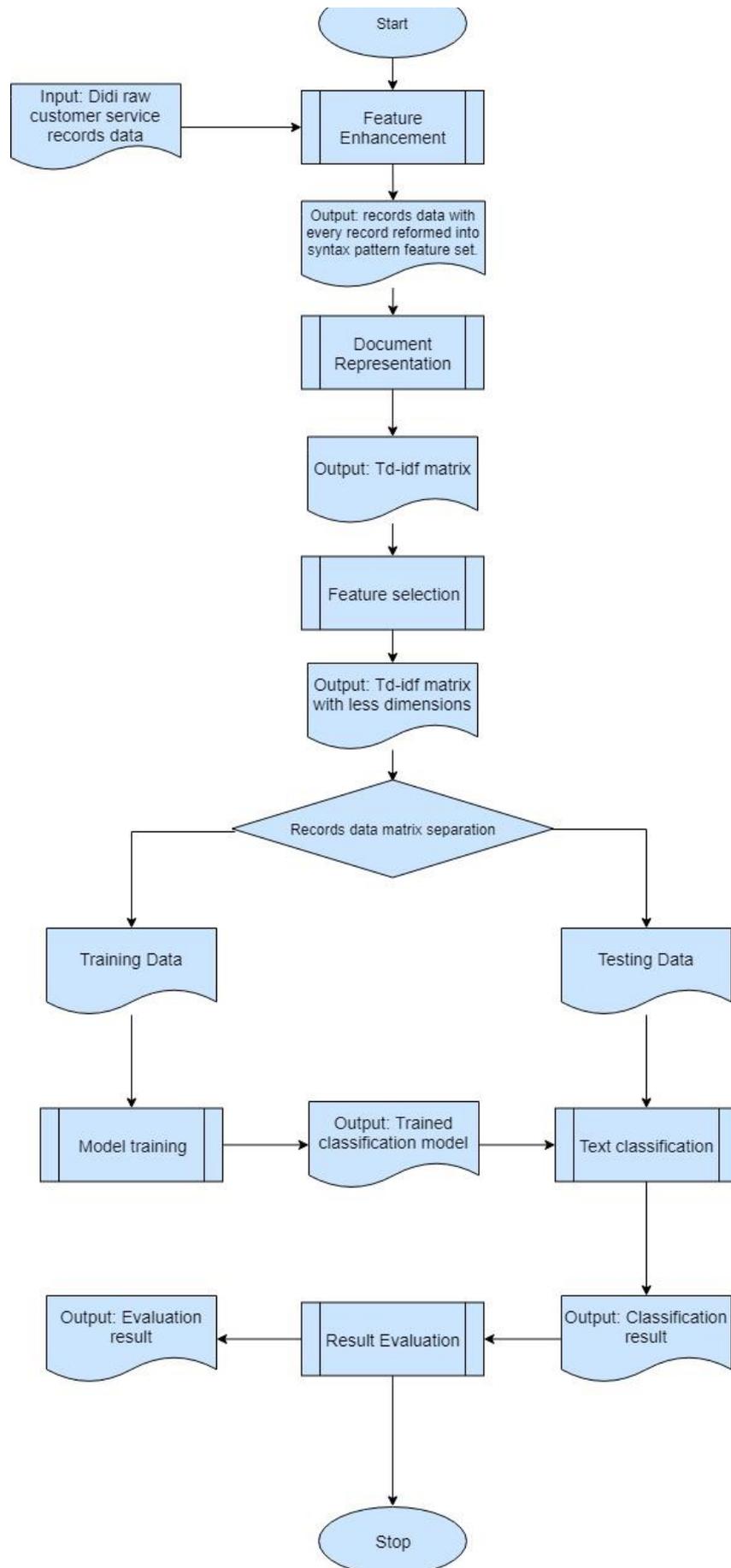


Figure 11 Design flow of the DiDi customer service records categorization solution

5.1. Implementation

The solution will be implemented with the Python programming language. Python is an open source programming language which is object-oriented and interpretive. It is first developed by Guido van Rossum in 1989 and released in 1991. Most important is, python has a very strong and rich standard library. For the traditional developers, it's easy to handle and implement the python encapsulation for the library developed in other languages. For the researchers or the beginners, the learning cost of python will not be more than a professional analysis tool while providing more possibilities to express their ideas. In recent years, with the rising popularity of machine learning and big data processing, python shows its remarkable expansibility with rich well-performed third-party libraries in both science calculation field and machine-learning field which attract increasing numbers of data scientists to choose python in their research.

5.1.1. Feature Enhancement

As mentioned at the beginning of this chapter the feature enhancement is a newly defined step which integrates the syntax pattern feature reformation into the pre-processing step. Specifically, the step contains the following executions: word segmentation, dependency parser, syntax pattern acquiring and new feature set reforming, stop-words removal. The entire processing flow of this step is presented below:

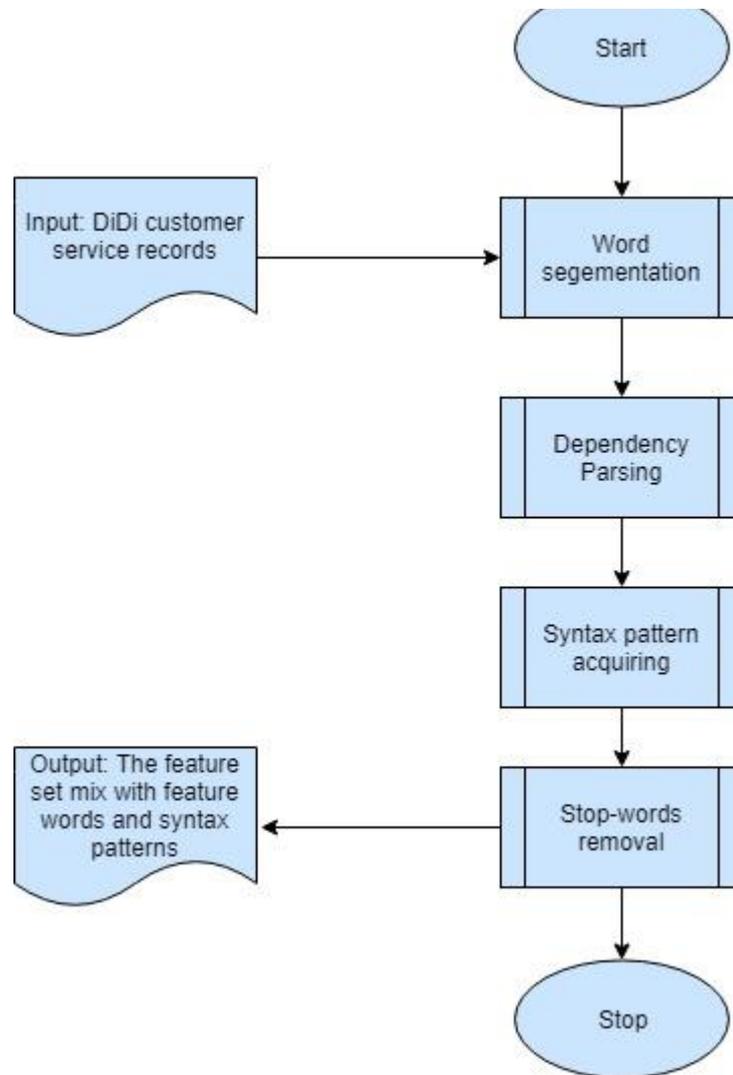


Figure 12 The flow of Feature enhancement

The word segmentation and dependency parser are implemented through the pyltp library. Pyltp library is the python encapsulation of the Language Technology Platform (LTP). As mentioned in 4.2, LTP is an open source Chinese Natural Language Processing system, developed by Harbin Institute of Technology. LTP [2018] defines the processing result expression using the XML language. It offers an entire set of the rich, highly-efficient, accurate Chinese natural language processing modules including word segmentation, Part-Of-Speech(POS) tagging, an identity of named entity, dependency parser and semantic role labelling. The LTP offers the models to support the interface to finish the word segmentation and the dependency parser. The return result of dependency parser contains the features in which every feature is marked with number index start with 1 from left to right, the dependency relations and the root index of every relation as three lists. Therefore, if combines the three return lists properly, the dependency tree can be easily derived from the return result. In this case, due to the indexes of the three return elements lists are the same, the output lists can be directly processed to obtain the dependency relation pairs. For instance, with the example sentence below:

司机 13318757313 来电表示提现未到帐，已告知会有延迟到账的现象

Driver 13318757313 called to complain that the withdraw deposit hasn't arrived, have told that the withdraw deposit will sometimes delay.

The dependency tree for the example is:

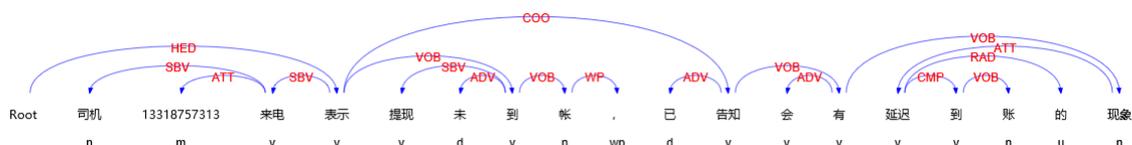


Figure 13 The dependency tree of the example sentence

The output dependency tree from the solution system will be:

['司机', '13318757313', '来电', '表示', '提现', '未', '到', '帐', '.', '已', '告知', '会', '有', '延迟', '到', '账', '的', '现象']

['SBV', 'ADV', 'HED', 'ATT', 'SBV', 'VOB', 'WP', 'COO', 'ATT', 'VOB', 'RAD', 'VOB', 'WP', 'FOB', 'ADV', 'COO', 'WP']

[3, 3, 0, 5, 6, 3, 3, 3, 12, 9, 9, 8, 8, 16, 16, 8, 3]

The syntax pattern analysis method is implemented to obtain the Enhancement feature pairs from the output result above. The method works with the steps below:

1. The dependency relation marks, mentioned in Figure 10, are used to match the relations and get the feature indexes of the required relations that the features belong to.
2. The indexes are used to get the root index of the required dependency relations from the third list.
3. The root indexes and the feature index are used together to obtain the head and tail features of the dependency relation feature pairs.

From the DiDi data description mentioned in the 2.3.2, it can be inferred that most of the document use the verb-object structure to express its core content. Therefore, in the DiDi case, the verb-object relation will be used to form the enhancement feature set.

Choosing the verb-object as the only syntax pattern here is due to the character of the short text, every document containing a small number of features. Therefore, using more dependency relations in the feature set reformation will lead to the high separation degree of the document vector which will decrease the accuracy of the classification. The syntax pattern selection method is implemented as below:

```

8  def syntax_side(a,b,c):
9  for i in range(len(a)):
10     if b[i]=='VOB' and b[c[i]-1] == 'VOB':
11         a[c[c[i]-1]-1] = a[c[c[i]-1]-1] + a[c[i]-1] + a[i]
12         a[c[i] - 1] = '/c'
13         a[i] = '/c'
14     elif b[i]=='VOB':
15         a[c[i]-1] = a[c[i]-1] + a[i]
16         a[i] = '/c'
17     elif b[i] == 'WP':
18         a[i] = ''
19     ystr = ' '.join(a)
20     ystr = re.sub(r'/c', '', ystr)
21     ystr = re.sub(r'\s+', ' ', ystr)
22     return ystr

```

Figure 13 The implementation of syntax selection method

The method takes the output result of the LTP dependency parser lists and outputs the enhancement feature set. The punctuations are removed at the same time. The example enhancement feature set is like below:

司机 13318757313 来电 表示到帐 提现 未 已 告知有现象 会 延迟 到账 的

The enhancement feature set data collection is the new data collection that will be used for text categorization.

5.1.2. Document Representation

After the reformation of the enhancement feature set, the new data collection invokes the `filterStopWords()` function to remove the stop-words as mentioned in the chapter 3.1 from the data collection. The stop-words list using, in this case, is supplied by Natural Language Processing & Information Retrieval Sharing Platform (NLPIR). NLPIR [2018] is one of the earliest platforms in China which offers both academic and business Chinese language processing service. Its research achievements affect many Chinese language processing service suppliers in China.

Before invoking the function, the data is read from the xlsx file by using the Pandas package. Pandas [2018] is an open source python data analysis tool which aims to enhance and simplify the python data analysis. Pandas offer three entirely new types of data structure which are Series, DataFrame and Panel. The three data structures are often mapped to accept the single-row or column data, table data and multi-table data from the database. As the data resource from DiDi is exported to xlsx, by using Pandas, the table in the data collection can be directly transferred into Pandas DataFrame (DF) which takes the original column names as column index and the first column as row index.

	class	content \
0	1	\n司机来电无法联系乘客，已告知到达接驾低点后耐心等待10分钟，十分钟还是...
1	1	司机来电查询订单 已告知 订单已支付。
2	1	司机来电查询订单信息 已告知迷你型订单只能接一个人
3	1	杨师傅 349063993 咨询完成指派订单数 已告知完成指派订单数
4	1	\n司机来电咨询乘客取消订单后，司机端没显示怎么办，\n已告知，正常点击下一步
5	1	司机来电查询订单信息 一高子涵客服关闭
6	1	司机来电咨询订单 已告知订单未支付，
7	1	阎师傅 15588663686 咨询订单时间 已告知订单时间
8	1	张师傅 17710366690 咨询订单是否取消 已告知订单已取消
9	1	蔡师傅 18826221665 咨询订单是否取消 已告知订单已取消
10	1	陈师傅 13925222124 咨询订单是否取消 已告知订单没有取消
11	1	乘客来电在服务中上车之后产生的费用较高 已告知现在看不到价格信息 下车之后再次致电客服...
12	1	关师傅 13911561165 咨询订单是否取消 已告知订单已取消
13	1	胡师傅 13819771186 咨询订单金额 已告知订单金额
14	1	李师傅 13808795596 咨询订单是否取消 已告知订单没有取消
15	1	李师傅 15965708970 咨询订单金额 已告知订单金额

Figure 14 The Sample of Pandas DataFrame

Figure 14 demonstrates that the Pandas DF is quite similar to the table in the original excel file. By invoking the DF object by **DF [column index] [row index]** all the members in the table can be easily located. By invoking **DF [column index]**, the entire column can be acquired. In this case, the output data collection after stop-words filtering will form a new column and is written back to the excel file for the further processing.

After removing stop words, the data collection is ready to be transferred to feature vectors by applying the VSM and using the TF-IDF to calculate the weight. In this case, the scikit-learn package will be applied. Scikit-learn is a very strong open source scientific calculation toolkit which based on SciPy. The first version was released by David Cournapeau in 2008 [Scikit-learn, 2018], and was originally one extending version of Scikits. Due to its keeping on extending in only the field of machining learning, this Scikit was named as Scikit-learn. Scikit-learn which is also known as sklearn contains six aspects of functions modules including classification, regression, clustering, dimensionality reduction, model selection and data pre-processing. The six modules

contain almost all the popular machine learning algorithms and their different implementations.

The sklearn offers the **CountVectorizer()** class to finish the document representation. With English or other data, the 'stop_words' can be set into the language and remove the stop-words while initializing the instance. After the processing work with the raw data, the **CountVectorizer()** could directly invoke the **fit_transform** method to build the basic bag-of-word feature matrix of the data collection. For calculating the TF-IDF value as the weight to the feature, the **fit_transform** method in **TfidfTransformer()** class can be used to transfer the basic feature matrix into the TF-IDF feature matrix. The TF-IDF matrix should then be transferred to Pandas DataFrame for further processing. The classes of the training data are required to be encoded to numbers and transferred to Numpy array for further processing. Below is the sample of the TF-IDF matrix:

	0	1	2	3	4	5	6	7	8	9	\
0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.000000	0.390386	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
5	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
7	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
11	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
12	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
13	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
15	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
16	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
17	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
18	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
19	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
20	0.0	0.275564	0.291225	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
21	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 15 Sample of TF-IDF matrix

5.1.3. Feature Selection and Model Training

The transferred TF-IDF matrix is theoretically the ready input data for the feature selection. However, if the feature's TF-IDF value is the larger than 1, it may affect the filtering of the lower TF-IDF value features. Therefore, the TF-IDF value features which

larger than 1 are all set into the value 1 to be put into the same weight level. Then, the output matrix is the final matrix to be ready for the feature selection and model training.

As mentioned in 5.1.2, sklearn offers most of the feature selection algorithms mentioned in section 3.3. As the TF-IDF value is applied as the feature weight, the DF feature selection method is abandoned in the testing. The rest of the methods including IG, CHI and PMI are all applied in testing. All the methods can be directly imported from the sklearn package. In much traditional text classification literature, the number of selected features is set into some specific number levels for testing. However, in the author's opinion, the testing result would be more general using the percentage because with the change of the data collection size the entire number of the features would in a long range. In this case, the different percentage of selected features, from 10 to 100 with 10 percentage increasing step is tested.

The selected data collection matrix is the final matrix for the model training and testing. According to the system flow, the data matrix needs to be split into a training set and a testing set. In the statistic field, the K-cross-fold-validation is the one of the most popular validation method to guarantee the randomness of the test set and result [Kohavi, 1995]. As mentioned in 3.1, the 10-cross-fold-validation is widely applied in the split of the training and testing set. This method split the whole data collection equally into 10 parts and randomly picks 1 part as the testing set, while the rest is used as a training set. This progress is randomly repeated 10 times in this method and the result will be the average result of the 10-time experiment.

The classifiers are also supplied by the sklearn and can be directly imported. All the classifiers mentioned in 3.4 are applied in the testing. Among the classifiers, the SVM is implemented with two algorithms which respectively are LinearSVC and SVC [Scikit-learn, 2018]. The two algorithms are implemented differently in the strategy of finding the hyperplane. The LinearSVC use the "one to rest" strategy to find the hyperplane which means when processing with one category the algorithm will try to find the best hyperplane between this category vectors and all the rest of the vectors. The SVC uses the "one to one" strategy to find the hyperplane which means when processing with one category the algorithm will try to find the best hyperplane between this category vectors and every other neighbor category vectors. From the two strategies, it can be indicated that the LinearSVC would perform faster and SVC would be more accurate. However, their practical performance needs to be tested to evaluate.

6. Testing and evaluation

The evaluation indicators are the indicators to evaluate the performance of the classification system and classifiers [Li, 2013]. The most popular evaluation indicators for the classification system performance are precision, recall, and F1 score.

The precision refers to the ratio of the number of correctly categorized documents to the total number of categorized documents, the precision is calculated with the following formula:

$$\text{Precision} = \frac{\textit{Correct categorized document amount}}{\textit{Total categorised document amount}}$$

The recall refers to the ratio of the number of correctly categorized documents amount to the total number of the document, the recall ratio can be calculated with the formula below:

$$\text{Recall} = \frac{\textit{Correct categorized document amount}}{\textit{Total document amount}}$$

The F1 score is a value to generalize the precision and the recall ratio to offer the high precision while low recall ratio result or the low precision while high recall ratio results in a new reference evaluation dimension. The F1 score can be calculated with the below formula:

$$\text{F1 score} = \frac{\textit{Precision} \times \textit{Recall} \times 2}{\textit{Precision} + \textit{Recall}}$$

6.1. Full features data collection testing

The original data set contains totally over 120000 records and consists of 114 manually classified test data categories. However, most of the records in the data collection contain less than 1000 records which are meaningless in generalizing the problems in the customer service. Therefore, only 43 categories which all contain more than 1000 pieces of records are picked out from data collection to guarantee the reference value of the result. After deleting the repeating records, the final shape of the matrix is (79796,5722), which means 79796 pieces of records and 5722 features. What's more, in case that some of the classifiers run too slow in the whole data test. The top 10 categories consist of 14332 pieces of records are also picked to run all the classifiers that to be considered as solution classifiers.

Below are the results of the classifiers' performance with all features used in 14332 items data collection

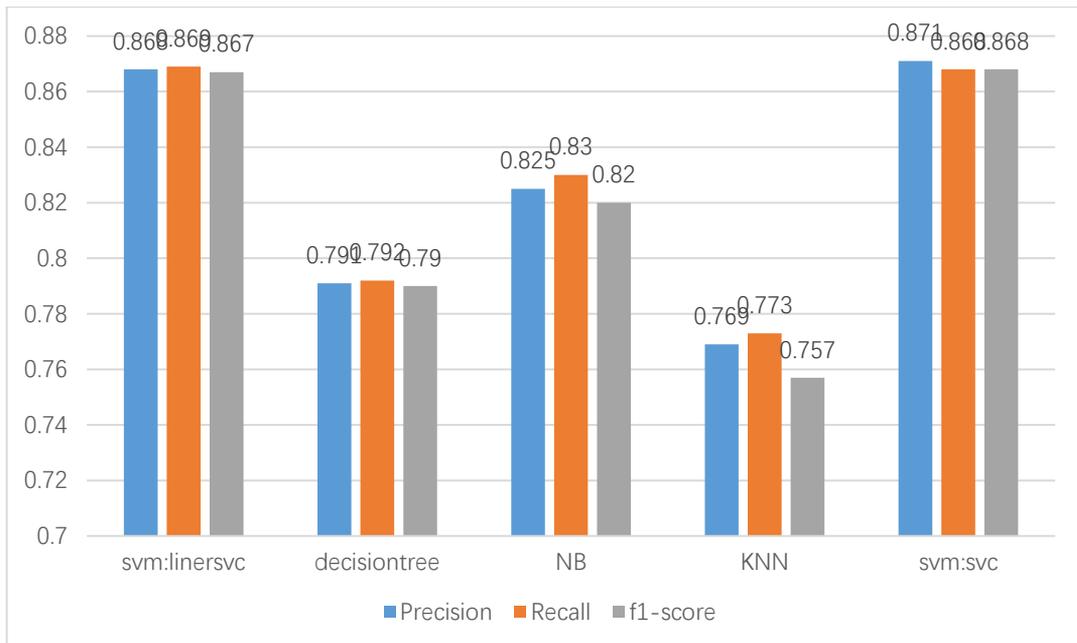


Figure 16 The classifiers' performance results with full features in small amount data collection

In the classifiers' performance test, the KNN was chosen to use the range of 15 to 20 neighbors which was recommended by many papers, in the testing the best result was achieved with 19 neighbors, presented in the figure above. The figure clearly demonstrates that within all the applied classifiers, the two types of SVM perform much better than the rest of the other classifiers. The SVC gains better precision while the LinearSVC achieves better recall score. From the F1 score, it can be inferred that the two types of SVM are similar in the general performance. However, considering of the time cost, the LinearSVC is obviously more suitable to be applied in the original data collection test. In the rest of the other classifiers, although recorded the best result in testing, the KNN is left far behind the other classifiers in all the testing aspects. The NB gains the best results among all the rest three classifiers which all the scores are no less than 0.82. The DT does not reach the 0.8 in all the three evaluation indicator scores but achieves good results above 0.79.

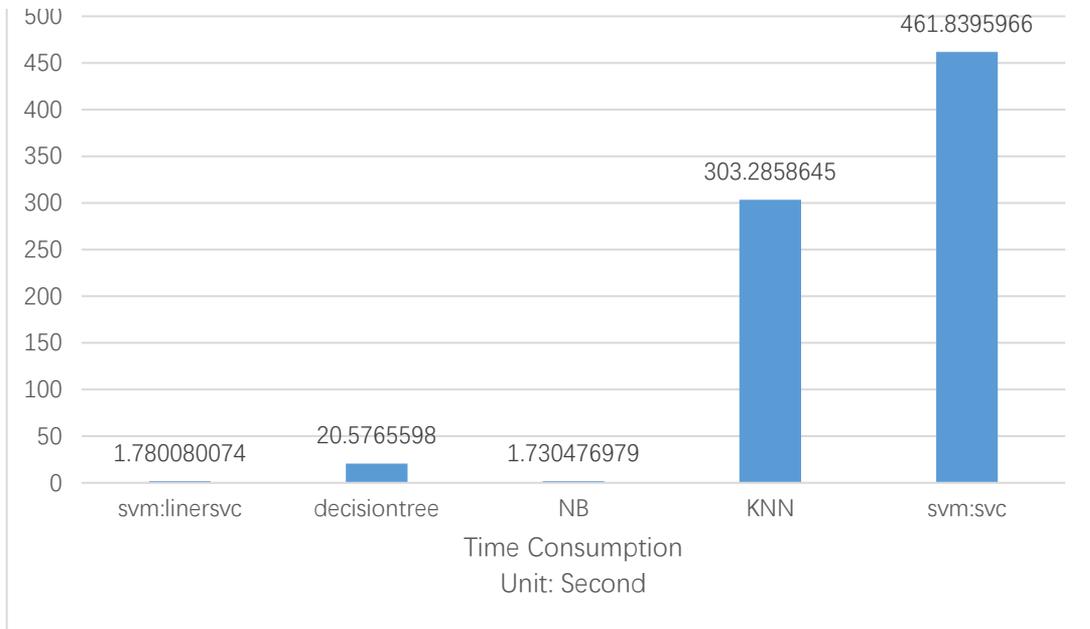


Figure 17 Classification time consumption of all classifiers

The figure above indicates the classification time consumption of all the classifiers. From the figure, the large distance of time cost in the model training between KNN, SVC and other classifiers are clearly demonstrated. Although the final solution would mainly focus on the classification result, too large time consumption will also be considered as an indicator to judge the performance of the classifiers. Therefore, generally considering the classifiers' classification evaluation results and time consumption, the KNN and SVC will not be applied in the full data collection testing.

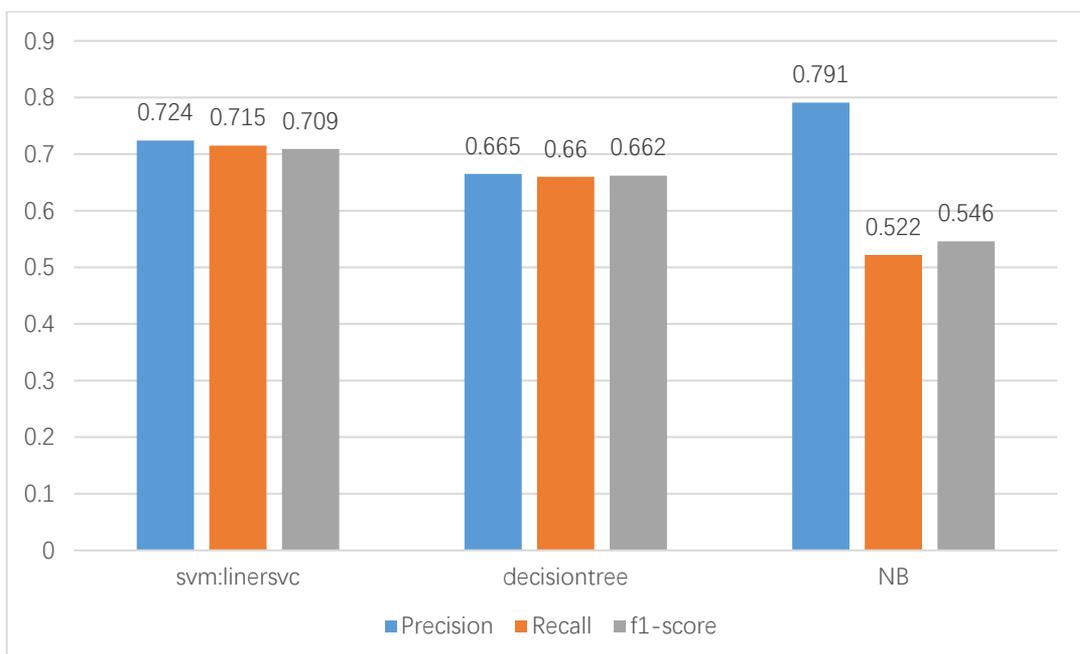


Figure 18 The classifiers' performance results with full features in full data collection

Figure 18 presents the classifiers' performance results with full features in full data collection. In the classification of a large amount of data collection, all the classifiers experienced a serious decrease in all the evaluation indicators. Among them, the LinearSVC and DT decrease by around 0.15 and 0.13 in all the three indicators. The balance of the recall score and precision is kept well in these two classifiers. However, a huge difference appears between the recall score and precision in the NB's result whose precision is 0.791 while the recall score is only 0.522. The difference leads to the low F1 score in NB which is only 0.548. The low recall score indicates that a large amount of testing data is failed in classification by NB model.

Within the three classifiers, the NB takes 15.410 seconds to finish the classification, while the LinearSVC takes 19.527 seconds. The DT is the slowest takes 166.925 seconds. Generally, in the full features, testing round the LinearSVC achieve the best comprehensive performance considering of both evaluation indicators aspect and time-consuming aspect.

6.2. Traditional feature selection methods testing

In this test, three feature selection methods mentioned above will be applied including CHI, IG, and PMI. The number of selected features will be based on the percentage steps. The range of the steps will be 10 to 90 with 10 percentage as step increment. All the three classifiers will be inquired in this test to see their performances in different feature numbers. Below are the results of the CHI method applied in three classifiers.

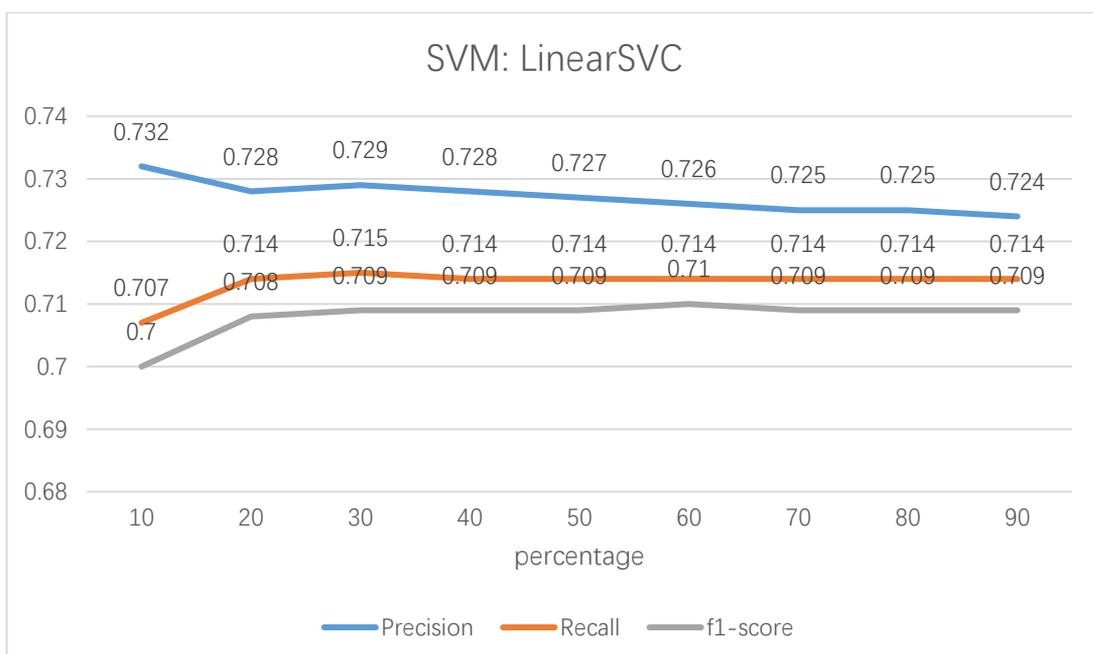


Figure 19 CHI method applied in LinearSVC classifier

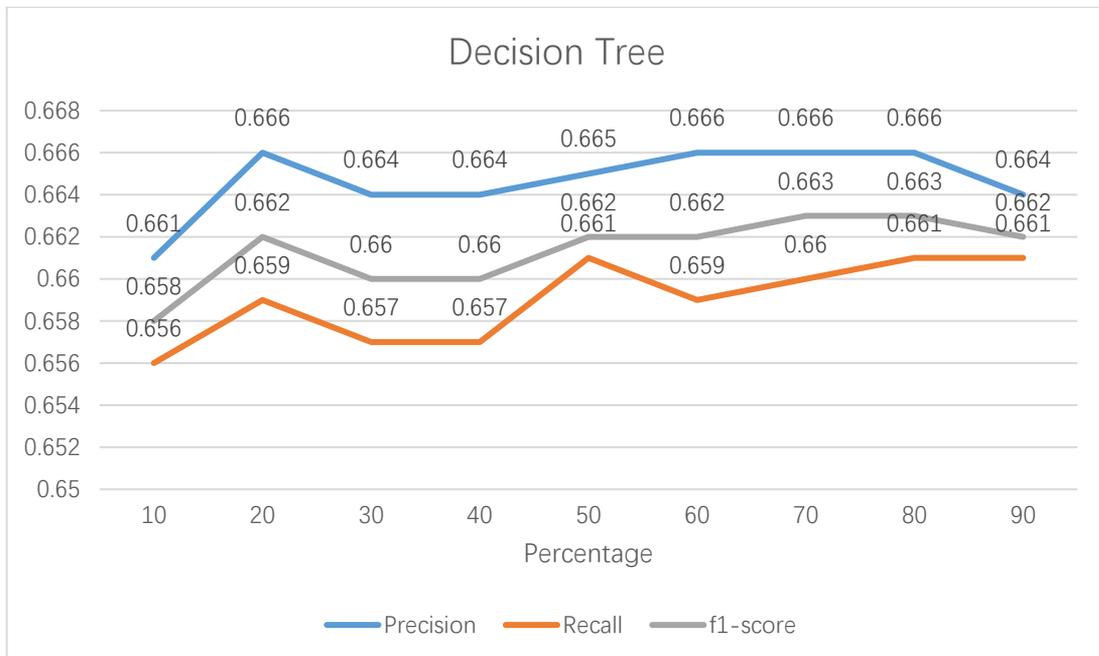


Figure 20 CHI method applied in Decision Tree classifier

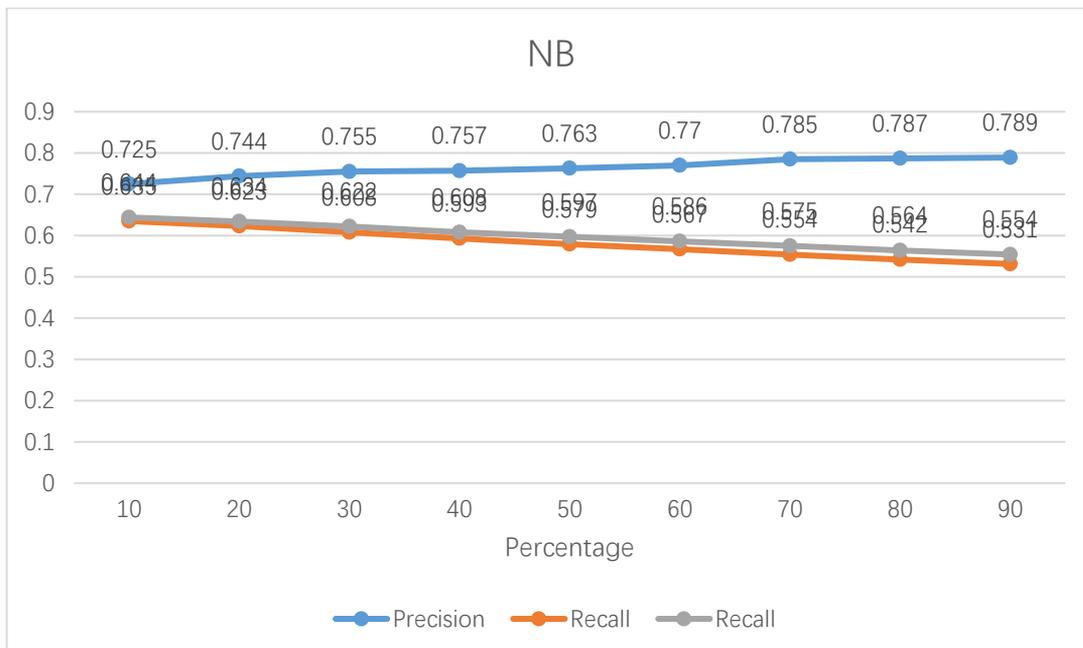


Figure 21 CHI method applied in the NB classifier

From the above figures, all the three classifiers achieve better results by applying the CHI method and keeping part of the features instead of using all the features for classification. The LinearSVC gains the best results by selecting 60% of the features while DT and NB gain their best results by selecting 80% and 50% of the features. According to the best results, the CHI does successfully maintain and even optimize a bit of the result in all

three classifiers by deducing a large part of the features. The results of the IG and PMI are listed below:

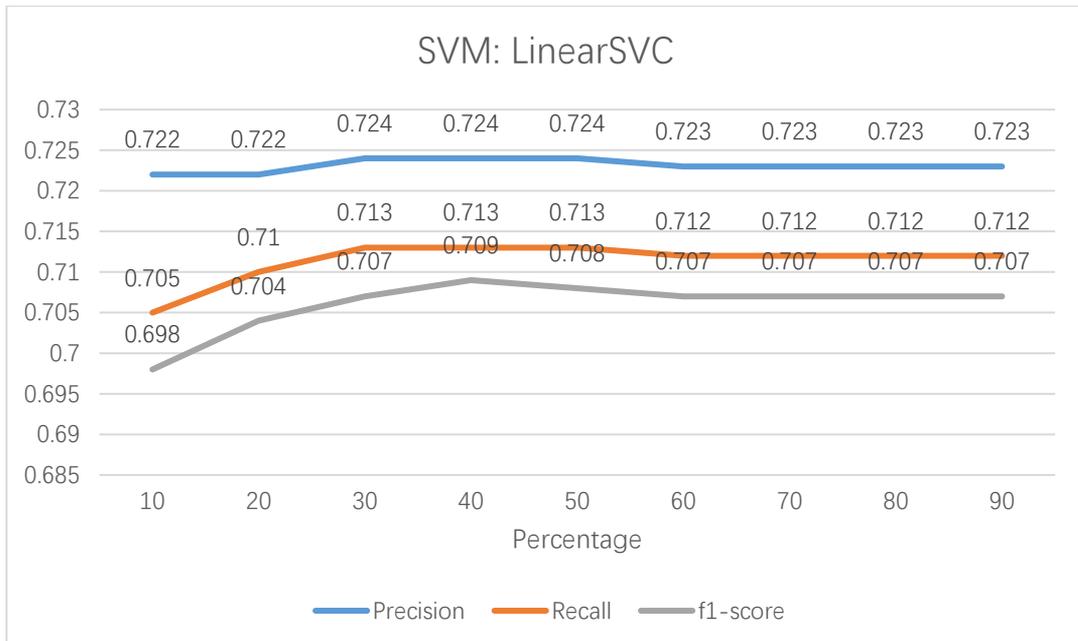


Figure 22 IG method applied in LinearSVC classifier

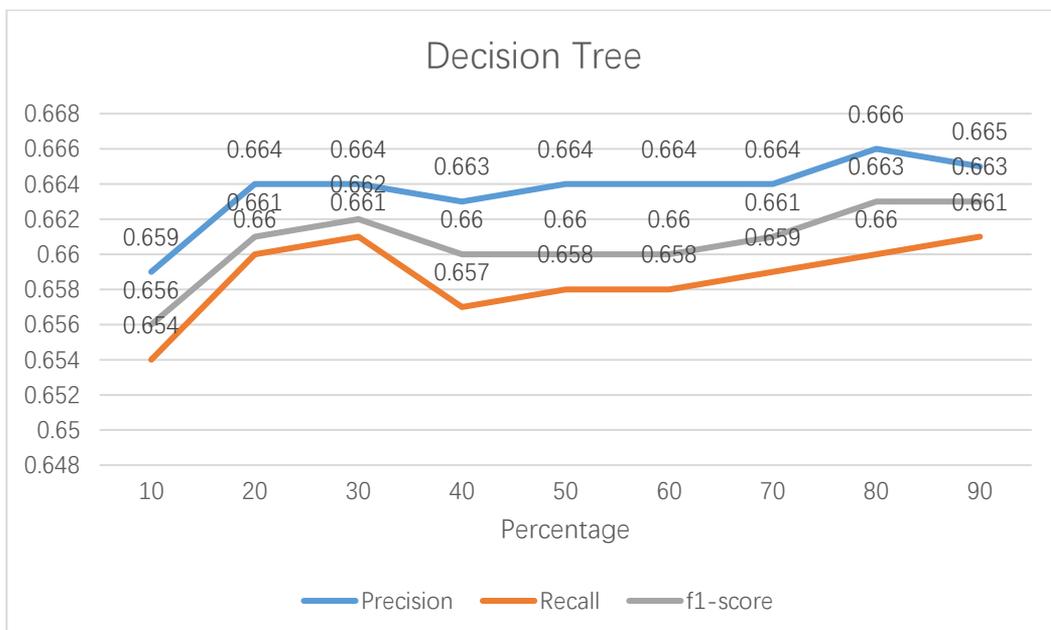


Figure 23 IG method applied in Decision Tree classifier

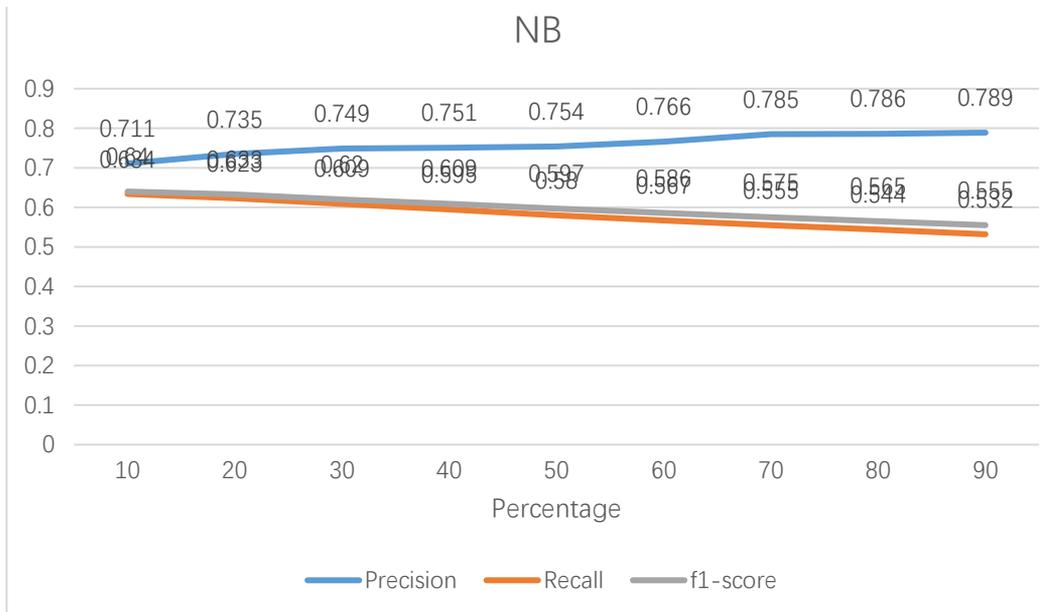


Figure 24 IG method applied in the NB classifier

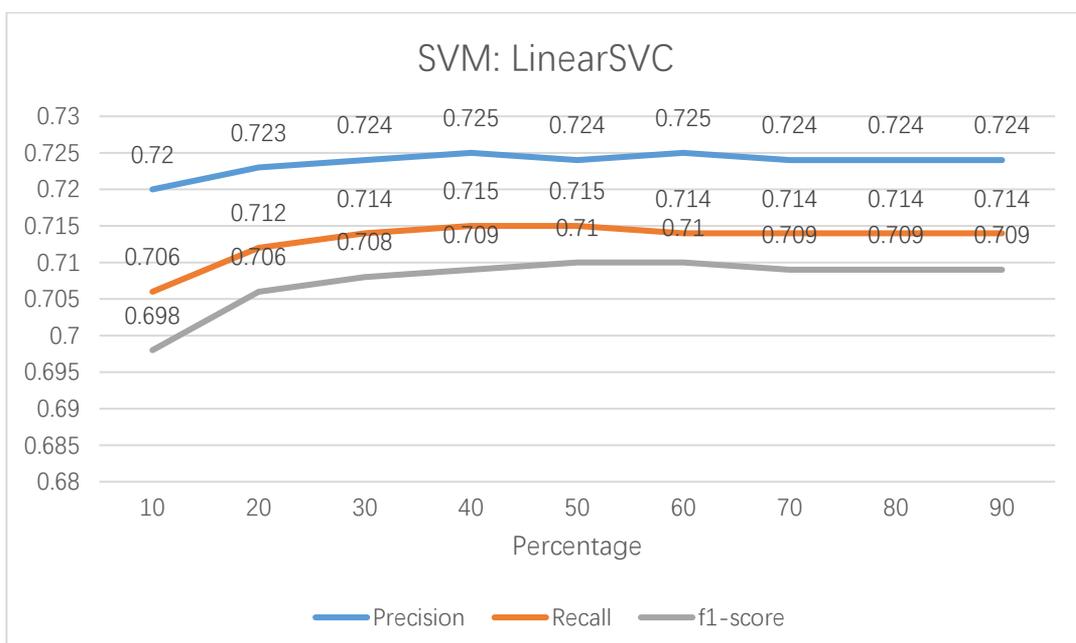


Figure 25 PMI method applied in LinearSVC classifier



Figure 26 PMI method applied in Decision Tree classifier

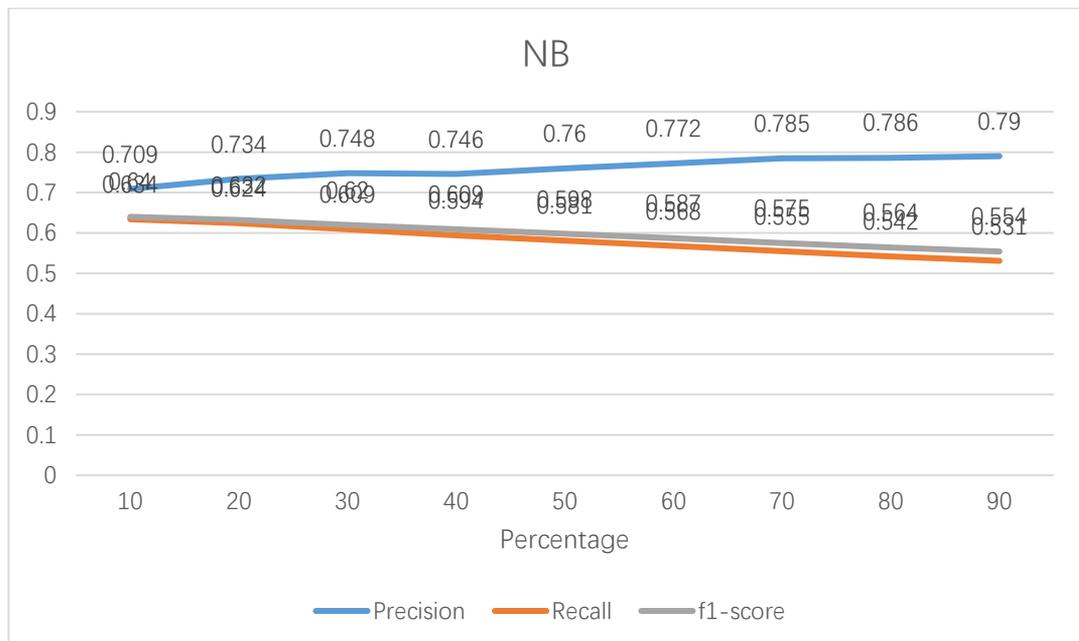


Figure 27 PMI method applied in NB classifier

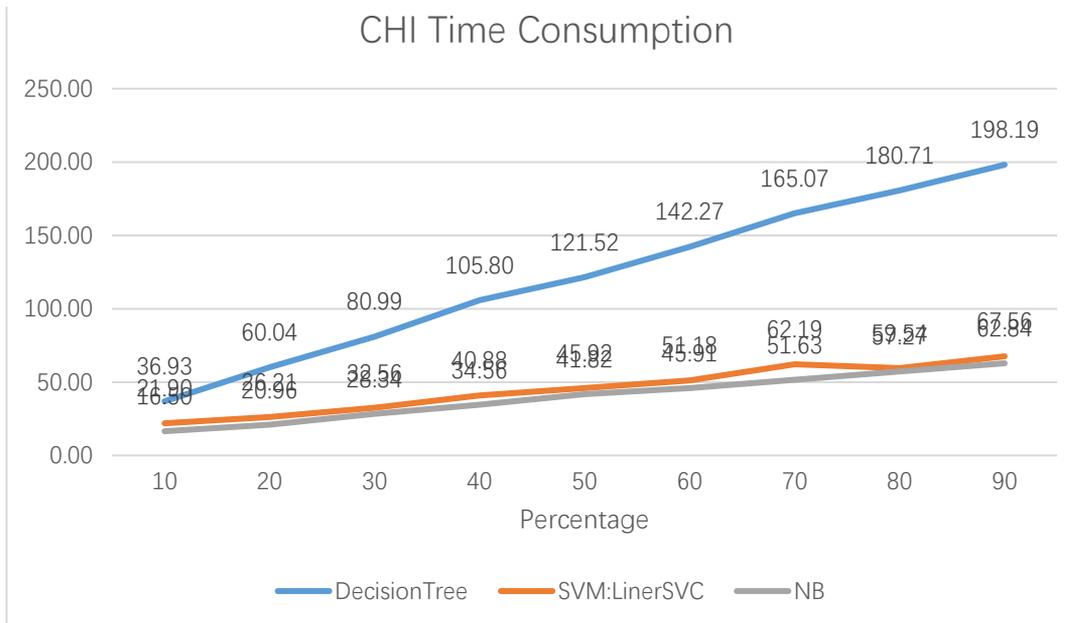


Figure 28 CHI method applied time consumption

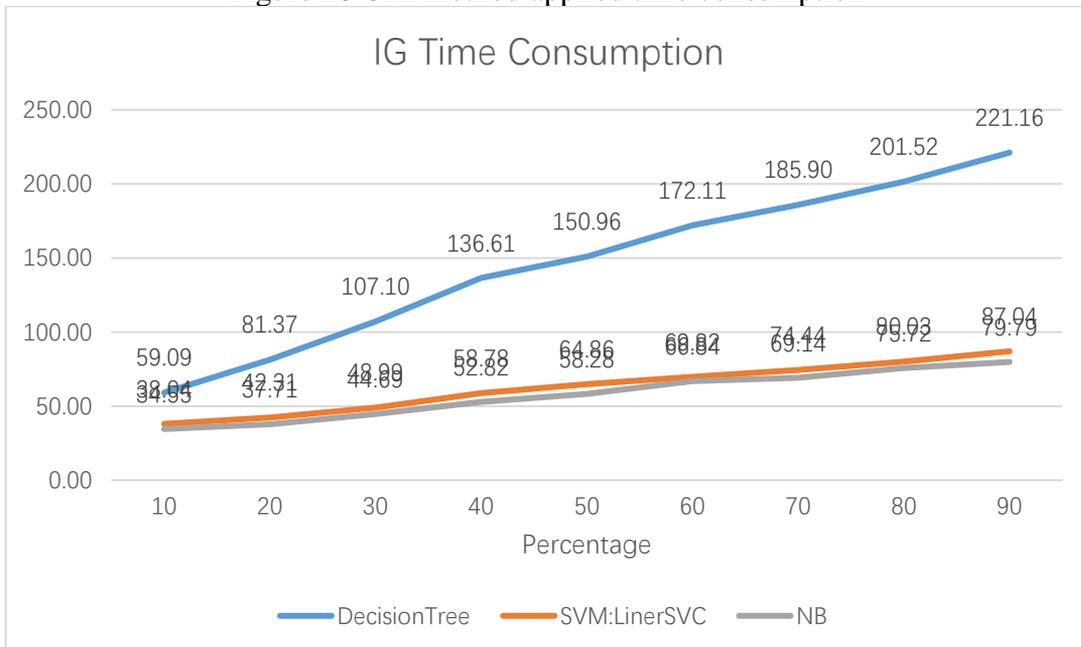


Figure 28 IG method applied time consumption

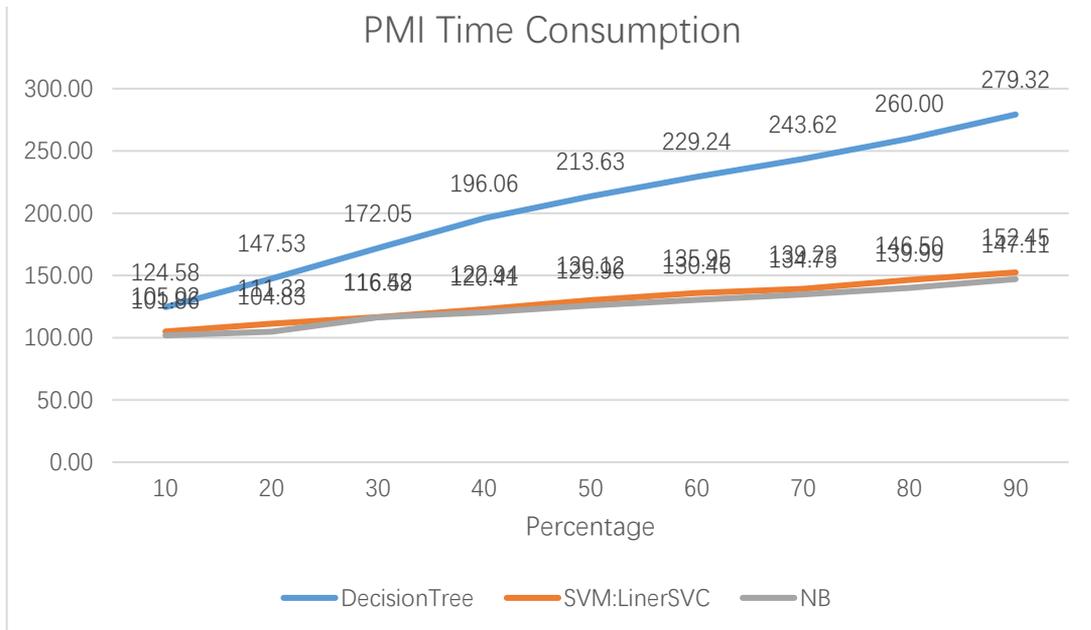


Figure 28 PMI method applied time consumption

The figures above show the testing results of IG and PMI applied in the three classifiers. In general, the IG and PMI gain a similar curve to CHI when applied to the three classifiers. However, for IG or PMI, the best results are no better than that of CHI. What's more, in the time consumption figures, the CHI cost much less time than PMI and IG in the classification progress. Therefore, in the enhancement feature set test, the CHI will be chosen as the feature selection method.

6.3. Enhancement features data collection testing

As mentioned in 5.1.1 the dependency parser will be used to reform the feature collection into new enhancement feature collection. The enhancement feature collection will be used in the classification system to evaluate the performance of this solution. Below are the testing results of the full enhancement feature collection used in three classifiers.

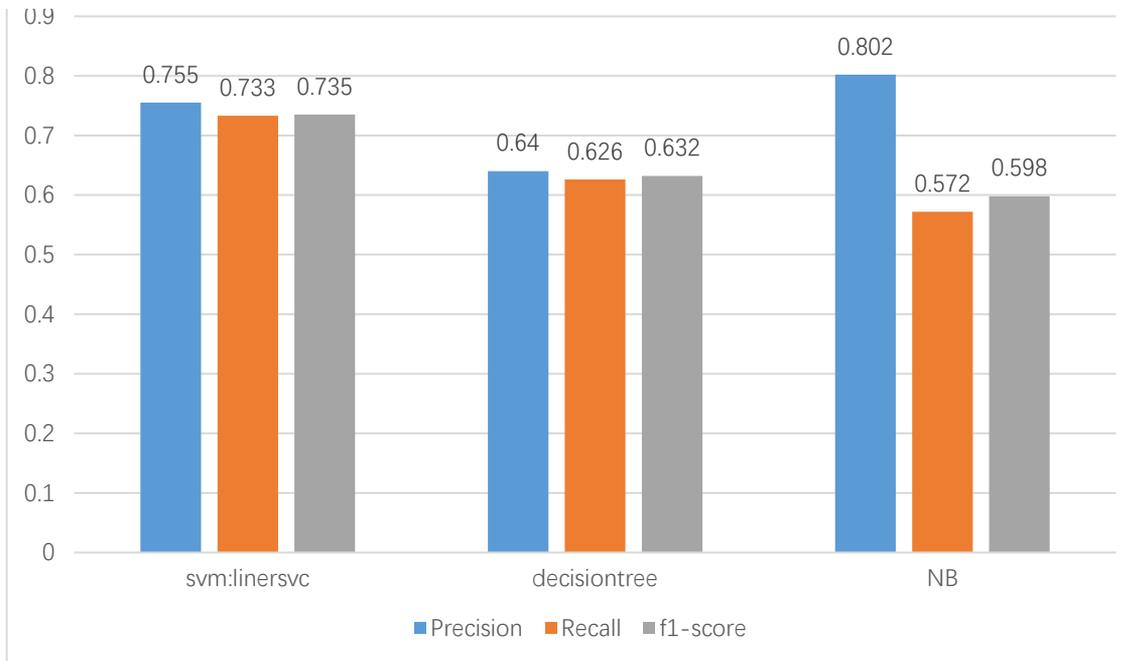


Figure 29 Classifiers' performance with enhancement feature collection

The performance testing results presented in the figure shows clearly that by using the enhancement feature collection, two classifiers achieve better scores in all testing indicators. Among them, the LinearSVC increase 0.31 in precision and 0.18 in recall score which makes the F1score grow by 0.26. The NB gains high precision of 0.802 which goes above 0.8 while reaching to the 0.572 in recall score which is increased by 0.5. However, the Decision Tree classifier decreased around 0.2 in all three indicators. According to the results, the LinearSVC is the most suitable to be the classifier in the final solution.

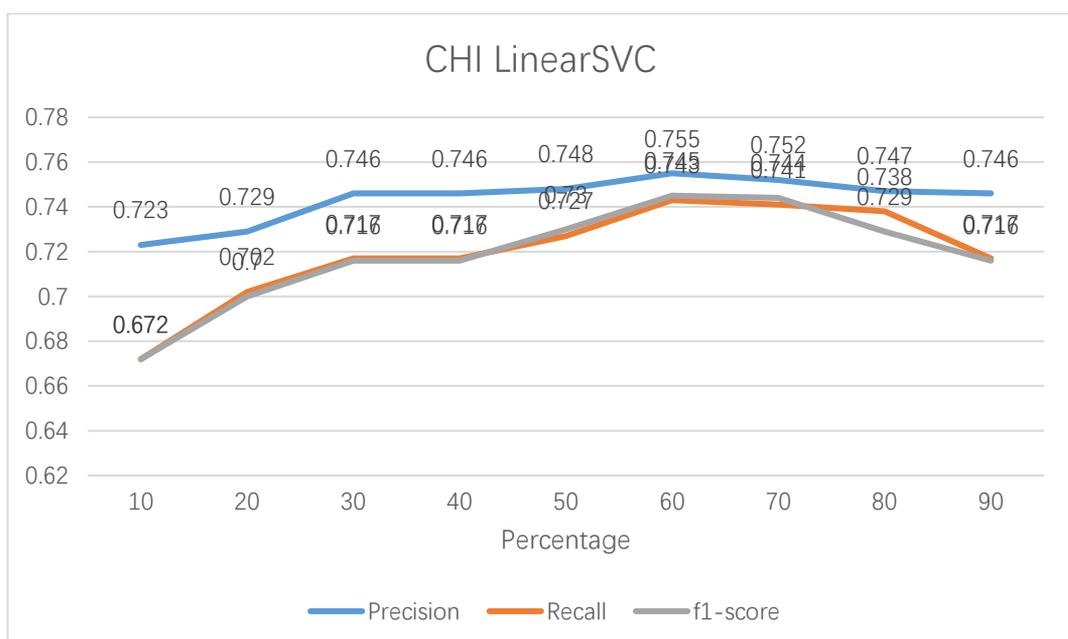


Figure 29 The performance testing of the final classification solution

The final solution of the classification system uses the enhancement feature collection as input feature collection. The CHI is chosen as the feature selection method with the LinearSVC chosen as the classifier. The figure demonstrates the performance of the final classification solution changes with different numbers of features. In the final solution, the classification achieves the best indicators results by keeping 60% of the features. Compared to using the original feature collection, the precision of the final solution increases by 0.28 while the recall score increases by 0.31 which results in the 0.36 increment in F1-score. The result proves that the performance optimization by applying the enhancement feature collection in the final solution is effective and successful.

6.4. Discussion

The proposed solution successfully manages to resolve the DiDi's customer service records data classification problem with supervised learning methods. In the solution testing, all the popular feature selection methods and classifiers have been tested to see their performance in processing the real data. The result is not only meaningful to the specific case of DiDi but also indicates their performance to the general short text data.

According to the special expression regulation of the customer service records data. The solution offers a way to apply the NLP field method Dependency Parser to combine the word features into short phrases features. This new way allows the human to intervene the classification progress and define the syntax pattern to make the feature collection more semantic and meaningful compared to fully rely on the statistical algorithms. The testing result proves that the new type of enhancement feature collection performs better than the original feature collection.

6.4.1. Limitation and future work

In the testing result, all the evaluation indicators of the proposed solution have reached around 0.75. However, the testing sample is just the data collection for a short period. Therefore, to be applied in the real processing, the results still need to be improved at least above the 0.8 for practical use. Obviously, there is still room for improvement in many aspects.

In terms of the supervised learning algorithm, apart from the classical algorithm introduced in the thesis, there recently appears more distribution, linear algebra based dimension deduction methods and regression, disperse based classifiers. By applying

more advanced algorithms in feature selection and classification, we may expect better performance of the classification system.

From the aspect of building the enhancement feature, using the SBV dependency relation as the syntax pattern is just an initial trial with the dependency parser. From the analysis result of the document sample sentence. Other relations are also worth trying to define the syntax pattern.

From the aspect of efficiency, with feature selection and enhancement feature collection building steps added in the system, the time consumption of solution is a couple times more than processing with the original feature collection. Therefore, whether the model training is the main time consumption part of the large data processing needs to be further tested to evaluate the value of feature selection. In the meantime, within the time of the big data, more and bigger data processing platforms are getting popular in distributed data processing. Implementing the solution in the big data platform such as Hadoop or Spark would also allow further exploration to improve the efficiency of the solution.

7. Conclusions

This thesis presents an entire solution for the Chinese O2O company DiDi to categorize their customer service records data into the pre-defined categories. The solution is based on the traditional text categorization flow and introduces a way to build an enhancement feature collection instead of using the original feature collection. Classic supervised learning algorithms in the traditional text categorization flow are also demonstrated in the thesis. Tests are conducted to validate the performance of inquiring different algorithms and feature collection in the system.

With the O2O industry entering the oligarchic era, many O2O companies in China like DiDi face the challenges of acquiring the helpful information to improve the service quality and obtain the new requirements. As big data processing and machine learning are getting popular, one of the direct ways to fulfill the needs of DiDi is to acquire the information from the customer service records. With that purpose, automatic classification of the customer service records data is the initial step.

The literature presented above is about the traditional supervised learning text classification flow including the steps of pre-processing, document representation, feature selection, and model training. Some classic algorithms of feature selection and model training are also specifically introduced. The related performance tests are conducted.

The traditional text classification flow is mainly implemented by applying the statistical algorithms which from the view of NLP are not semantic and linguistic. By analyzing the sample documents from the data collection, the author found that the expression of the whole records is following some general regulation which offers the possibility to the solution to add the step which can make it more semantic and human intervening.

Finally, the solution includes a way for a human to be able to pre-define the syntax to reform the feature collection. The method from the NLP field named dependency parser is introduced to obtain the pre-defined syntax. The reformed feature collection is named enhancement feature collection to be input into the traditional text classification flow instead of the original feature collection.

Tests were conducted to test the performance of the solution and all the introduced feature selection algorithms and classifiers. Three indicators including precision, recall score and F1-scores are used to evaluate the system performance. Test results indicated that the most effective and best performance solution based on the traditional text classification

flow which chooses the enhancement feature collection as the input features while CHI as feature selection method and SVM: LinearSVC as the classifier. The performance figure is acceptable but still has room for improvement.

In general, the solution successfully fulfills the requirement from DiDi to automatically categorize the customer service records data. It also offers a new way to apply text analysis approaches in optimizing the result and make the solution more semantic, linguistic and human intervening.

References

- [Stan, C. 2011] An Entrepreneur in E-commerce. Zhihu. November 2011
- [Shen, C., & Wang, Y. 2014] Shen, Chentao & Yongle Wang. Online to Offline Business Model.
- [Hui, S. C., & Jha, G. 2000] Hui, Siu Cheung & G. Jha. Data mining for customer service support. *Information & Management*, 38(1), 1-13.
- [Pohl, K. 2010] Pohl, Klaus. Requirements engineering: fundamentals, principles, and techniques. Springer Publishing Company, Incorporated.
- [Tan, A. H. 1999] Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*. Vol. 8. 1999.
- [Feldman, R., & Sanger, J. 2007] Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.
- [Rampell, A. 2010]. Rampell, Alex. Why Online2Offline Commerce is a trillion dollar opportunity. *Techcrunch.com*. Available online at <http://techcrunch.com/2010/08/07/why-online2offline-commerce-is-a-trillion-dollaropportunity/> (2010).
- [Wang, L.Y. 2013] What is O2O. *IT Industry*, Vol.20, No.1, 2013
- [Yang, X., Liang, M., & Wu, J.,2014] Yang Xu, Liang Min, Wu Jun Model of Customer Trust in O2O E-commerce Based on Fuzzy Comprehensive Evaluation Method. *Journal of Beijing University of Posts and Telecommunications: Social Sciences Edition*, 2014 (3): 45-51.
- [Li, D.Z., & Dou, W.F., 2013] Li DongZhen & Dou WanFeng. A Study of Organization Model Based on O2O E-commerce. *Electronic World* 4 (2013): 10-11.

- [Guo, Y.Q., 2015] Guo, YongQiang Study of O2O E—commerce in the Mobile Internet Era Hunan Normal University, 2015
- [Lu, Y.Q. & Li, C., 2013] Lu, YiQing & Li, Chen Study of O2O Business Mode and Development Prospect. Enterprise Economy. 32.11 (2013): 98-101.
- [Tian, A.G., 2017] Tian, AiGuo The Analysis of the Present Situation and Further Development of O2O E-Commerce in the Context of Mobile Internet. Journal of Chongqing University of Posts and Telecommunications(Social Science Edition) Vol. 29 No. 4 (2017): 104-112.
- [Li, H.S., Ren, M.Y. & Xu, L.P., 2016] Li, HongShuang, Ren, MingYang & Xu, LiPing The Development Status, Problems and Solutions of O2O Development Model. Modern Economic Information 2016(7): 306
- [iResearch, 2017] iResearch Chinese Local Life O2O Industry Research Report of 2017. Available at: <http://www.iresearch.com.cn/report/3024.html>.
- [ebrun, 2017] ebrun 2017 The Second half of O2O Development Analysis Report. Available at: <http://www.ebrun.com/20170515/230683.shtml>
- [Qi, P.C. & Yang, J.Z., 2016] Qi PengCheng & Yang JianZheng The Collaborative and Monopoly of O2O Magnates “merging the similar terms”. Electronic Commerce 2016(9): 10-12
- [DiDi., 2017] DiDi About DiDi. Available at: <http://www.xiaojukeji.com/website/about.html>
- [DiDi., 2017] DiDi About DiDi Products. Available at: <http://www.didichuxing.com/en/aboutus/products>
- [Mohri, M., Rostamizadeh, A., & Talwalkar, A. 2012] Mohri Mehryar, Afshin Rostamizadeh and Ameet Talwalkar Foundations of machine learning. MIT press.2012
- [Wikipedia., 2017] Wikipedia. The explanation of “Unsupervised Learning” available at: https://en.wikipedia.org/wiki/Unsupervised_learning

- [Michael, I. J. & Christopher, M. B., 2004] Michael, I. Jordan & Christopher M. Bishop. Neutral Networks. In Tucker, Allen B., ed. Computer science handbook. CRC press, 2004: 66
- [Chaovalit, P., & Zhou, L.2005] Chaovalit, Pimwadee & Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on (pp. 112c-112c). IEEE.
- [Lee, C. H., & Yang, H. C., 2009] Lee, Chung-Hong & Hsin-Chang Yang. Construction of supervised and unsupervised learning systems for multilingual text categorization. Expert Systems with Applications, 2009, 36(2), 2400-2410.
- [Dai, L., Huang, H., & Chen, Z. 2004] Dai, Liuling, Heyan Huang, and Zhaoxiong Chen. A comparative study on feature selection in Chinese text categorization. Journal of Chinese Information Processing, 18(1), 26-32.
- [Sparck Jones, K. 1972] Sparck Jones, Karen. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11–21.
- [Sparck Jones, K. 2004] Sparck Jones, Karen. IDF term weighting and IR research lessons. Journal of Documentation, 60(6), 521–523.
- [Jin X.B. 2006] Jin XiaoBo. A Survey on Text Categorization. Automation Panorama, (z1), 24-29.
- [Zhou, Q. Q., Sun, B. D., & Wang, Y. 2005] Zhou QinQiang, Sun BingDa & Wang Yi. Study on New Pretreatment Method for Chinese Text Classification System. Application Research of Computers, 2, 85-86.
- [Uğuz, H. 2011] Uğuz, Harun. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024-1032.
- [Mitchell, T. M. 1997] Mitchell, Tom M. Machine learning. WCB.
- [Zheng, W. & Wang R. 2007] Zheng Wei & Wang Rui. Comparative Study of Feature Selection in Chinese Text Categorization. Journal of Hebei North University: Natural Science Edition, 23(6), 51-54.

- [Yang, Y., & Pedersen, J. O. 1997] Yang, YiMing & Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In ICML (Vol. 97, pp. 412-420).
- [Church, K. W., & Hanks, P. 1990] Church, Kenneth Ward & Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), 22-29.
- [Fan, X.L. & Liu, X.X. 2010] Fan, Xiao-Li, and Xiao-Xia Liu. Study on mutual information-based feature selection in text categorization. *Computer Engineering and Application*, 46(34), 123-125
- [Li, T.B. 2013] Li TaiBai. Research of Feature Selection Algorithm in Short Text Classification. Diss. Chongqing: Chongqing Normal University.
- [Shen, H., Lu, B.L., Utiyama, M., & Isahara, H. 2006] Shen Hong, Lu BaoLiang, Utiyama Masao & Isahara Hitoshi. Comparison and Improvement of Feature Extraction Methods for Text Categorization. *Computer Simulation*, 23(3), 222-224.
- [Wang, X. W., Fan, X. H., & Zhao, J. 2009] Wang XiWei, Fan XinHua & Zhao Jun. Method for Chinese Short Text Classification Based on Feature Extension. *Journal of Computer Applications*, 29(3), 843-845.
- [Agrawal, R., Imieliński, T., & Swami, A. 1993] Agrawal, Rakesh, Tomasz Imieliński, & Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.
- [Lewis, D. D. 1998] Lewis, David D. Naive (Bayes) at forty: The Independence Assumption in Information Retrieval. In *European Conference on Machine Learning* (pp. 4-15). Springer Berlin Heidelberg.
- [Quinlan, J. R. 1986] Quinlan, J. Ross. Induction of Decision Trees. *Machine learning*, 1(1), 81-106.
- [Hunt, E. B., Marin, J., & Stone, P. J., 1966] Hunt, Earl B., Janet Marin, & Philip J. Stone. Experiments in induction.
- [Wang, Y., 2006] Wang, Yu. Study on Text Categorization Based on Decision Tree and K Nearest Neighbors. Doctoral dissertation, University of Tianjin, 2006

- [Quinlan, J. R., 1983] Quinlan, J. Ross. Learning efficient classification procedures and their application to chess end games. In *Machine Learning, Volume I, 1983* (pp. 463-482).
- [Cui, Z, Y. 2011] Cui, Zheng Yan. Research of Chinese Short-Text Classification [D]. Doctoral dissertation, Kaifeng: University of Henan, 2011
- [Tripathy, A., Agrawal, A., & Rath, S. K. 2016] Tripathy, Abinash, Ankit Agrawal, & Santanu Kumar Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57, 117-126.
- [Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. 1992] Brown, Peter F., et al. Class-based n-gram models of natural language. *Computational linguistics*, 18(4), 467-479.
- [Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. 2014] Sidorov, Grigori, et al. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3), 853-860.
- [Rak, R., Stach, W., Zaïane, O. R., & Antonie, M. L. 2005] Rak, Rafal, et al. Considering re-occurring features in associative classifiers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 240-248). Springer, Berlin, Heidelberg.
- [Liu, H. 2007] Liu Hua. Text Categorization Based on Key Phrases. *Journal of Chinese Information Processing*, 21(4), 34-41.
- [Yan, H. U., Hu-zi, W. U., & Zhong, L. 2007] Yan, H. U., W. U. HuZi, & Zong Luo. Research of Feature Extraction Methods Based on Part of Speech in Chinese Documents Classification [J]. *Journal of Wuhan University of Technology*, 4, 037.
- [Lin L.L. 2014] Lin LanLan. Review feature selection based on syntax pattern. *Journal of Guangdong Technical College of Water Resources and Electric Engineering*, (4), 24-26.
- [Zheng-yu, Z. H. U., Cun-qing, L. I., & Peng, Z. 2010] Zheng-yu, Z. H. U., L. I. Cun-qing & Zhang Peng. Topic Words and Opinion Words Extraction from Chinese

Product Reviews Based on Syntax Pattern [J]. Journal of Chongqing University of Technology (Natural Science), 5, 018.

[Tesnière, L. 1953] Tesnière, Lucien. Elements of structural syntax. John Benjamins Publishing Company.

[Robinson, J. J. 1970] Robinson, Jane J. Dependency structures and transformational rules. Language, 259-285.

[Bai, M.Q., Zheng, J.H. 2004] Bai MingQing & Zheng JiaHeng. Study on Ways of Verb-Verb Collocation. Computer Engineering and Applications, 40(27), 70-72.

[Tang, X.B., & Xiao, L. 2014] Tang XiaoBo & Xiao Lu. Research of Text Feature Extraction on Dependency Parser. Data Analysis and Knowledge Discovery, 30(11), 31-37.

[LTP, 2018] LTP. Syntax Dependency Relation Tags and Definitions. Available at: <https://www.ltp-cloud.com/intro/>

[Python, 2018] Python. Overview. Beginners' Guide of Python. Available at: <https://wiki.python.org/moin/BeginnersGuide/Overview>

[NLPIR, 2018] NLPIR. About NLPIR Available at: <http://www.nlpir.org/?action=aboutus>

[Pandas, 2018] Pandas. Powerful Python Data Analysis Toolkit. Available at: <http://pandas.pydata.org/pandas-docs/stable/#pandas-powerful-python-data-analysis-toolkit>

[Scikit-learn, 2018] Scikit-learn. User Guide. Available at: http://scikit-learn.org/stable/user_guide.html

[Kohavi, R., 1995] Kohavi, Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

[Scikit-learn, 2018] Scikit-learn API Reference. Available at: <http://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm>