National College of Ireland

Sentiment Analysis using machine learning algorithms: online women clothing reviews

MSc Research Project
Data Analytics

Shuangyin Xie
Student ID:
x18126634

School of Computing
National College of Ireland

Supervisor: Bahman Honari

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shuangyin Xie
Student ID:	x18126634
Programme:	Msc Data Analytics
Year:	2019
Module:	Research Projcet
Supervisor:	Bahman Honari
Submission Due Date:	12/12/2019
Project Title:	Sentiment Analysis using machine learning algorithms:online women clothing reviews
Word Count:	XXX
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis using machine learning algorithms: online women clothing reviews

Shuangyin Xie
X18126634

Abstract

Internet technology has been closely related to life. It not only convenient people's lives but also allows people to share information, especially in the field of e-commence. People leave message and share their feelings online. As a result, sentiment analysis becomes more and more attracted. Accurate sentiment analysis not only allows customers to better understand the product, but also enables the company to get better feedback from the market. In this paper, we use data set from online women clothing reviews to conduct sentiment analysis, which can be downloaded from Kaggle. The machine learning methods used in this research are Support Vector Machine, Logistic Regression, Random Forest, Naive Bayes. All experiments were done in this research using python. We evaluate the model in terms of accuracy, precision, recall, F1-score and Area Under Curve(AUC). This study provides us with sentimental analysis of various women clothing opinions dividing them Positive, Negative and Neutral behaviour. These data suggest that the Naive Bayes gives highest accuracy to classify the Reviews, which is 93%.

Keywords: sentiment analysis, machine learning, Support Vector Machine, Logistic Regression, Random Forest, Naive Bayes.

1 Introduction

With the explosive growth of social media on the web(Liu and Zhang; 2012),large amounts of data and information are produced and shared across the social media every day(Ali et al.; 2019). There is huge number of products online and each product may have hundreds of reviews.How can customers gain useful information from so many reviews? How can organizations gain useful feedback about the product business? Sentiment analysis comes out. The process of sentiment analysis is to determine whether review texts are negative, positive or neutral(Alrehili and Albalawi; 2019).

1.1 motivation and background

Sentiment analysis can often quantify the positive or negative level of the main subject of the text(Gräbner et al.; 2012). For example, a mother wants to buy clothes for her kids, she searches online. There are many reviews, some people say I like this clothes very much, the color can make me more beautiful, some say it is so bad that when you wear it, you look older at least 10 years old, some say the material is comfortable, the size is suitable for me but in some details it is so bad, and some say when you wear it outside, you are the coolest people in the street. New customer may be confused with so

many review texts. As a result, they may lose patience to read all the reviews or even probably give up to purchase the product. However, sentiment analysis can give a direct suggestion such as positive, neutral and negative or recommend and not recommend to customers.

However, some inappropriate comments will not only decrease the true score of the product but may also mislead customers to reduce their desire to buy(shah et al.; 2018). Therefore, accurate sentiment analysis about customer reviews is particularly important.

The purpose of this paper is to find a reliable classification method of customer reviews based on online women clothing reviews by applying sentiment analysis, which can improve accuracy.

1.2 research question

Which machine learning algorithm can improve the accuracy of classifying sentiment about online women clothing reviews?

1.3 research objectives

To solve research question, four classification algorithms which were Support Vector Machine, Logistic Regression, Random Forest and Naive Bayes were selected to build the model. Implemented the algorithms and evaluated them. Compared with their accuracy and got the result.

The remainder of the paper is structured as followed: In section 2, related literature review and previous study will be discussed. Methodology and design specification are presented in section 3 and section 4 respectively. In section 5 shows how to implement the algorithms and methods. In section 6 evaluates the experimental results . Finally we make conclusions and discuss future work.

2 Related Work

With the development of the Internet, more and more people choose to shop online. After shopping, people like to share their experience of using products on the network, and then provide some suggestions for other customers. At the same time, more researchers focus on these reviews to do sentiment analysis, hoping to give more accurate classifications or predictions.In this section we will discuss previous research papers as following aspects:

2.1 Classification of customers' sentiment based on POS(part-of speech) taggers

The POS tagger is the process of matching tagged words in the corpus to specific parts of the speech based on the context. [1]

(Pankaj et al.; 2019) selected all the objective content as sentiment sentences and used POS taggers to classify word and then identify the positive or negative opinion. The data source was Amazon's online product reviews, which performed sentiment analysis through preprocessing, bias, data accuracy and other functions. Different from (Pankaj et al.; 2019) model, (Abulaish et al.; 2009) classified sentiment by linguistic and

[1] https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-marko

sentiment analysis. There were five major modules which were document processor, objectivity analyser, document parser, feature and opinion learner and review visualizer. Position different types of information inside the document using the POS taggers. Feature extraction was mainly performed through semantic and linguistic analysis of text documents. Senti-WordNet used the polarity score of opinion words to establish the polarity of sentences, and then generated characteristic documents.

2.2 Classification of customers' sentiment based on Machine Learning

Machine learning becomes more and more popular to solve problems, classifying customers' sentiment using machine learning become a hot topic, many researchers show interests in it.(Agarap and Grafilon; 2018) proposed a Recurrent Neural Network(RNN) with long-short term memory(LSTM) to research whether it recommends or not and sentiment analysis. The results indicated that F1-score can get 0.88 and 0.93 for recommendation classification and sentiment classification respectively.(Ali et al.; 2019) used deep learning method to do sentiment analysis, which combined long-short term memory(LSTM) and Convolutional Neural Network (CNN). Compared with LSTM and CNN, in order to prove the applicability of the model, the IMDB data set containing 50,000 film reviews was used, of which 50% were positive reviews and 50% were negative reviews. The hybrid model got a higher accuracy. Similarly, (Jain et al.; 2018) also used CNN and LSTM to classifiy customer reviews. However, (Jain et al.; 2018) concentrated on the advantages of using deep models for sentiment analysis in customer reviews. In addition, (Jain et al.; 2018) also studied the applicability of deep neural network strategies to extract current emotions, and used deep networks trained with weak supervision strategies to make predictions. (Lal et al.; 2018) proposed a deep learning algorithm like Autoencoder Neural Network, whose neural vector was trained to reproduce input vector as output vector. The model first trained the neural network and then fine-tunes it. One comment was selected as the query, and the other comments were ranked based on the cosine of the angle between the codes. Other test comments also followed this step. When evaluating, (Lal et al.; 2018) draw the number of retrieved comments based on the proportion of categories in the same label as the query document. Compared with (Ali et al.; 2019) and (Jain et al.; 2018), (Lal et al.; 2018) had less minimal constraints on the task for sentiment analysis. The result performed better than Naive Bayes and SVM as well. (Burns et al.; 2011) compared effect of Naive Bayes and dynamic language model on balanced and unbalanced data set. The bag of words method was used on the TV data set, and its classification result was obviously better than other classifiers. It also showed that semantics were not very important here. The result indicated both algorithms performed better on unbalanced dataset. Different from (Burns et al.; 2011), (Jagdale et al.; 2019) used Naive Bayes and SVM to classify reviews that were positive or negative. Data sets are reviews from Amazon cameras, laptops, phones, tablets, TVs, video surveillance. Both algorithms achieved good results. (Alrehili and Albalawi; 2019) also used SVM and Naive Bayes, however, (Alrehili and Albalawi; 2019) used ensemble method voting which combined Naive Bayes, Support Vector Machines (SVMs), Random forest, Bagging and Boosting. The ensemble model was implemented in six different scenarios. All experiments were completed by Weka and used 6 completely different scenario tests to evaluate the model. The final result showed that the random forest technology can be as accurate as 89.87% when using unigram. (KHAN

et al.; 2019) also proposed a framework which contained data acquisition, pre-processing, and feature extraction.SVM, NB and Decision Tree(DT) algorithms were used to classify customer sentiment. The framework contributed to researchers, service providers, and decision makers. Using the airline's data set to evaluate the framework, results displayed that the SVM accuracy was as high as 90.3%, which was significantly higher than other technologies. (Kiritchenko et al.; 2014) used Passive-Aggressive (PA) algorithm, SVM to detect the sentiment expressed in terms of terms and aspects in the customer reviews was reached. Besides, (Kiritchenko et al.; 2014) also generated a dictionary from the corpus and then calculated the emotional score of each word in each corpus.

2.3 Classification of customers' sentiment based on unsupervised learning

Unsupervised learning can solve recognition problems by training samples of unlabeled categories.[2] Clustering included.

(Bagheri et al.; 2013) proposed a novel unsupervised and domain-independent model for detecting explicit and implicit aspects for sentiment analysis. Firstly, he used heuristic rules to check the impact of opinion words on detection. Then scored aspects using a new bootstrap iterative algorithm of mutual information and aspect frequency. Next, two pruning methods were used in order to remove incorrect aspects. Finally, the implicit aspect was identified mainly by using explicit aspects and insight words. The highlight of the model was it successfully solved domain dependencies, the need for tagged data, and the main bottlenecks in hermits. Meanwhile, the result showed it can be performed efficiently especially under the circumstance of high precision. (Gamon et al.; 2005) scheduled a pulse model to mine topics and sentiment analysis from customer reviews. The main idea was to find the cluster of keywords in the sentence by clustering methods such as k-means, entropy-based and n-gram feature vectors, and collect the scores of the customer's emotions from the sentiment classifier. By using this method, customers can quickly find the information they need in a large amount of text.

2.4 Classification of customers' sentiment based on fuzzy

(Sun et al.; 2019) proposed a fuzzy product ontology mining algorithm, which explored products from a fine-grained level of online customer reviews. The novel algorithm can not only help a company improve their prodcuts but make better decisions for customers. (Yang et al.; 2018) proposed an evolutionary fuzzy deep belief networks with incremental rules (EFDBNI)algorithm based on fuzzy mathematics and genetic algorithm to figure out the problem with a small number of marker comments. The results showed that EFDBNI had a significant improvement over existing methods. This method had achieved good results in sentiment classification problems with a few labeled comments. In contrast to previous studies, the performance of existing deep learning architectures was significantly improved.

2.5 Others

(Gräbner et al.; 2012)proposed an original method of fine-grained hierarchical sentiment analysis of massive user reviews. Based on functional and contextual sentiment analysis

[2] `https://en.wikipedia.org/wiki/Unsupervised_learning`

and a large number of user reviews on the Internet, semantics were extracted from online customer reviews with positive and negative labels, and a semi-supervised fuzzy product ontology mining algorithm was implemented. Compared with the baseline method, this method had obvious performance improvement. The data set was about tourist domain. (Gräbner et al.; 2012) only used target labels such as good, neutral and bad to analyse. (shah et al.; 2018) designed a system used HTML,CSS,JAVASCRIPT,.NET FRAMEWORK and SQL to analysis customer reviews. The highlight of this system was efficiently increased the customers reliability with result and data sufficiency to improve prospects toward the product customer want to purchase. (Markus et al.; 2019) designed a probit model grounded on a Nagelkerke pseudo R-square measure to explain the overall star rating. The result showed the probit model performed better in star rating. Moreover, the model was easy to interpret and valuable for analysing customer assessments. The model effectively addressed existing methods for explaining overall star ratings that often fail to address methodological issues related to these star ratings and ignore comment text, which contains valuable information about the customer's assessment of different aspects of the rated item .

2.6 Conclusion

This section was described from five different aspects. We can see that many methods have been applied in sentiment analysis, but each method has its own advantages and disadvantages. Based on previous research and combining mechanical learning, this article wants to find a method to improve the accuracy of sentiment analysis.

3 Methodology

Probability theory and statistics are the basis of data mining. Using models to represent simple, descriptive statistics makes it easier to help people understand what they are researching.

Many procedure models such as KDD(Knowledge Discovery in Database), SEMMA and CRISP-DM(Cross-industry Standard Process For Data Mining) have already been used in data mining(Huber et al.; 2019).CRISP-DM is widely used in data mining as a standard process model.

The purpose of this research is to predict customer sentiment from reviews on women clothing e-commerce. Based on this aim, CRISP-DM is chosen to perform as a methodology for this research. There are six stages in CRISP-DM which are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment, as shown in Figure 1

Following steps will explain the six stages in CRISP-DM in detail about our current research.

3.1 Business Understanding

Business understanding is the first stage in CRISP-DM, which requires to know the requirements and ultimate destination of the project from a business perspective. As for our research, finding the best models using machine learning algorithms to predict customer sentiment from reviews in order to help company know how their customers

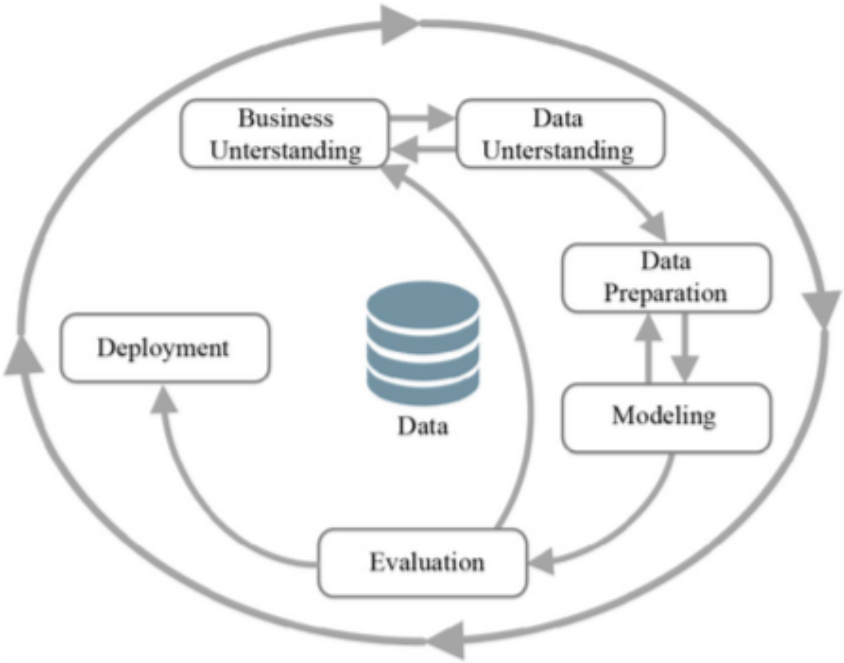


Figure 1: CRISP-DM
Huber et al. (2019)

perceive their products and how to improve their businesses. Meanwhile, customers can better know the product from previous customers.

3.2 Data Understanding

Data understanding is to select relevant data or samples from the original database and select relevant data for the target search of knowledge discovery, including the conversion of different schema data and the unification and aggregation of data. Data selection is to identify the data sets that need to be analyzed, reduce the scope of processing, and improve the quality of data mining. The data for this research can be obtained from `https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews`. It is the customers' reviews for Women's Clothing E-Commerce. Because this is a real commercial data, in the review text, retailer replaced the company's name. There are 23486 rows and 10 feature variables which are Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, Division Name, Department Name and Class Name respectively in this dataset.
Clothing ID: It refers to the clothing name and it is integer categorical.
Age: It indicates the age of reviewers' and it is also integer categorical.
Title: It shows the title of the review and it is a string variable.
Review Text: It represents what the customer write about the product and it is string variable.
Rating: It ranks for the product from 1 to 5, which 1 is the worst and 5 is the best.
Recommended IND: It is a binary variable stating and it means if the customer recommends the product where 1 is recommended, 0 is not recommended.
Positive Feedback Count: Number of other customers who affirmed this review.
Division Name: It shows the category name of the product's advanced division.

Department Name: It is the product department's name such as jackets, dresses and so on.
Class Name: It is the name of the product category such as pants, dresses and so on.

3.3 Data Preparation

Data preparation is the most important stage in data mining, which should spend amount of time and efforts on it. Good data preparation can help build good models. The original data set should be analysed and transformed to the final data set, which can meet the requirements to build models. In this research, data preparation including data exploration, missing values, remove unexpected features, data encoding and feature selection. The aim of data exploration is to help us better understand the data. The mainly method is using histograms, heat map, bar chart, etc.

Figure 2 shows the process of data preparation of this research.

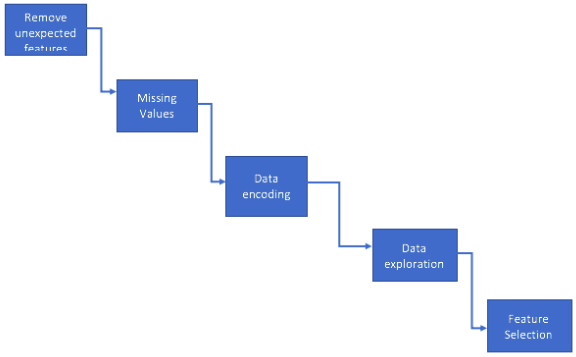


Figure 2: Process of Data Preparation

3.4 Modeling

Choosing the right model using the cleaned data and finding the meaning behind the data is the core task of the modeling phase. For the same business problem, choosing different models will have different results, and different parameters of the same model will also have different results. For our study, SVM, Logistic Regression, Random Forest, and Naive Bayes are chosen to build the model.

3.4.1 Support Vector Machine

Support Vector Machine (SVM) is a binary classification algorithm which supports linear and nonlinear classification. Its input is vector space and the output is positive or negative (0 or 1)(KHAN et al.; 2019). It belongs to supervised learning. SVM is widely used in various fields such as portrait recognition, text classification, handwritten character recognition, and bioinformatics.

3.4.2 Logistic Regression

Logistic regression is to obtain the category of the object by inputting the sequence of attribute features of the unknown category object. The result of logistic regression is a probability between 0 and 1, which is easy to use and explain. Logistic regression is widely used in data mining, automatic disease diagnosis, economic forecasting and other fields.

3.4.3 Random Forest

Random forest trains multiple trees in order to predict samples. Because of this, it has many advantages, for example, When there are many input variables, it can filter the input ability to handle these high-dimensional features without reducing the number of dimensions, it can also get good results for data sets which have missing values or default values.

3.4.4 Naive Bayes

Naive Bayes method is a classification method based on Bayes' theorem and the independent assumption of feature conditions. In Naive Bayes, there isn't any attribute variables making a large proportion of the decisions. Vice versa. It does not show too much differences for different types of data sets. It is widely used in text classification, spam classification, credit evaluation, phishing website detection and so on. Following is the equation.

$$P(c \mid f) = \frac{P(c)P(f \mid c)}{P(f)}$$

f means independent variables, c means class.

3.5 Evaluation

There are many metrics to evaluate a model, in this research, accuracy, precision, recall, F1-score, Area Under Curve (AUC) and receiver operating characteristic (ROC) are chosen to evaluate the models. High accuracy means the model can be well predicted sentiment from customer reviews online. AUC is a probability value, which between 0.1 and 1.0. It can evaluate the quality of the classifier, the higher, the better. Precision defines the proportion of rating the reviews classified correctly to all review text classified. Recall describes the proportion of rating correctly to all review text selected.

3.6 Deployment

The final step is to implement this research in real-world. The plan for data analysis implementation is applied to the business system, data and results feedback. At this

stage, it is important to know the definition of each step and maintenance of the entire process.

4 Design Specification

Figure 3 shows the work flow chart of this research.There are four stages. First of all, collecting data set from Kaggle and pre-processing data such as data exploration, missing values, special characters, data encoding and feature selection. Next, different classifications are built to train the data set. Then evaluate the result, accuracy, precision, recall, f1-score, AUC and ROC curve are selected as metrics to evalute the models. In the end, show the result with Excel in order to understand the results easily.

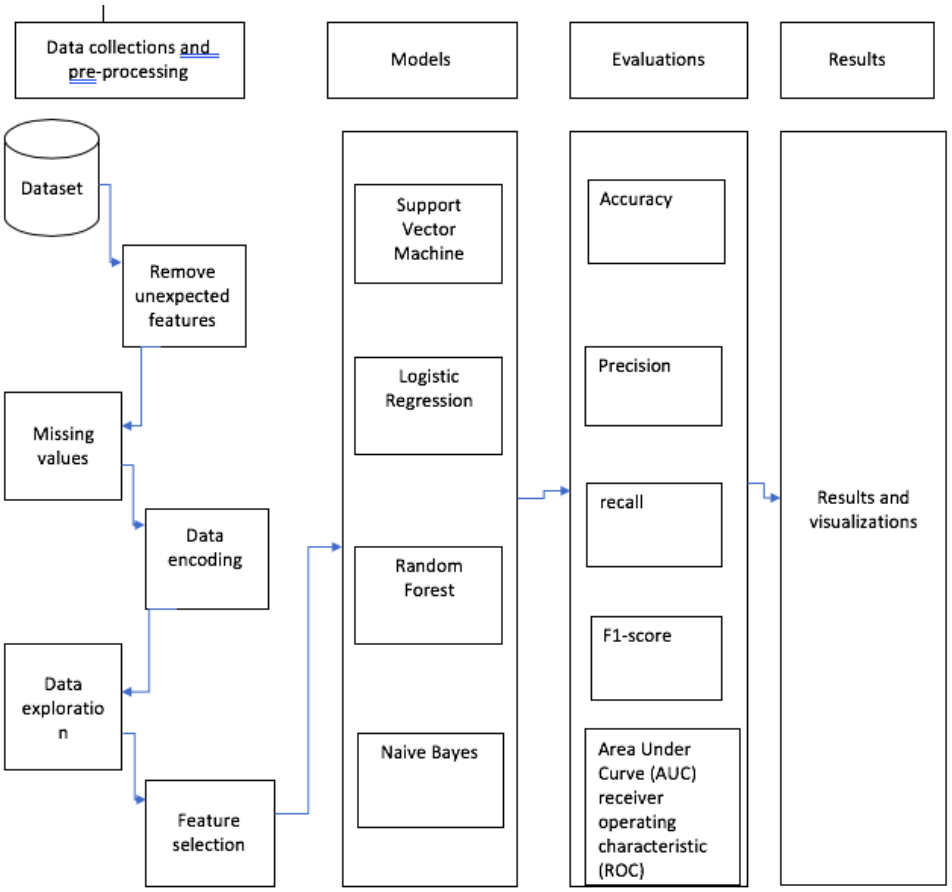


Figure 3: Work Flow Chart

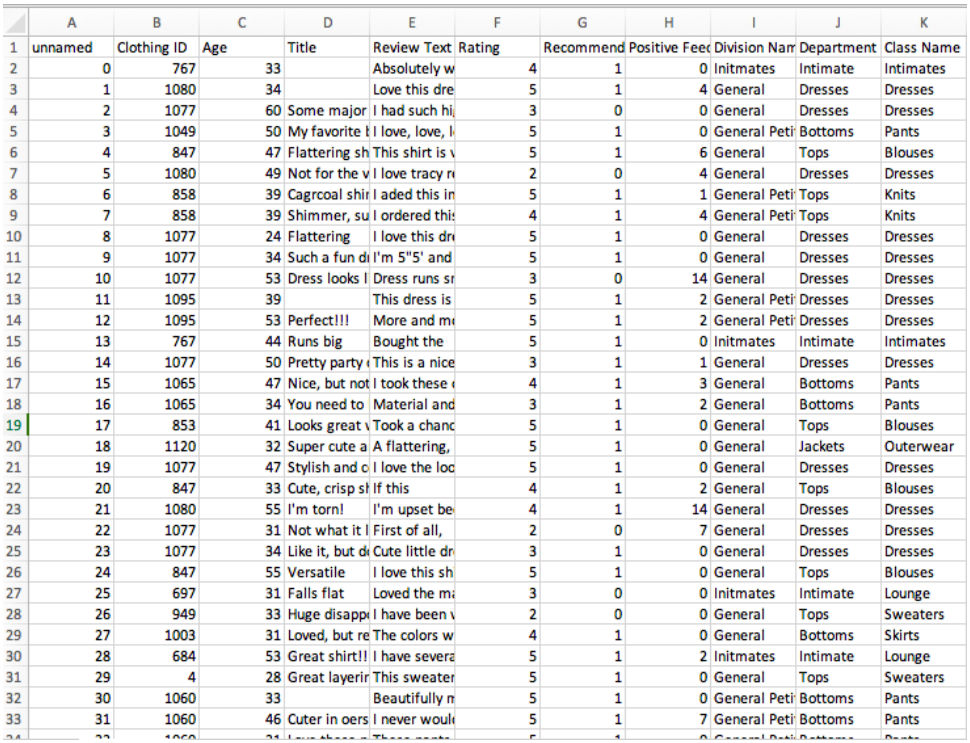
5 Implementation

5.1 Data collection and pre-processing

5.1.1 Data collection

The data in this research can be obtained from Kaggle. As explained in section 3.2. It is a csv file which includes customer reviews for Women's clothing. Figure 4 shows the raw

data



	A	B	C	D	E	F	G	H	I	J	K
1	unnamed	Clothing ID	Age	Title	Review Text	Rating	Recommend	Positive Fee	Division Nam	Department	Class Name
2	0	767	33		Absolutely w	4	1	0	Initmates	Intimate	Intimates
3	1	1080	34		Love this dre	5	1	4	General	Dresses	Dresses
4	2	1077	60	Some major	I had such hi	3	0	0	General	Dresses	Dresses
5	3	1049	50	My favorite b	I love, love, l	5	1	0	General Peti	Bottoms	Pants
6	4	847	47	Flattering sh	This shirt is v	5	1	6	General	Tops	Blouses
7	5	1080	49	Not for the v	I love tracy r	2	0	4	General	Dresses	Dresses
8	6	858	39	Cagrcoal shir	I aded this in	5	1	1	General Peti	Tops	Knits
9	7	858	39	Shimmer, su	I ordered this	4	1	4	General Peti	Tops	Knits
10	8	1077	24	Flattering	I love this dr	5	1	0	General	Dresses	Dresses
11	9	1077	34	Such a fun d	I'm 5"5' and	5	1	0	General	Dresses	Dresses
12	10	1077	53	Dress looks l	Dress runs sr	3	0	14	General	Dresses	Dresses
13	11	1095	39		This dress is	5	1	2	General Peti	Dresses	Dresses
14	12	1095	53	Perfect!!!	More and m	5	1	2	General Peti	Dresses	Dresses
15	13	767	44	Runs big	Bought the	5	1	0	Initmates	Intimate	Intimates
16	14	1077	50	Pretty party	This is a nice	3	1	1	General	Dresses	Dresses
17	15	1065	47	Nice, but not	I took these	4	1	3	General	Bottoms	Pants
18	16	1065	34	You need to	Material and	3	1	2	General	Bottoms	Pants
19	17	853	41	Looks great v	Took a chanc	5	1	0	General	Tops	Blouses
20	18	1120	32	Super cute a	A flattering,	5	1	0	General	Jackets	Outerwear
21	19	1077	47	Stylish and c	I love the loc	5	1	0	General	Dresses	Dresses
22	20	847	33	Cute, crisp sl	If this	4	1	2	General	Tops	Blouses
23	21	1080	55	I'm torn!	I'm upset be	4	1	14	General	Dresses	Dresses
24	22	1077	31	Not what it l	First of all,	2	0	7	General	Dresses	Dresses
25	23	1077	34	Like it, but d	Cute little dr	3	1	0	General	Dresses	Dresses
26	24	847	55	Versatile	I love this sh	5	1	0	General	Tops	Blouses
27	25	697	31	Falls flat	Loved the ma	3	0	0	Initmates	Intimate	Lounge
28	26	949	33	Huge disapp	I have been v	2	0	0	General	Tops	Sweaters
29	27	1003	31	Loved, but re	The colors w	4	1	0	General	Bottoms	Skirts
30	28	684	53	Great shirt!!	I have severa	5	1	2	Initmates	Intimate	Lounge
31	29	4	28	Great layerir	This sweater	5	1	0	General	Tops	Sweaters
32	30	1060	33		Beautifully m	5	1	0	General Peti	Bottoms	Pants
33	31	1060	46	Cuter in oers	I never woul	5	1	7	General Peti	Bottoms	Pants

Figure 4: Raw data

5.1.2 Remove Unexpected features

Figure 5 provides the raw data contained unexpected features such as column unnamed.

	unnamed	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

Figure 5: Raw data including unexpected features

Likewise, special characters such as \n , stopwords, numeric, punctuation, numbers and returns list of words in the review texts should be removed.

5.1.3 Missing Values

Missing values can have an influence on the performance of the model. Before building the model, checking missing values is important. Missing values such as NaN can be replaced by 0 or delete the missing values. Figure 6 shows the missing values in this data set. Title and Review Text have more missing values. Deleting missing observations for

Class Name, Division Name and Department Name variables because they just have 14 missing values.

```
Clothing ID                 0
Age                         0
Title                    3810
Review Text               845
Rating                      0
Recommended IND             0
Positive Feedback Count     0
Division Name              14
Department Name            14
Class Name                 14
dtype: int64
```

Figure 6: Missing Values

Meanwhile, data type also should be checked. Figure 7 indicates the data type of each feature. There are 5 features belonging to int and 5 features belonging to the object. Changing Review Text variables into string in order to prepare for data exploration.

5.1.4 Data Encoding

The aim of data encoding is to quantify variables that cannot be quantified.

In the data set, Recommend and not Recommend cannot be recognised by models, it has to be encoded. Column Recommend IND 0 represents not recommend, 1 represents recommend. In this research, sentiment refers to the rating. Rating of 4 or higher means positive. Rating of 2 or lower represents negative and rating of 3 is neutral. Encoding Class Name, Division Name and Department Name in order to prepare for the data exploration.

5.1.5 Data Exploration

All the pre-processing and data exploration stage were implemented in ANACONDA NAVIGATOR jupyter notebook 6.0.0 which is an edition people can write code. Pandas, numpy, matplot.lib, seaborn, wordcloud packages were installed and performed. There are many functions and methods in Pandas which can help us analyze data set.[3] Numpy is an open source numerical computing extension of Python, which can be used to store and process large matrices.[4] Matplot.lib is used to draw visualizations such as plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc.[5]Seaborn is a graphical

[3] https://pandas.pydata.org
[4] https://numpy.org
[5] https://matplotlib.org

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22628 entries, 0 to 23485
Data columns (total 10 columns):
Clothing ID                 22628 non-null int64
Age                         22628 non-null int64
Title                       19662 non-null object
Review Text                 22628 non-null object
Rating                      22628 non-null int64
Recommended IND             22628 non-null int64
Positive Feedback Count     22628 non-null int64
Division Name               22628 non-null object
Department Name             22628 non-null object
Class Name                  22628 non-null object
dtypes: int64(5), object(5)
memory usage: 1.9+ MB
```

Figure 7: Type of data

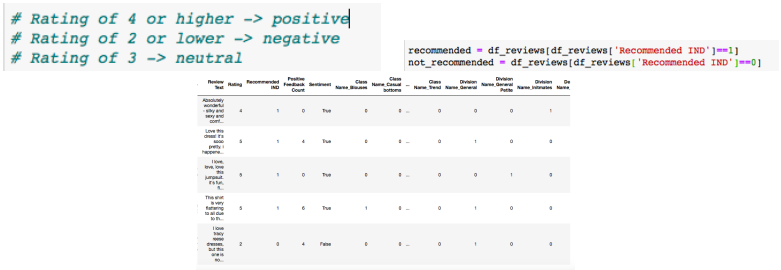


Figure 8: Data Encoding

visualization python package based on matplotlib.[6]

Data Exploration: To have a better understanding of the data set, data exploration has been undertaken. Because in the review text, they are all string. To explore this, WordCloud was performed. We add word counts to the dataframe in order to use these counts to reach some useful information.

It can be observed from the Figure 9 that the most common words were dress, love, size, top, fit, like, wear, great, would and fabric.

Figure 10 showed the data exploration results, it was easy to find that age from 25-56 is likely to review online, especially people in 39, the number of reviews in 39 years old were more than 1200. General gained the most reviews in division part and Initmates gained the least reviews. The amount of Tops reviews was the highest in Department, which more than 10000. Compared with Tops, the number of Trend reviews was the lowest, which nearly 200. From Class Name prospective, the amount of Dresses was the highest, which was twice than Blouses. Casual bottoms and Chemises were the lowest. In the Initmates division, most of the people choose to recommend, only a few not recommend. Compared with Initmates, General division had the opposite trends, the amount of recommend and not recommend almost the same. The number of recommend item in Department was higher than not recommend except Dresses. Recommend in

[6] `https://seaborn.pydata.org`

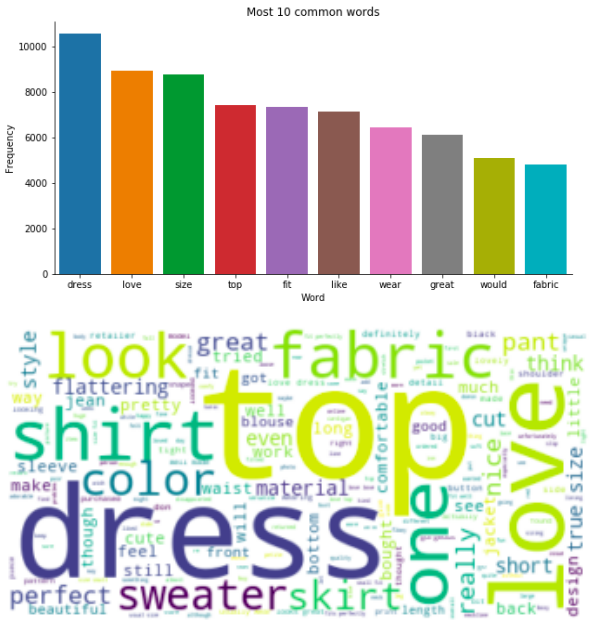


Figure 9: Wordcloud

each Class had the same trend with recommend in Department. Only the amount of not recommend Jeans was higher than recommend. From the bar chart, it was easy to observe that item 1078 gained the highest popularity.

5.1.6 Feature Selection

There are many features in the data set, however, not all the features can be useful to build the model. Therefore, feature selection can choose the most important features and merge some similar features. In our research, Pearson correlation coefficient and heat map contributes to feature selection.

Figure 11 showed the correlation between the variables. From the maps, it suggested a strong correlation between Recommended IND and Rating. It seemed there wasn't correlation between Positive Feedback Count and Rating and Recommended IND.

5.2 Models

Through data exploration and heat map we can defined the features we would use in the models were Rating, Review Text, Class Name, Age and Sentiment. The new data set was chosen 80% as train data and the other 20% as test data. Sklearn.model_selection import train_test_split was applied for splitting data set.

5.2.1 Support Vector Machine

The main idea of Support Vector Machine(SVM) is to look for a hyperplane in space that can divide the data set into different categories, and distance between all the data and hyperplane should be the shortest. For this research, sklearn SVC function was selected to build SVM. Kernel, class_weight, probability, random_state are parameters of this model. Kernel chose linear, class_weight chose balanced, probability was true and

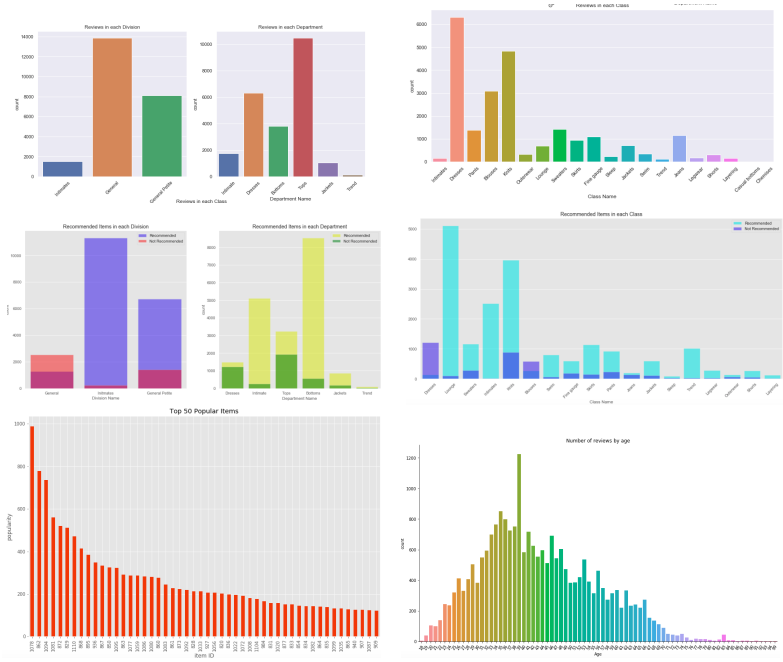


Figure 10: Data Exploration

random_state was 111. Figure 12 shows the result of the SVM. The accuracy of SVM was 91%.

5.2.2 Logistic Regression

The logistic regression model is a classification model which expressed in the form of conditional probability distribution P(Y/X). For our research, sklearn package can help us make the logistic regression model. Figure 13 shows the result of the Logistic Regression. The accuracy of Logistic Regression was 91%.

5.2.3 Random Forest

Random forest can avoid over-fitting because it can use the bag method to generate multiple training sets, and use each training set to construct the tree.

For this research, n_estimators were 1000, the max depth of the decision was 5, Figure 14 shows the result of the Random Forest. The accuracy of Random Forest was 87%.

5.2.4 Naive Bayes

Naive Bayes algorithm is based on Bayes' theorem,whose characteristic condition is independent. It is assumed that the n features of X are conditionally independent under the conditions determined by the class. Figure 15 shows the result of the Naive Bayes. The accuracy reached 93%.

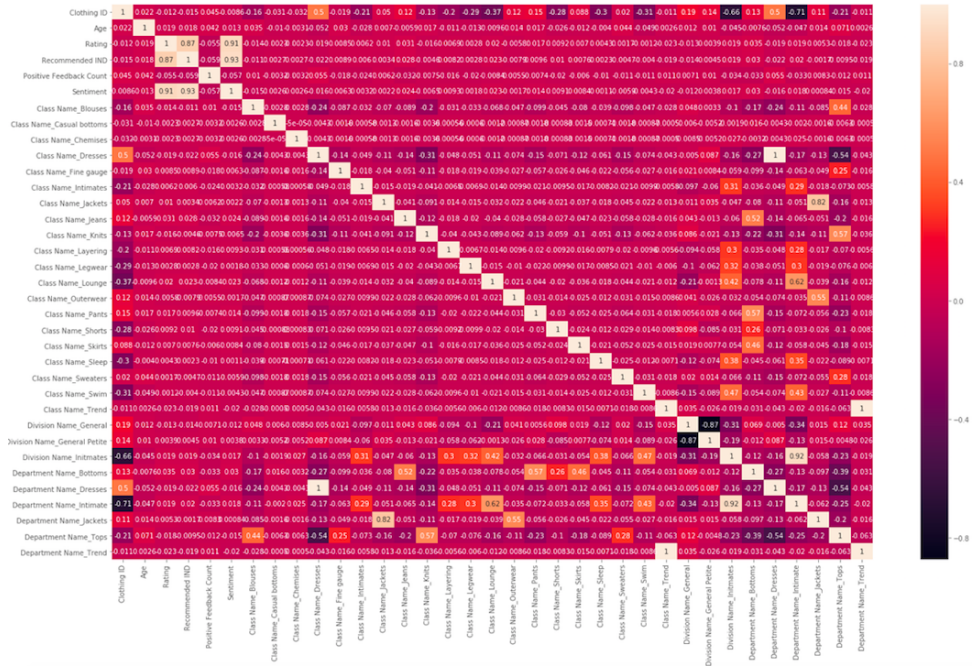


Figure 11: Heat map

```
# svm
from sklearn.svm import SVC
start=dt.datetime.now()
svm = SVC(C=1.0,
          kernel='linear',
          class_weight='balanced',
          probability=True,
          random_state=111)
svm.fit(X_train,y_train)
```

```
SVC(C=1.0, cache_size=200, class_weight='balanced', coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
    kernel='linear', max_iter=-1, probability=True, random_state=111,
    shrinking=True, tol=0.001, verbose=False)
```

```
# evaluate the model
from sklearn.svm import SVC
import re
test_predictions = svm.predict(X_test)
print(classification_report(y_test, test_predictions, svm.classes_ ))
```

```
              precision    recall  f1-score   support

       False       0.61      0.69      0.65       450
        True       0.96      0.94      0.95      3511

    accuracy                           0.91      3961
```

Figure 12: the result of SVM

```
# logistic regression
from sklearn.linear_model import LogisticRegression
start=dt.datetime.now()
lr = LogisticRegression(class_weight='balanced',
                            random_state=111,
                            solver='lbfgs',
                            C=1.0)
lr.fit(X_train,y_train)
```

```
LogisticRegression(C=1.0, class_weight='balanced', dual=False,
                   fit_intercept=True, intercept_scaling=1, l1_ratio=None,
                   max_iter=100, multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=111, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
from sklearn.linear_model import LogisticRegression
import re
test_predictions = lr.predict(X_test)
print(classification_report(y_test, test_predictions, lr.classes_ ))
```

```
              precision    recall  f1-score   support

       False       0.59      0.78      0.67       450
        True       0.97      0.93      0.95      3511

    accuracy                           0.91      3961
```

Figure 13: the result of Logistic Regression

```
# random forest
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(n_estimators=1000, max_depth=5,
                                  class_weight='balanced', random_state=3)
rf_model.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight='balanced',
                       criterion='gini', max_depth=5, max_features='auto',
                       max_leaf_nodes=None, min_impurity_decrease=0.0,
                       min_impurity_split=None, min_samples_leaf=1,
                       min_samples_split=2, min_weight_fraction_leaf=0.0,
                       n_estimators=1000, n_jobs=None, oob_score=False,
                       random_state=3, verbose=0, warm_start=False)
```

```
from sklearn.ensemble import RandomForestClassifier
import re
test_predictions = rf_model.predict(X_test)
print(classification_report(y_test, test_predictions, rf_model.classes_ ))
```

```
              precision    recall  f1-score   support

       False       0.47      0.84      0.60       450
        True       0.98      0.88      0.93      3511

    accuracy                           0.87      3961
```

Figure 14: the result of Random Forest

```
# Naive Bayes
from sklearn.naive_bayes import MultinomialNB
start=dt.datetime.now()
nb = MultinomialNB()
nb.fit(X_train,y_train)
```

```
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

```
from sklearn.naive_bayes import MultinomialNB
import re
test_predictions = nb.predict(X_test)
print(classification_report(y_test, test_predictions, nb.classes_ ))
```

```
              precision    recall  f1-score   support

       False       0.73      0.64      0.68       450
        True       0.95      0.97      0.96      3511

    accuracy                           0.93      3961
```

Figure 15: the result of Naive Bayes

6 Evaluation

The performance of algorithms will be evaluated by accuracy, precision, recall, F1-score and AUC. In Table 1, it showed the Confusion Matrix.

Table 1: Confusion Matrix.

Algorithm	**Predicted Positive**	**Predicted Negative**
Actual Positive	True Positive(TP)	False Positive(FP)
Acutual Negative	False Negative(FN)	True Negative(TN)

In this report, Accuracy means the proportion of correctly predicted sentiment to the total number of predicted sentiment.

Accuracy= (TN+TP)/(TN+TP+FN+FP)

Precision means the the proportion of positive sentiment were identified correctly.

Precision=TP/(TP+FP)

Recall means the proportion of actual positives were correctly identified.

Recall=TP/(TP+FN)

F1-score is based on Recall and Precision.

F1-score=(2*Precision*Recall)/(Precision+Recall)

ROC is a curve, which horizontal axis represents FPR (False Positive Rate) - the probability of erroneously predicted as a positive example, and the vertical axis represents TPR (True Positive Rate) - the probability of correctly predicting as a positive example. Following are the formulas of FPR and TPR.

FPR=FP/(FP+TN)

TPR=TP/(TP+FN)

AUC (Area Under Curve) is a numerical value. It can be intuitively seen from the meaning of its representation. The ROC-AUC represents the area enclosed by the ROC curve and the coordinate axis. Apparently, the larger the TPR and the smaller the FPR, the better the model effect. Therefore, If we want the model to work better, the ROC

curve should be closer to the upper left corner. In Table 2, it showed the algorithms, Accuracy, Precision, Recall, F1-score and AUC.

Table 2: Algorithm and metrics.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC**
Support Vector Machine	0.91	0.92	0.91	0.92	0.92
Logistic Regression	0.91	0.93	0.91	0.92	0.95
Random Forest	0.87	0.92	0.87	0.89	0.94
Naive Bayes	0.93	0.93	0.93	0.93	0.95

6.1 Experiment / Support Vector Machine

From Table 2, it was obvious that the SVM model was not too bad and not too good. Its accuracy was 91% and the precision was 92%. AUC and F1-score also had the same value 92%. The value of Recall was 91%.

6.2 Experiment / Logistic Regression

Compared with SVM, the accuracy, Recall and F1-score of logistic regression was the same as SVM, which was 91% ,91% and 92% respectively. However, Logistic regression had the highest precision of four algorithms, which was 93%. The AUC was 95%, which means Logistic regression classifier was better than SVM classifier.

6.3 Experiment / Random Forest

Random Forest had the lowest accuracy in the four classifiers, whose value was 87%. Likewise, recall and F1-score were the lowest of four classifiers, which was 87% and 89% respectively. Random Forest was a little lower than Logistic Regression in AUC, which was 94%.

6.4 Experiment / Naive Bayes

Compared with other classifiers, Naive Bayes seems better. It had the highest accuracy, AUC, recall and F1-score, whose value was 93%,95%,93% and 93% respectively. Its precision had the same value as Logistic Regression, which was 93%.

6.5 Discussion

Experimental results showed the that Naive Bayes classifier had the best accuracy, precision, recall, F1-score and AUC level as compared to Support Vector Machine, Logistic Regression and Random Forest. Both Logistic Regression and Naive Bayes had the highest precision, however, Random Forest had the lowest accuracy, recall and F1-score.

Through the above research and analysis, we got the satisfied results and we can find that the NB method can improve the accuracy of customer sentiment classification in this report. In reality, for the customer, they can clearly know whether the product is what they want from the classification results; for the company, the company can clearly know

the demand for those products from the results, those Products are customer-rejected, and the company can better reposition the product market to meet the requirements of customers and help the company make better profits.

Different from the previous literature, in this experimental report, we identified several features that contributed more to modeling by exploring the data, instead of using all features to model as (Pankaj et al.; 2019). Moreover, this research used the same data set as (Agarap and Grafilon; 2018). Both (Agarap and Grafilon; 2018) and our research, We can see that the model had a relatively stronger predictive performance for the positive sentiments(True represents positive, False represents negative).Compared with (Agarap and Grafilon; 2018), the algorithms we used in our research can save amounts of time.

7 Conclusion and Future Work

This research used four machine learning algorithms: Support Vector Machine, Logistic Regression, Random Forest and Naive Bayes to classify customer review texts. We concentrated on online women clothing reviews features such as rating, class name, age and review texts. Moreover, we compared our results with previous research and our results indicated that Naive Bayes was the preferred classifier.

Previous studies such as (Agarap and Grafilon; 2018) used Bidirectional Recurrent Neural Network to do sentiment analysis, our research used four machine learning algorithms and was able to achieve better result, achieving more than 90% accuracy for all the algorithms. Both (Agarap and Grafilon; 2018) and our study used the same data set, however, when explored data we used different methods. (Agarap and Grafilon; 2018) used NLTK to do sentiment analysis. Compared with (Agarap and Grafilon; 2018),we used a heat map and data exploration to do feature selection and then built models.

In the future, more factors such as region, occupation, salary would be considered to better classify customers' sentiment. Besides, the data set is unbalanced data set, there are more True than False. If it is possible, we should focus on unbalanced and balanced data set to explore the sentiment analysis.

References

Abulaish, M., Jahiruddin, Doja, M. N. and Ahmad, T. (2009). Feature and opinion mining for customer review summarization, *PReMI 2009: Pattern Recognition and Machine Intelligence* pp. 219–224.

Agarap, A. F. and Grafilon, P. M. (2018). Statistical analysis on e-commerce reviews, with sentiment classification using bidirectional recurrent neural network, *https://www.researchgate.net/publication/323545316* .

Ali, N. M., Hamid, M. M. A. E. and Youssif, A. (2019). Sentiment analysis for movies reviews dataset using deep learning models, *International Journal of Data Mining Knowledge Management Process (IJDKP)* **9**(2/3): 19–27.

Alrehili, A. and Albalawi, K. (2019). Sentiment analysis of customer reviews using ensemble method, *2019 International Conference on Computer and Information Sciences (ICCIS)* .

Bagheri, A., Saraee, M. and de Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews, *Knowledge-Based Systems* .

Burns, N., Bi, Y., Wang, H. and Anderson, T. (2011). Sentiment analysis of customer reviews: Balanced versus unbalanced datasets, *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems,Springer-Verlag Berlin Heidelberg 2011* **6881**: 161–170.

Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. (2005). Pulse: Mining customer opinions from free text, *International Symposium on Intelligent Data Analysis IDA 2005: Advances in Intelligent Data Analysis VI* pp. 121–132.

Gräbner, D., Zanker, M., Fliedl, G. and Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis, *Information and Communication Technologies in Tourism 2012* pp. 460–470.

Huber, S., Wiemer, H., Schneider, D. and Ihlenfeldt, S. (2019). Dmme: Data mining methodology for engineering applications – a holistic extension to the crisp-dm model, *12th CIRP Conference on Intelligent Computation in Manufacturing Engineering* **79**: 403–408.

Jagdale, R. S., Shirsat, V. S. and Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques, *Cognitive Informatics and Soft Computing, Advances in Intelligent Systems and Computing 768* pp. 639–647.

Jain, V. K., Kumar, S. and Mahanti, P. (2018). Sentiment recognition in customer reviews using deep learning, *International Journal of Enterprise Information Systems* **14**: 77–78.

KHAN, D. M., Rao, T. A. and Shahzad, F. (2019). The classification of customers' sentiment using data mining approaches, *Global Social Sciences Review (GSSR)* **IV**(IV): 198–212.

Kiritchenko, S., Zhu, X., Cherry, C. and Mohammad, S. M. (2014). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* pp. 437–442.

Lal, M., Jain, A. and Avatade, M. (2018). Sentiment analysis on customer reviews using deep learning, *International Journal of Computer Sciences and Engineering* **6**: 1023–1024.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis, *Springer Science+Business Media, LLC 2012* pp. 415–463.

Markus, B., Bernd, H., Mathias, K., Andreas, O. and Alexander, S. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews, *Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden.* .

Pankaj, Pandey, P., Muskan and Soni, N. (2019). Sentiment analysis on customer feedback data: Amazon product reviews, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* pp. 320–322.

shah, A., karan shah, hridesh shah and dhruv shah. (2018). Research article data analysis on customer review., *International Journal of advanced research(IJAR)* **6**(10): 1487–1492.

Sun, Q., Niu, J., Yao, Z. and Yan, H. (2019). Exploring ewom in online customer reviews: Sentiment analysis at a fine-grained level, *Engineering Applications of Artificial Intelligence* **81**: 68–78.

Yang, P., Wang, D., Du, X.-L. and Wang, M. (2018). Evolutionary dbn for the customers' sentiment classification with incremental rules, *Industrial Conference on Data Mining ICDM 2018: Advances in Data Mining. Applications and Theoretical Aspects* pp. 119–134.