# AMAZON USER REVIEW ANALYSIS BY NLP


## ABSTRACT

Opinion information is very important for businesses and manufacturers. They often want to know in time what consumers and the public think of their products and services. However, it is not realistic to manually read every post on the website and extract useful viewpoint information from it. If you do it manually, there is too much data. Sentiment analysis allows large-scale processing of data in an efficient and cost-effective manner. In order to know more about sentiment analysis, this author explores the application of sentiment analysis on business to understand its strengths and limitations. This paper used dataset of the Amazon Alexa reviews, and then built a model to predict the sentiment of the comment given the comment declaration by using Python and machine leaning algorithm- Naïve Bayes and logistic regression. SMOTE is used to cope with the unbalanced dataset and AUC/ROC are used to evaluate which method is best.

# 1. INTRODUCTION

Today, individuals, businesses, institutions, and governments are increasingly using information from social media to inform their decisions (Liu, 2012). When a person wants to buy a product, he no longer only asks people around him, but gets a lot of comments, discussions and other information about the product from the Internet. For an organization, the evaluation of its products and services can also be obtained through the Internet. Similarly, it is easy for governments to get public feedback on their policies, as well as to learn about important events in other countries. However, due to the proliferation of various websites, online discovery and monitoring of opinion websites and extracting information from them is still a difficult task. The average human reader will have a hard time identifying relevant sites, extracting and summarizing comments from them. Therefore, an automated emotional analysis system is needed (Liu, 2012).

This paper selects dataset “Amazon Alexa Reviews” from Kaggle (Siddhartha, 2018) and would like to do a sentimental analysis about the Amazon Alexa reviews by Python. The author will detect the positive and negative feedback and visualize the contents by word cloud and predict the ratings from the reviews by Naïve Bayes and Logistic regression model, and then compare the accuracy of different machine learning algorithms.

Speech is the most natural form of human interaction. After the invention of computer, it has become a goal for people to enable machines to "understand" human language. Alexa is Amazon’s voice control system, which allow people to say commands and then have them fulfilled; Alexa can be linked up with lights, curtains, sound system, television, kindle device and more(Komak, 2017). Below are the electronic devices with Alexa.

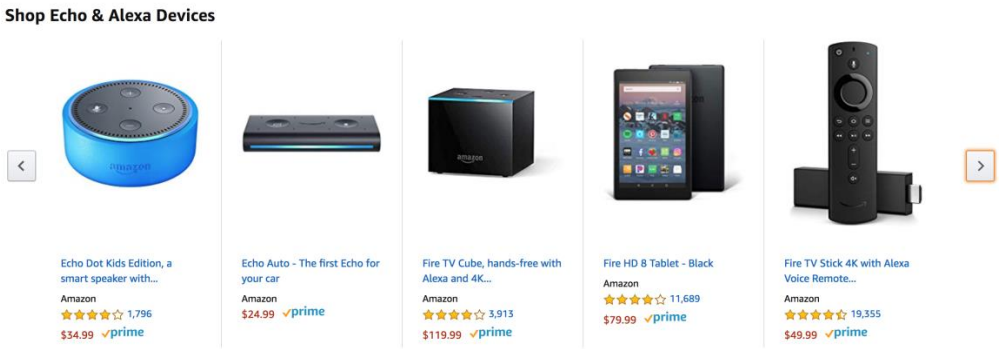



**Figure1: Shop Echo & Alexa Devices (Amazon.com ,2019)**

# 2. RELEVANT RESEARCH

Sentiment analysis, also known as opinion mining. The goal of this research field is to analyse people's views, feelings, evaluations, attitudes and emotions towards entities and their attributes from the text, and these entities can be a variety of products, services, institutions, individuals, events, issues, or topics(Liu, 2012). As publicly available information on the Internet continues to grow, there are a large number of opinions online. With the help of sentiment analysis system, this unstructured information can be automatically translated into structured data about products, services, brands, politics, or other topics that people can express(Liu, 2012). This data is useful for marketing analysis, public relations, product reviews, customer service, etc. The classification of emotional polarity has three levels: document level, sentence level and entity and aspect level. The document level focuses on whether a document as a whole expresses negative or positive emotions; while the sentence level focuses on the emotional classification of each sentence; The entity level and the aspect level are the views that people like or don't like about them (Liu,2012; Fang and Zhan, 2015).

There are many types of sentiment analysis, ranging from systems that focus on polarity (positive, negative, neutral) to systems that detect emotions (anger, happiness, sadness, etc.) or identify intent (interested and not interested) (Monkeylearn, n.d.). However, the quality of some reviews online cannot be guaranteed for spam are not relevant to the topic or meaningless, which may impede the sentiment analysis (Fang and Zhan, 2015). Meanwhile, the opinions are subjective, so it is important to collect the opinions from many people than just only a few people(Liu, 2012).

In recent years, most of the work of sentiment analysis is to develop more accurate classification algorithms and constantly solve some major challenges and limitations in this field. Natural language processing (NLP) had done little research on sentiments until the year 2000, when we human were getting to have a huge volume of opining text online(Liu, 2012). Natural language processing (NLP) provides the necessary technology for text mining to automatically extract knowledge from these texts(Liddy, 2000). The goal of NLP is to develop a language that allows computers to understand unstructured text and help them understand the language. However, sentiment analysis is a highly restricted NLP problem because NLP can only understand some aspects of it, such as positive or negative emotions(Liu, 2012). NLTK is the natural language processing toolkit for Python, one of the most commonly used Python libraries in the NLP world. This paper will use NLTK to do the research.

# 3. DATA UNDERSTANDING AND EXPLORATION

## Data understanding

This dataset has 3150 Amazon alexa customer comments (text) on amazon Alexa products in about 3 months. Here is the header of the dataset. There are 5 columns: “rating” column is the comment scores; “date” is the comment date; the “variation” column is various Alexa products like Charcoal Fabric, Alexa Echo, Echo dots, Alexa Fire sticks etc.; in “feedback” column, if the comment score is greater than or equal to 3 it's one, and if it's less than three it's zero.

Table1: The header of the dataset

| | rating | date | variation | verified_reviews | feedback |
|---|---|---|---|---|---|
| 0 | 5 | 31-Jul-18 | Charcoal Fabric | Love my Echo! | 1 |
| 1 | 5 | 31-Jul-18 | Charcoal Fabric | Loved it! | 1 |
| 2 | 4 | 31-Jul-18 | Walnut Finish | Sometimes while playing a game, you can answer... | 1 |
| 3 | 5 | 31-Jul-18 | Charcoal Fabric | I have had a lot of fun with this thing. My 4 ... | 1 |
| 4 | 5 | 31-Jul-18 | Charcoal Fabric | Music | 1 |

Before diving into data analysis, it is necessary to take a look at the dataset and clean data if necessary. For sentimental analysis, it needs more steps to prepare the text for later on analysis. Here is the workflow of data analysis by Python in this project.

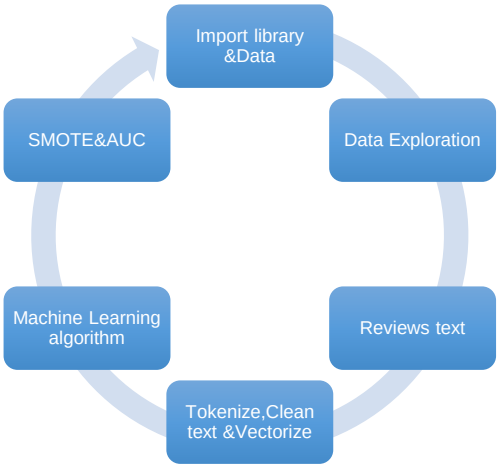



**Figure2: The workflow of the project**

## Data exploration

After importing the library and data, we can do some data exploration to generally understanding the dataset. Here is the describe of the reviews (figure3). We can see that the average rating is about

4.46 with standard variance 1.06, and the feedback is about 0.92 with standard variance 0.27, which means these products are highly rated by most of the customers. We can know more of the detail by plotting the distribution of the rating(figure4). More than 2100 reviews are rated 5 scores and 500 reviews rated 4 scores.

| | rating | feedback |
|---|---|---|
| count | 3150.000000 | 3150.000000 |
| mean | 4.463175 | 0.918413 |
| std | 1.068506 | 0.273778 |
| min | 1.000000 | 0.000000 |
| 25% | 4.000000 | 1.000000 |
| 50% | 5.000000 | 1.000000 |
| 75% | 5.000000 | 1.000000 |
| max | 5.000000 | 1.000000 |

**Figure3: Describe of the reviews**

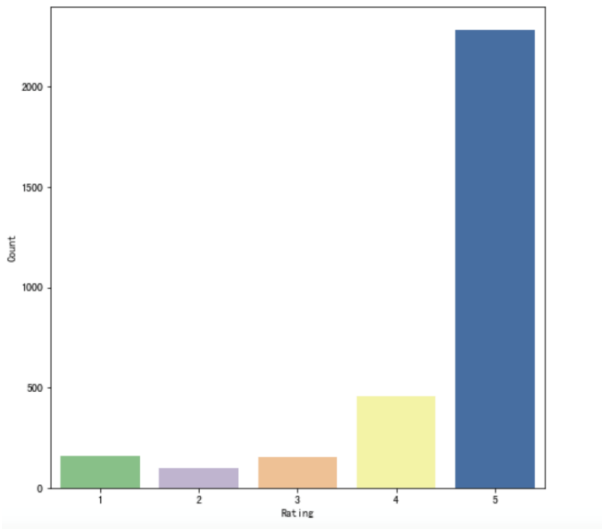



**Figure4: Distribution of the rating**

Because there are various Alexa products, let's check the average rating by variation. The picture below shows that the average rating based on Alexa variation are more than 4 for each devices, which also prove that most of the customers observed were fond of the Alexa products.

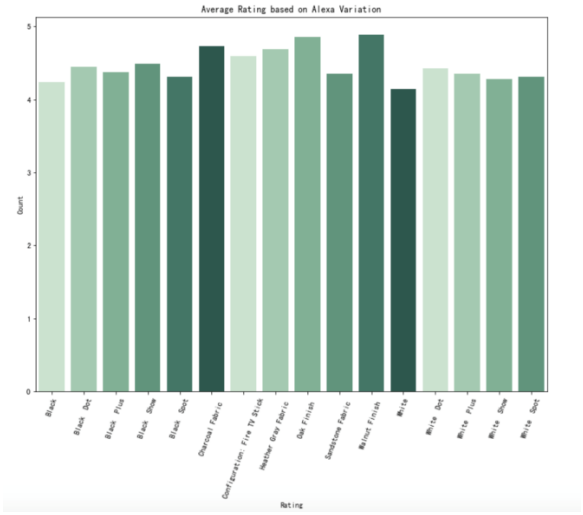



**Figure5: The average rating by variation**

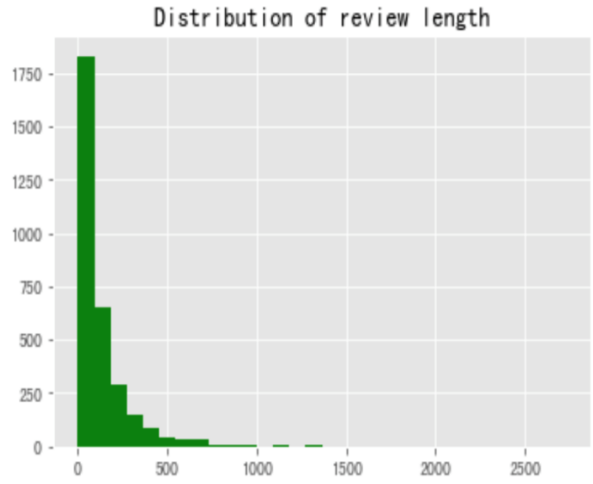
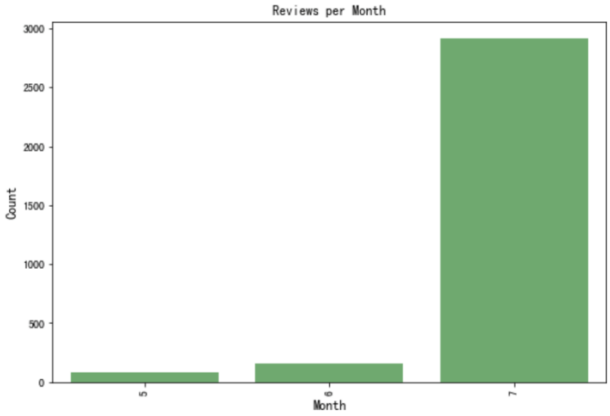



**Figure6: The distribution of review length**

## Review text

Let's figure out the distribution of the review length in figure6. Most of the reviews length are under 500 words. The mean length of the reviews is about 128 words and the maximum is 2730 words (by Python). Next, we can take a look at the comment date of the reviews and the average ratings on each weekdays.



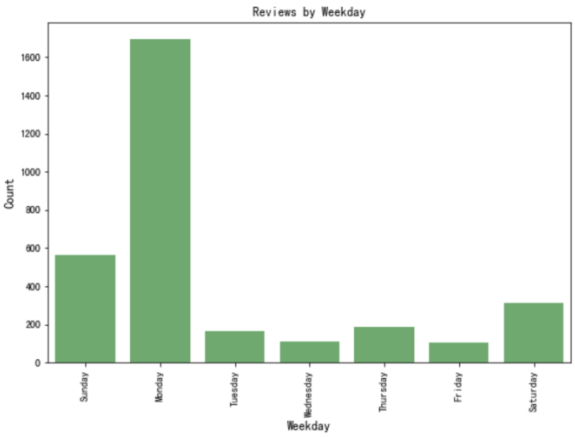




**Figure7: Reviews per month** **Figure8: Average rating over the day**

From the figure 7 and 8, we can see that most of the views are from July and the average ratings did not vary with the weekdays, which means the reviews collected did not affected by the time.

As the final exploration step, we will visualize the most frequently words in the reviews by word cloud.

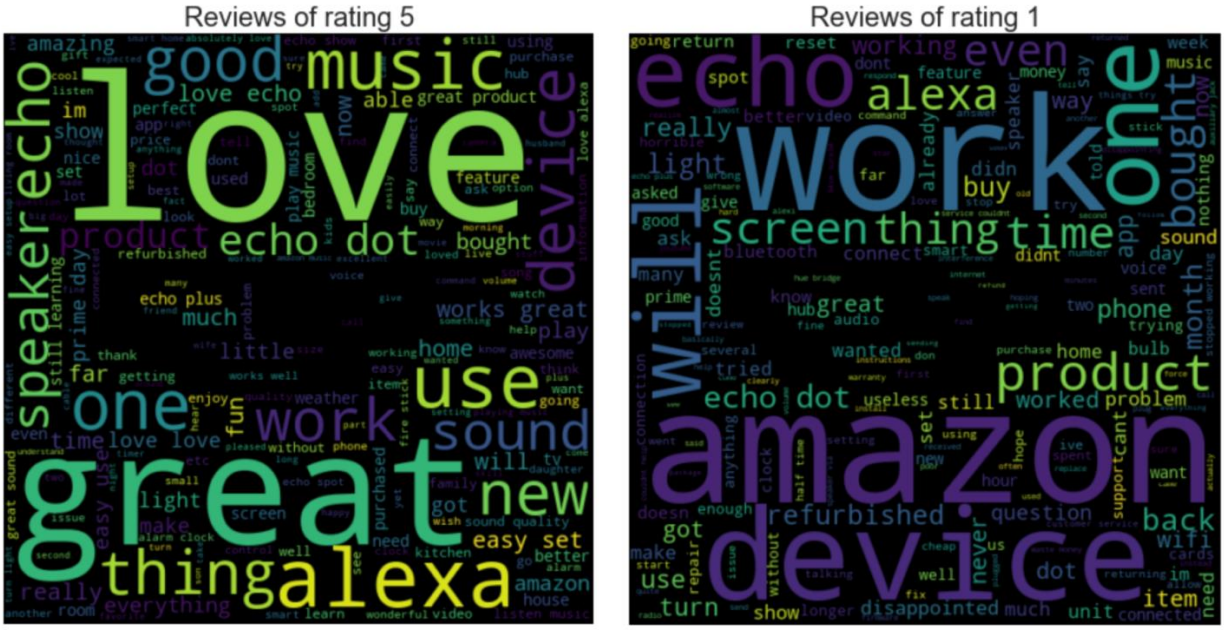



**Figure9: Word cloud of reviews of rating 5 and rating 1**

This paper would like to extract the reviews with rating 5 points and with 1 point, so that we can look at the best the worst reviews at the same time. From figure9, it can be seen that the outstanding words like “love”, “echo”, “great”, “alexa”, “music”, “like”, “use”, “sound”, “new” “easy” and so on are in the rating 5. For the reviews of rating 1, we can see the words “amazon”, “device”, “work”, “disappointed”, “echo”. It says the people who gave rating 1 will describe the facts and the feelings together. And the world could be just a big picture of the words, we need to do more work to find more valuable things.

# 4. MODELLING AND DATA ANALYSIS

In this part, the NLTK library will be used to analyse the sentiment analysis. In order to getting the bag of words, which is a way of extracting features from text for modelling(Brownlee, 2017), tokenisation, stemming and removing the punctuation, and vectorization the text are required.

## Tokenization

Tokenization is to separate word-like units from text(Grefenstette and Tapanainen, 1994), which is also known as word segmentation(Palmer, n.d., p. 2). When doing text mining, the first preprocessing is word segmentation. English words are naturally separated by Spaces and can be easily separated by Spaces, but sometimes multiple words need to be treated as one word. For example, " New Jersey ", need to be looked as one word. In Chinese, participles are a special problem to be solved because there is no space. Whether in English or Chinese, the principle of word segmentation is similar. Modern participles are all based on statistical participles, and statistical sample content comes from some standard corpus (Liu, 2017a).

## Stemming

For some changing words, we have to deal with different forms of the same word, and make the computer understand that these different words have relative forms. For example, the word “sing” can come in many forms as “sang”, “singer”, “singing”, which means the same thing. Humans can easily identify the basic and derivative forms of these words. When analysing text, it's useful to extract these basic forms, which will let us extract useful statistics to analyse the input text (Nie, 2018).

## Punctuation

There are many meaningless words in the English text, such as "the", "in", some short words, and some punctuation marks, which we need to remove them. These invalid words are stop words(Liu, 2017b).

## Vectorization

In the step of counting the word frequency, we will get the word frequency of all the words in the text. With the word frequency, we can use the word vector to represent the text(Liu, 2017b). The vectorization method is easy to use and straightforward, but it is difficult to use in some scenarios such as the vocabulary after word segmentation is very large. Hash Trick is a commonly used method to reduce the dimension of text features.

## Applying the machine learning model to predict ratings

With the feature vector of the reviews, we can use the dataset to establish machine leaning model to predict the rating. The dataset provided the feedback and ratings for each reviews, which can be used to create a semimetal score: positive or negative. In this paper, the ratings 4 or above is used for positive and ratings less than 4 for negative. Then we need to separate the dataset into two parts: training and test dataset. This paper split the data 70% for training 30% for testing. The training dataset is the sample dataset for learning, while the test dataset is the sample dataset for performance evaluation (“Training, validation, and test sets,” 2019).

This paper used two machine learning algorithm: Naive Bayes and Logistic Regression. Because the naive Bayes classifier assumes that all attributes are independent, that is why the so-called “Naive” and it has been successfully applied to document classification in many research efforts (McCallum and Nigam, 1998). Logistic regression is a algorithm which can deal with binary classification (Hosmer and Lemesbow, n.d.; Liu, 2017c). The training speed of Logistic regression is faster than support vector machine (SVM), and it is enough to solve common classification problems (Liu, 2017c).

# 5. RESULTS AND EVALUATION

After applying the Naive Bayes and Logistic Regression, we can get the results of each model. The accuracy of the Naive Bayes model is 87.1%, and the confusion matrix is below:

**Table2: The confusion matrix of the Naive Bayes**

| | Predicted negative | Predicted positive |
|---|---|---|
| Actual negative | 2 | 121 |
| Actual positive | 1 | 821 |

From table2, it is seen that the Naive Bayes model works better at predicting the positive reviews than negative reviews. All the negative reviews are wrongly predicted as positive. The reason is that the dataset is imbalanced with most of the comments as positive.

The accuracy of the logistic regression model is 87.4%, and the confusion matrix is in table3.

**Table3: The confusion matrix of the Logistic Regression**

| | Predicted negative | Predicted positive |
|---|---|---|
| Actual negative | 5 | 118 |
| Actual positive | 1 | 821 |

The accuracy of Logistic model is just a little better than Naïve Bayes, which means the logistic regression model is also good at predicting the positive reviews in this project. Machine learning algorithms are usually evaluated by their predictive accuracy, but this is not appropriate when the data is unbalanced(Chawla et al., 2002). Next this paper will deal with the problem.

# Dealing with class-imbalance

Class imbalance is an imbalance of class distribution of training set(Determined22, 2017). The ideal situation is that the number of positive and negative samples is similar. However, if there are 995 positive samples and only 5 negative samples, it means that there is class imbalance. From the perspective of training model, if the number of samples in a category is small, the "information" provided by this category is too small, then the model does not learn how to identify a few classes. In this paper, a classical oversampling algorithm called SMOTE (Synthetic Minority Oversampling) is adopted-the minority class is over-sampled by creating “synthetic” examples rather than substitution (Chawla et al., 2002). Then the confusion matrix has been changed into below.

**Table4: The confusion matrix of the Naive Bayes with SMOTE**

| | Predicted negative | Predicted positive |
|---|---|---|
| Actual negative | 85 | 38 |
| Actual positive | 90 | 732 |

**Table5: The confusion matrix of the Logistic Regression with SMOTE**

| | Predicted negative | Predicted positive |
|---|---|---|
| Actual negative | 91 | 32 |
| Actual positive | 123 | 699 |

Comparing the confusion matrix before and after SMOTE, it can be observed that the model prediction in negative has been improved. The Receiver Operating Characteristic (ROC) curve can summarize the performance of classifier and the Area Under the Curve (AUC) is a traditional performance metric for a ROC curve(Chawla et al., 2002). AUC is the area under the Roc curve, between 0.1 and 1. As a numerical value, AUC can directly evaluate the quality of classifier. The larger the value, the better the model.

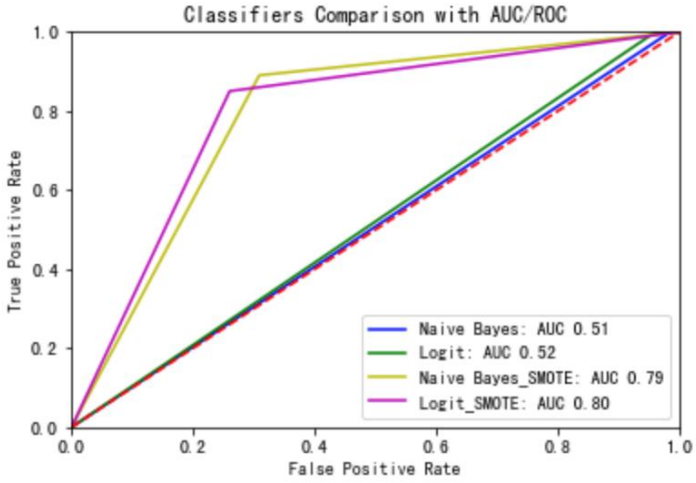



**Figure10: Classifiers comparison**

From the figure10, it can be seen that the AUC scores of the two models has improved from 0.5 to about 0.8 after applying the SMOTE, meaning the model turned to be better after the adjustment.

# 6. CONCLUSION

This paper has done a sentimental analysis about the Amazon Alexa reviews, and built a model to predict the sentiment of the comment given the review text. Because the imbalance distribution of the dataset, SMOTE is used to solve the problem and AUC/ROC are used to evaluate the performance of the model. Although it is text data, the distribution and characteristics of data and the selection of machine algorithm affect the accuracy of machine learning prediction. Word clouds can be used to

look at the high frequency of words, but can hardly to do more detail analysis. The limitation of the project is the accuracy of the sentiment analysis prediction is about 80%. Meanwhile, computer programs have trouble identifying such things as sarcasm, irony, jokes and hyperbole, which a person has little difficulty identifying. And not realizing that can distort the facts(Boothroyd, 2018). Actually, the sentiment analysis is not an easy task even for humans, which means sentiment analysis classifiers may not be as accurate as other classifiers. Nevertheless, Sentiment analysis tools can automatically and quickly recognize and analyze texts, it is useful for monitoring public opinion and customers likes and dislikes in such an exponentially growing information age.

# References


1. Boothroyd, A., 2018. The benefits (and limitations) of online sentiment analysis tools [WWW Document]. Typely Blog. URL https://typely.com/blog/making-use-of-sentiment-analysis/ (accessed 5.17.19).
2. Brownlee, J., 2017. A Gentle Introduction to the Bag-of-Words Model [WWW Document]. URL https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (accessed 5.15.19).
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953
4. Determined22, 2017. Machine learning -- class imbalance problem and SMOTE oversampling algorithm [WWW Document]. URL https://www.cnblogs.com/Determined22/p/5772538.html (accessed 5.15.19).
5. Fang, X., Zhan, J., 2015. Sentiment analysis using product review data. J. Big Data 2, 5. https://doi.org/10.1186/s40537-015-0015-2
6. Grefenstette, G., Tapanainen, P., 1994. What is a word, What is a sentence? Problems of Tokenization 10.
7. Hosmer, D.W., Lemesbow, S., n.d. Applied logistic regression, 2nd ed.
8. Komak, R., 2017. Alexa: Over 497 of the Funniest Questions to Ask Alexa on Amazon Echo, Echo Dot, and Amazon Tap!/Ross Komak-所有类别-Amazon.com [WWW Document]. URL https://www.amazon.cn/dp/1542608406/ref=sr_1_3?__mk_zh_CN=%E4%BA%9A%E9%A9

%AC%E9%80%8A%E7%BD%91%E7%AB%99&keywords=alexa&qid=1557732176&s=gateway&sr=8-3 (accessed 5.13.19).

9. Liddy, E.D., 2000. Text Mining. Bull. Am. Soc. Inf. Sci. Technol. 27, 13–14. https://doi.org/10.1002/bult.184
10. Liu, B., 2012. Sentiment Analysis and Opinion Mining.
11. Liu J., 2017a. The segmentation principle of text mining [WWW Document]. URL https://www.cnblogs.com/pinard/p/6677078.html (accessed 5.14.19).
12. Liu J., 2017b. Summary of English text mining preprocessing process [WWW Document]. URL https://www.cnblogs.com/pinard/p/6756534.html (accessed 5.15.19).
13. Liu J., 2017c. Logical regression principle summary. Log. Regres. Princ. Summ. URL https://www.cnblogs.com/pinard/p/6029432.html (accessed 5.15.19).
14. McCallum, A., Nigam, K., 1998. A comparison of event models for Naive Bayes text classification, in: In Aaai-98 Workshop on Learning for Text Categorization. AAAI Press, pp. 41–48.
15. Monkeylearn, n.d. Sentiment Analysis: nearly everything you need to know | MonkeyLearn [WWW Document]. URL https://monkeylearn.com/sentiment-analysis/ (accessed 5.17.19).
16. Nie, chaoxiong, 2018. Artificial intelligence: python implementation chapter 10, the first day of NLP introduction and use of stemming reduction vocabulary [WWW Document]. CSDN Blog. URL https://blog.csdn.net/qq_34494334/article/details/79321363 (accessed 5.14.19).
17. Palmer, D.D., n.d. Chapter 2: Tokenisation and Sentence Segmentation 23.
18. Siddhartha, M., 2018. Amazon Alexa Reviews [WWW Document]. URL https://kaggle.com/sid321axn/amazon-alexa-reviews (accessed 5.13.19).
19. Training, validation, and test sets, 2019. . Wikipedia.